I. INTRODUCTION

Al-Qur'an is the holy book of Muslims which is a miracle revealed by Allah to the prophet Muhammad SAW. The Qur'an contains the words of Allah revealed in stages over a period of 22 years 2 months 22 days to Muhammad through the angel Gabriel (*Jibril*). The Qur'an consists of 30 parts (*juz*), 114 chapters (*surah*) and 6236 verses (Tim Redaksi, 2008). The Qur'an serves as the way of life for all Muslims in the world, so that many Muslims are eager to understand its contents. Nevertheless, the Qur'an contains many words that have more than one meaning, presenting certain difficulties in understanding. For example, the word أَزْوَاجًا has two equivalents as it might be translated either "*jodoh*" ("mate") as in Surah An-Nahl (16: 72: 6) or "*golongan-golongan*" ("groups") as in Surah Al-Hijr (15: 88: 8). Such case is known as word sense, a word or words that have more than one meaning(Samhith et al., 2016). For this reason, a word senses set was built which contains a collection of word senses from the Qur'an vocabulary, so that it can help facilitate understanding of vocabulary that has more than one meaning in the Qur'an. This study focuses on the vocabulary in the Qur'an which has more than one equivalent in Bahasa Indonesia due to the different context where it is used.

Computational linguistics, particularly involving Natural Language Processing (NLP) and Princeton WordNet (PWN)(Miller et al., 1991), is one of the most popular and most widely used lexical databases. It is used as the research materials in the construction of WordNet for Arabic language (Elkateb et al., 2006), set of synonyms (synset) for Bahasa Indonesia (Gunawan & Saputra, 2010), synset of Qur'an vocabulary using WordNet approach (Gupitasari, 2019), arabic corpus (Al-Thubaity et al., 2013) and so forth. But for the arabic corpus the Qur'an is still currently still lacking even though in the last few years there has been the development of arabic corpus which are free of access. These corpus can be said not enough to encourage linguists to apply them to their corpus-based (Al-Thubaity et al., 2013). In previous research as well, namely a construction of wordnet for Arabic (Elkateb et al., 2006). Discussion about the set of word senses is only explained as a form of lexical ambiguity but was not included in the wordnet which was made as a set of word senses. Research on set of synonyms (synset) for Indonesian (Gunawan & Saputra, 2010) explains about making synonyms sets (synset) using clustering technique that can be used as a reference for this research using the same technique. However, this research does not collect synonyms from Indonesian language set but collecting the set of Arabic word senses. Another difference from this study with previous research (Gupitasari, 2019) is that this study collects Qur'an words which have different meanings but with the same lemma. Whereas the previous research discussed about vocabulary that has similar meaning.

Related to the previous studies mentioned above, this research was conducted to construct a set of word senses from Arabic vocabulary used in Al-Qur'an by utilizing lexical semantic similarity on PWN. Previous research about construction set of the synonyms from vocabulary in Al-Qur'an by using a WordNet approach has successfully constructed a lexical database focusing on a set of synonyms (Gupitasari, 2019). The research for a set of word senses from Al-Qur'an vocabulary to build Al-Qur'an thesaurus is still rarely found, while the thesaurus is not only contains synonyms, but also includes the word senses and it is deemed incomplete enough. Therefore this research becomes important to build a collection of word senses that can be used as a prototype to build Al-Qur'anic thesaurus. The construction of this set of word senses from Al-Qur'an vocabulary is using the same techniques as those used in the construction of synonym sets for Indonesian (Gunawan & Saputra, 2010)and the construction of synonym sets for Arabic

vocabulary (Gupitasari, 2019), namely clustering techniques. The clustering technique that is used in this research is hirerarchical clustering. The primary reasons of choosing hierarchical clustering are because the result of the clusters can not be predicted before the clustering process is done (therefore, the partitional clustering is inappropriate for this need) and the simplicity concept of hierarchical clustering itself. Furthermore, the method of hierarchical clustering used is the agglomerative method. Then all members of the set of word senses that have been formed would be evaluated by comparing them with the gold standard, and the accuracy would be calculated using the F measure method.

The purpose of this research is to help linguists to collect the word senses set of vocabulary Quran, it is expected that the construction of this set of word senses can be used to facilitate understanding of the vocabulary in the Qur'an and the result of word senses set from this research can be used as a prototype to create a corpus of the Qur'an and tesaurus in order to increase the resources of Qur'an research.