OPEN ACCESS
Indo-JC Indonesia Journal of Computing
ISSN 2460-9056
socj.telkomuniversity.ac.id/indojc

Ind. Journal on Computing
Vol. 5, Issue. 2, September 2020. pp. 99-106
doi:10.21108/indojc.2020.5.2.496

Anaphora Resolution on Al-Quran with Indonesian Translation

Arlinda Dwi Ardiyani [#1], Moch Arif Bijaksana [*2], Arif Fatchul Huda [#3]

[#] *Department of Informatics, Telkom University*
Jl. Telekomunikasi Terusan Buah Batu, Bandung, West Java 40257, Indonesia

[1] arlindadwia@student.telkomuniversity.ac.id

[2] arifbijaksana@telkomuniversity.ac.id

[3] afhuda@uinsgd.ac.id

Abstract

Al-Quran is the holy book of Islam, in Al-Quran we often find many cases of *anaphora*. *Anaphora* is a pronoun, for example "it" which refers to an object (antecedent) in the previous sentence. *Antecedent* of a pronoun is very important to understand the Al-Quran. *Coreference Resolution* with the classification model using the *Support Vector Machine* method are needed to find out the *antecedent*. In this research, we use i feature and j feature for the extraction process. Based on the evaluation results, the system is able to find the *antecedent* of an *anaphor* with the best accuracy value of 88.08%.

Keywords: *anaphora resolution, coreference resolution, support vector machine*, al-quran.

Abstrak

Al-Quran merupakan kitab suci umat islam. Pada Al-Quran sering ditemukan kasus anaphora. *Anaphora* merupakan kata ganti seperti "it" yang mengacu pada suatu objek (antecedent) dikalimat sebelumnya. Mengetahui *antecedent* dari sebuah kata ganti sangat penting untuk memahami Al-Quran. Untuk mengetahui *antecedent* tersebut diperlukan *Coreference Resolution* dengan model pengklasifikasi menggunakan metode *Support Vector Machine*. Penelitian ini menggunakan fitur i dan fitur j. Berdasarkan hasil evaluasi, diketahui bahwa sistem mampu mengetahui *antecedent* dari sebuah *anaphor* dengan nilai akurasi terbaik sebesar 88.08%.

Kata Kunci: *anaphora resolution, coreference resolution, support vector machine*, al-quran.

Received on xxx, accepted on xxx, published on xxx

I. Introduction

ALL Religious people have a way of life, especially Muslims who have a way of life of the Quran. Al-Quran is the main source of Islamic teachings [10]. Al-Qur'an is the Muslim holy book which was revealed using Arabic [15]. Al-Quran has a unique and distinctive characteristics, with writing or style of language [6].Especially the Quran uses the original Arabic (Al-Quran) which is translated into various languages such as Indonesian, in the Koran often found cases of *anaphora* [13]. *Anaphora* is a pronoun, for example "it" which refers to an object (antecedent) in the previous sentence. *Anaphora Resolution* is a way to show a substitute word to an object that is meant previously [14]. One example of anaphora is the Surah Al-Fatihah, the insured verse which is "Hanya kepada Engkaulah kami menyembah dan hanya untuk engkaulah

kami meminta pertolongan". We can see the sentence "hanya kepada Engkaulah", which refers to a particular object (antecedent). To find out the *antecedent*, we could use *Coreference Resolution* with the classification model using the *Support Vector Machine* method.

Coreference Resolution is the process of identifying a collection of noun phrases that refer to one and the same real-world entity [9]. According to Andrew Radford in 1988, *Coreference Resolution* is a relation between several terms that have the same reference. *Coreference Resolution* is a subtask of *Natural Language Processing (NLP)*, the task is to find all expressions that refer to the same entity. This is an important step for *Natural Language Processing* assignments that involve understanding *Natural Language* [1].

After doing Coreference Resolution and produce pairs of words that have been labeled. Furthermore, the classification process is carried out, the classification process in this study uses the *Support Vector Machine* method, where the method developed by Boser, Guyon, Vapnik, and presented for the first time at the Annual Workshop on Computational Learning Theory in 1992. *Support Vector Machine (SVM)* works using the principle of *Structural Risk Minimization (SRM)* with the aim of finding the best hyperplane to separate two classes of data [7]. *Support Vector Machine* method for the *Coreference Resolution* case has been done before, carried out by Ayu Linggar Sari on *Coreference Resolution* by using the *SVM* method in Indonesian novels. Based on these results, it can be concluded that the *SVM* method is actually suitable for *Coreference Resolution* in Indonesian novels, but equipped with training data that has different patterns and is equipped with compound pronoun detection to get high accuracy values [8]. However, no research has been conducted on *Coreference Resolution* in the Al-Quran with Indonesian translations using the *SVM* method so that its accuracy is still unknown.

This research can make it easier for readers to find out the contents of Al-Quran without any misinterpretation in Al-Quran, because the results of Arabic translation into other languages such as Indonesian cause confusion to determine which pronouns in the content of the Al-Quran verses refer to where. For example in surah al-fatihah the fifth verse which is "Hanya kepada Engkaulah kami menyembah dan hanya kepada Engkaulah kami meminta pertolongan". In that verse there is the phrase or words "Hanya kepada Engkaulah" which doesn't know where to refer. This research helps the reader to understand that the phrase or word "Hanya kepada Engkaulah" (antecedent) refers to the word "Allah" (anaphor) in the first verse in surah Al-Fatihah. Knowing antecedent from a anaphor can minimize errors in understanding the contents of the Al-Quran. Based on the background or introduction that has been described, the formulation of the problem in this study is how to determine the *antecedent* of an *anaphor* which focuses on the PRON pronouns on the Al-Quran by using the *SVM* method. Knowing the *antecedent* of a pronoun is very important to understand the Qur'an [13]. Then the purpose of this research is to determine the *antecedent* of an *anaphor* in the Al-Quran by using the *SVM* method.

II. Literature Review

The journal published by Ayu Linggar Sari is about *Coreference Resolution* using the *Support Vector Machine* method in Indonesian novels. This journal uses Indonesian-language novel documents as input and uses the *Support Vector Machine (SVM)* method for *Coreference Resolution*. This research uses 7 features, they are *i-pronoun, i-propernoun, j-pronoun and j-propernoun*. The *distance of words* feature, *distance of sentences* feature and *string match* feature are used for relations between two of them. These stages are the training stage, the testing stage, the classification stage using *SVM* as well as the calculation of the resulting *coreference resolution* accuracy. Based on the results of research with 2 training data in the form of novels with a total of 10440 pairs of *coreference* pairs and test data in the form of 30 novel quotations containing pronouns or named entities produce an average accuracy of 50.14% with the highest accuracy of 63% while the lowest accuracy is equal to 38.8%. This is due to the limitations of the training data used because each novel has a different pattern and limited detection for compound pronouns. So

it can be concluded that the *Support Vector Machine* method is actually suitable for Coreference Resolution in Indonesian novels, but is equipped with training data that has different patterns and is equipped with compound pronoun detection to get high accuracy values [8].

The journal entitled Research of Noun Phrase Coreference Resolution, published by Gao Junwei, discusses an approach-based learning for *coreference resolution* with noun phrases in Chinese. In his research, he used 15 features, some of the features used include *Distance, String Match, Alias, Appositive, i-pronoun* and others. There are several features that provide a high enough value, namely *String Match, Alias* and *Similarity*. The results of the study by Gao Junwei had a *precision* performance of 52.97%, a *recall* of 49.95% and an *F-measure* of 51.41% [4].

III. Research Method

Below is a flowchart in this research

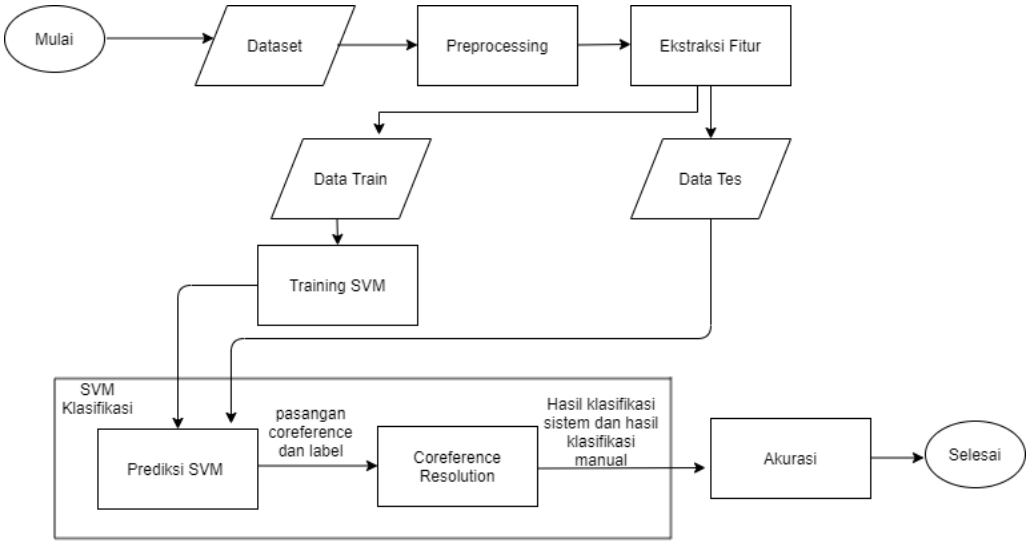

Figure 1: Flowchart System

Figure 1 shows the flow of the system running process that starts from the dataset, then *preprocessing*, then feature extraction process is performed, after that the dataset is divided into *Train* data and *Test* data. The *SVM* classification process is then performed. The results of the classification process will be calculated for their accuracy.

A. Anaphora Resolution

Anaphora comes from the Greek language, *Anajora*. Ana means back and Jora means brings. So *Anaphora* is the act of bringing back, which means that *anaphora* is to show a substitute word to an object referred to in the previous sentence. *Anaphora* which represents the relation between terms is called *anaphor*, while *anaphora* that represents another object is called *antecedent* [12]. For the example is in surah Al-Fatihah verse (1:5:14) إِيَّاكَ which means "hanya kepada Engkaulah" which refers to verse (1:1:2) اللَّهِ meaning "Allah". Which is the pronoun "hanya kepada Engkaulah" here is a pronoun from the word "Allah".

B. Coreference Resolution

Coreference Resolution is the process of identifying a collection of noun phrases that refer to one and the same real-world entity [3]. According to Andrew Radford in 1988, *Coreference* is a relation between several terms that have the same reference. *Coreference Resolution* is a sub-task of *Natural*

Language Processing (NLP) whose job is to find all expressions that refer to the same entity. This is an important step for NLP assignments that involve understanding *Natural Language* [4]. The following is an example of the *Coreference Resolution* process[4]. The following is an example of the *Coreference Resolution* process :

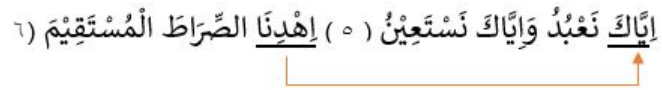

Figure 2: Coreference Resolution Process

After completing the *Coreference Resolution* process in Figure 2, the results of the *Coreference Resolution* process in Table I are as follows:

Table I: The result of Coreference Resolution Process

Sequence_i	Arab_i	i	Tag_i	Sequence_j	Arab_j	j	Tag_j	Manual
(1:5:14)	إِيَّاكَ	Hanya kepada Engkaulah	PRON	(1:1:2)	اللَّهِ	Allah	PN	Coreference
(1:5:16)	وَإِيَّاكَ	dan hanya kepada Engkaulah	PRON CONJ	(1:1:2)	اللَّهِ	Allah	PN	Coreference
(1:5:16)	وَإِيَّاكَ	dan hanya kepada Engkaulah	PRON CONJ	(1:5:14)	إِيَّاكَ	Hanya kepada Engkaulah	PRON	Bukan Coreference
(1:6:18)	اِهْدِنَا	Tunjukilah kami	PRON V	(1:1:2)	اللَّهِ	Allah	PN	Bukan Coreference
(1:6:18)	اِهْدِنَا	Tunjukilah kami	PRON V	(1:5:14)	إِيَّاكَ	Hanya kepada Engkaulah	PRON	Bukan Coreference

In Table I, the column sequence_i and sequence_j are the location of the word, where the first number indicates the surah of the word, the second number indicates the verse of the word, then the third number indicates the position of the number of words in one surah in question. In Table I in the first row there are the words “Hanya kepada Engkaulah” and the word “Allah”. If the word i has a PRON tag and the word i is the pronoun of the word j, then the *coreference*, and if the word i is not the pronoun of the word j, then it is not *coreference*.

C. Preprocessing

1) Tokenization: Tokenization is the process of cutting an input string by breaking a document into several parts. *Tokenization* can be done by looking at the delimiter, such as the type of capitalization, the presence of digits, punctuation marks, special characters and so on. Splitting a document into single words is done by scanning the document and each word is identified or separated by another word by a space separator [2].

Table II: The result of Tokenization

Sebelum	**Sesudah**
Hanya kepada Engkaulah	Hanya_kepada_Engkaulah
dan hanya kepada Engkaulah	dan_hanya_kepada_Engkaulah

The explanation of Table II is in Arabic the word ”Hanya kepada Engkaulah“ is one word, whereas in the Indonesian dictionary the word “Hanya kepada Engkaulah “ consists of three words. So that the meaning of the Arabic word is not reduced and so that the system of reading the word consists of one feature or one word, then the process is done to change the space into an underscore. So that the word ”Hanya kepada Engkaulah” in the system will read into one feature or one word.

2) Combine feature i with feature j:

Table III: Feature i and Feature j Combination

Sebelum		Sesudah
Hanya_kepada_Engkaulah	Allah	Hanya_kepada_Engkaulah Allah
dan_hanya_kepada_Engkaulah	Allah	dan_hanya_kepada_Engkaulah Allah

Table III combines feature i with feature j because in general the classification text detects word words in one text. So that the system reads data on one line consisting of two features, a merge of feature i and feature j is performed.

D. Extraction Feature

Extraction Feature is the process of extracting important characteristics or information from data for use in the classification process. Feature is an attribute that is used to represent the data used, where the feature is the result of extraction which is assumed to provide information about the data. [4] [8]. In this study using two features, namely feature i and feature j. After the feature extraction process is carried out, the two features will be used for the classification process using SVM. Examples of extraction feature results can be seen in table .

Example of extraction feature results can be seen in Table IV :

Table IV: The result of Extraction Feature

	agama_kalian	aku	Allah	apa_yang_dia_miliki_hartanya	api_neraka
Hanya_kepada_Engkaulah Allah	0	0	1	0	0
dan_hanya_kepada_Engkaulah Allah	0	0	1	0	0
dan_hanya_kepada_Engkaulah Hanya_kepada_Engkaulah	0	0	0	0	0
Tunjukilah_kami Allah	0	0	1	0	0
Tunjukilah_kami Hanya_kepada_Engkaulah	0	0	0	0	0

E. Support Vector Machine

Support Vector Machine (SVM) is a machine learning method for converting text into vector data. This method was developed by Boser, Guyon, Vapnik in 1992. SVM is a technique for finding hyperplane that can separate two data sets from two different classes (Vapnik, 1999). Hyperplane is the boundary separating data between classes, while the margin is the distance between the hyperplane and the closest data in each class. The data closest to the hyperplane in each class is called support vector [11]. This study uses a library of sklearn that supports the use of the SVM model as a classification model. In this library, there are several parameters that have been set by default by the Sklearn library. Among the most common are like the kernel, gamma, C etc. With this the researcher uses the library by adjusting several parameters such as the kernel, by default the "rbf" kernel is set for the researchers themselves to use a "linear" kernel because researchers expect a linear Hyperplane representation. Then parameters such as C (Regularization parameter), gamma etc. still use the default values from each library.

Table V: Data training in SVM format

Format SVM	Label
0 0 1 0 0	1
0 0 1 0 0	1
0 0 0 0 0	0
0 0 1 0 0	0
0 0 0 0 0	0

In Table V, label one is used for pairs that are coreference and zero labels are used for pairs that are not coreference.

F. Dataset

The data used in this research were obtained from www.corpus.quran.com. The dataset consists of 13 letters in the Al-Quran and 437 word pairs.

Table VI: *Train* Data

Sequence_i	Arab_i	i	Tag_i	Sequence_j	Arab_j	j	Tag_j	Manual
(1:5:14)	إِيَّاكَ	Hanya kepada Engkaulah	PRON	(1:1:2)	اللَّهِ	Allah	PN	Coreference
(1:5:16)	وَإِيَّاكَ	dan hanya kepada Engkaulah	PRON CONJ	(1:1:2)	اللَّهِ	Allah	PN	Coreference
(1:5:16)	وَإِيَّاكَ	dan hanya kepada Engkaulah	PRON CONJ	(1:5:14)	إِيَّاكَ	Hanya kepada Engkaulah	PRON	Bukan Coreference
(1:6:18)	اهْدِنَا	Tunjukilah kami	PRON V	(1:1:2)	اللَّهِ	Allah	PN	Bukan Coreference
(1:6:18)	اهْدِنَا	Tunjukilah kami	PRON V	(1:5:14)	إِيَّاكَ	Hanya kepada Engkaulah	PRON	Bukan Coreference

Table VI shows some *Train* data. *Train* data consists of 350 pairs of words.

Table VII: *Test* Data

Sequence_i	Arab_i	i	Sequence_j	Arab_j	j	Manual	System	Result
(109:6:23)	لَكُمْ	bagi kalian	(109:4:14)	أَنَا	aku	Bukan Coreference	....	....
(109:6:23)	لَكُمْ	bagi kalian	(109:4:17)	عبَدْتُمْ	kalian sembah	Bukan Coreference	....	....
(109:6:23)	لَكُمْ	bagi kalian	(109:5:19)	أَنْتُمْ	kalian	Bukan Coreference	....	....
(109:6:24)	دِينُكُمْ	agama kalian	(109:1:1)	قُلْ	katakanlah (Muhammad)	Bukan Coreference	....	....
(109:6:24)	دِينُكُمْ	agama kalian	(109:1:3)	الْكَافِرُونَ	orang-orang kafir	Coreference	....	....

Table VII shows a portion of the Test data that will be used for system testing. The amount of test data used was 87 pairs of words.

G. Evaluation

In this research, the data used were taken from www.corpus.quran.com as many as 13 letters. The dataset consists of 437 pairs of words. This study uses data that has been classified by SVM to calculate its accuracy.

The accuracy calculation is done using the following formula :

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100\% \quad (1)$$

Equation 1 has the following information :

- TP (True Positive) is the amount of data with a positive true value and a positive predictive value.
- TN (True Negative) is the amount of data with a negative true value and a negative predictive value.
- FP (False Positive) is the amount of data with a negative true value and a positive predictive value.
- FN (False Negative) is the amount of data with a positive true value and a negative predictive value [5] [3].

IV. Results and Discussion

A. Test Result

The test carried out is to compare the results of the *Coreference Resolution* detection analysis using a system that has been made and detection manually. System testing is done to ensure that the system is made in accordance with the design of the system that has been made before. The test results can be seen in Table VIII.

Table VIII: The result of Detection Analysis

Sequence_i	Arab_i	i	Sequence_j	Arab_j	j	Manual	System	Result
(109:6:23)	لَكُمْ	bagi kalian	(109:4:14)	أَنَا	aku	Bukan Coreference	Bukan Coreference	Benar
(109:6:23)	لَكُمْ	bagi kalian	(109:4:17)	عَبَدْتُمْ	kalian sembah	Bukan Coreference	Bukan Coreference	Benar
(109:6:23)	لَكُمْ	bagi kalian	(109:5:19)	أَنتُمْ	kalian	Bukan Coreference	Bukan Coreference	Benar
(109:6:24)	دِينُكُمْ	agama kalian	(109:1:1)	قُلْ	katakanlah (Muhammad)	Bukan Coreference	Bukan Coreference	Benar
(109:6:24)	دِينُكُمْ	agama kalian	(109:1:3)	الْكَٰفِرُونَ	orang-orang kafir	Coreference	Coreference	Benar

Based on Table VIII shows that there is a difference between detecting manually and detecting the system at *Coreference Resolution.*

Accuracy values are obtained using equation 1, which is as follows :

Table IX: Accuracy Calculation

y_prediksi	0	0	0	0	0	0	0	0	0	0
y_tes	0	0	0	0	1	0	0	0	0	1
Value	TN	TN	TN	TN	FN	TN	TN	TN	TN	FN

$$Accuracy = \frac{0+8}{0+8+0+2} * 100\%$$

B. Analysis of Testing Results

Then an evaluation using the *Cross Validation* method is divided all data into several parts (fold). The first scenario is to use K of 5 Fold, from the amount of data of 437 pairs of words obtained an average accuracy of 88.08%. The second scenario is then performed using K at 10 Fold with an average accuracy of 88.03%. In Table X and Table XI can be seen the results of the accuracy of each Fold.

Table X: Accuracy Result K 5

Fold	Accuracy
1	92,05%
2	94,32%
3	89,66%
4	82,76%
5	81,61%
Accuracy Average Result	**88,08%**

Based on Table X and Table XI, it can be seen that the largest average accuracy is obtained from K 5, which is 88.08%. Obtaining the highest accuracy value is influenced by several factors, one of which is the amount of data availability or the diversity of data patterns. The more diverse patterns of data entered into the model, the better the resulting model and the higher the accuracy.

Table XI: Accuracy Result K 10

Fold	Accuracy
1	88,64%
2	95,45%
3	95,45%
4	93,18%
5	86,36%
6	93,18%
7	93,18%
8	72,09%
9	76,74%
10	86,05%
Accuracy Average Result	**88,03**%

V. Conclusion

Based on the research results obtained, that by using the *Support Vector Machine (SVM)* method as a classification method, researchers are able to determine the antecedent of an anaphor in the verses of the Al-Quran by using an Indonesian translation. The result of this method is a *coreference* value (refers to a reference) and not a *coreference* (not referenced), thus *SVM* is able to determine an antecedent from an anaphor in the verses of the Koran from these results.

References

[1] Indra Budi, Stephane Bressan, and Nasrullah. Co-reference resolution for the indonesian language using association rules. In *iiWAS*, pages 117–126, 2006.
[2] Herny Februariyanti and Eri Zuliarso. Klasifikasi dokumen berita teks bahasa indonesia menggunakan ontologi. *Dinamik*, 17(1), 2012.
[3] Mercury Fluorida Fibrianda and Adhitya Bhawiyuga. Analisis perbandingan akurasi deteksi serangan pada jaringan komputer dengan metode naïve bayes dan support vector machine (svm). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN*, 2548:964X, 2018.
[4] Junwei Gao, Fang Kong, Peifeng Li, and Qiaoming Zhu. Research of noun phrase coreference resolution. In *2011 International Conference on Asian Language Processing*, pages 93–96. IEEE, 2011.
[5] Jiawei Han, Micheline Kamber, and Jian Pei. Data mining concepts and techniques third edition. *Morgan Kaufmann*, 2011.
[6] Rahil Helmi. Majas yang terkandung dalam al-quran terjemahan surat al-baqarah. 01 2017.
[7] Evi Fitria Umi Latifah et al. Perbandingan kinerja machine learning berbasis algoritma support vector machine dan naive bayes (studi kasus: Data tanggapan mengenai traveloka melalui media sosial twitter). 2018.
[8] Ayu Linggar Sari. *Coreference Resolution Dengan Menggunakan Metode SVM Pada Novel Berbahasa Indonesia*. PhD thesis, Universitas Komputer Indonesia, 2017.
[9] Vincent Ng. Unsupervised models for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 640–649. Association for Computational Linguistics, 2008.
[10] Linda Norhan and Laras Sanjaya. Aplikasi pembelajaran menyusun ayat sebagai metode menghafal al-qur'an (juz 30). *Jurnal Online Informatika*, 1(2):87–91, 2016.
[11] Pusphita Anna Octaviani, Yuciana Wilandari, and Dwi Ispriyanti. Penerapan metode klasifikasi support vector machine (svm) pada data akreditasi sekolah dasar (sd) di kabupaten magelang. *Jurnal Gaussian*, 3(4):811–820, 2014.
[12] TW Roeper and B Lust. Studies in the acquisition of anaphora, vol. 1: Defining the constraints. 1986.
[13] Abdul-Baquee M Sharaf and Eric Atwell. Qurana: Corpus of the quran annotated with pronominal anaphora. In *LREC*, pages 130–137. Citeseer, 2012.
[14] Astria Kurniawan Sumantri, Indra Budi, and H. Kurniawan. Perbandingan decision tree, maximum entropy, dan association rules pada resolusi koreferensi untuk bahasa indonesia. 2010.
[15] Ida Latifatul Umroh. Keindahan bahasa al-qur'an dan pengaruhnya terhadap bahasa dan sastra arab jahily. *DAR EL-ILMI: Jurnal Studi Keagamaan, Pendidikan dan Humaniora*, 4(2):49–65, 2017.