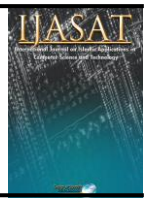

Multiword corpus of the Holy Quran

Mohammed Zeki Khedher
Al-Mishkat Center for Research, Studies and Consultations
khedher@gmail.com

Abstract
Text of Quran has been served in many forms. Recent attempts have presented such serves in corpus form. However no previous attempt made to deal with multiword in Quran in corpus form. This paper is based on presenting multiword corpus of Quran based on roots of the word rather than the word itself. This combines the relations between same words in addition to other words with common roots. This enriches the corpus since Arabic language word structure is based on root of the word. This corpus should enable future research in going deep in analyzing Quran ontology, Arabic language studies and inter relations between Quran and other Islamic resources, e.g. Hadith, Fiqh, Usool etc.
***Keywords*: Multiword corpus, Holly Quran, Quran, Corpus, Arabic Multiword**

1. INTRODUCTION

Corpora are the main knowledge base in corpus linguistics. The analysis and processing of various types of corpora are also the subject of much work in computational linguistics, speech recognition and machine translation, where they are often used to create hidden Markov models for part of speech tagging and other purposes. Corpora and frequency lists derived from them are useful for language teaching. Corpora can be considered as a type of foreign language writing aid as the contextualized grammatical knowledge acquired by non-native language users through exposure to authentic texts in corpora allows learners to grasp the manner of sentence formation in the target language, enabling effective writing (Yoon and Hirvela, 2004).

The Quranic Arabic Corpus (Corpus, 2011) is an international collaborative linguistic project initiated at the University of Leeds. This open source resource includes part-of-speech tagging for the Quran, morphological segmentation and a formal representation of Quranic syntax using dependency graphs. Version 0.4 of the corpus had been released in May 2011. Further description of the corpus will be discussed later in this paper.

The meaning of a word is not to be known by examining it in isolation, but by the company that it keeps. It is described by the associations of words rather than by individual words. Examples of meaning by collocation such as ‘one of the meanings of night is its collocability with dark'. The test of collocability refers to the notion that words are collocates when they are found to be associated with sufficient frequency to exclude the possibility that they are chance co-occurrences (Greaves and Warren, 2010).

Multiword expression (MWE) 2 or more orthographic words/lexemes that function together as an idiomatic whole. Idiomatic means not fully predictable in form, function, and/or frequency (Schneider and Smith, 2015).

A considerable interest has been given to Multiword Expression (MWEs) identification and treatment. The identification of MWEs affects the quality of results of different tasks heavily used in natural language processing (NLP) such as parsing and generation. Different approaches for MWEs identification have been applied such as statistical methods which employed as an inexpensive and language independent way of finding co-occurrence patterns (Meghawry et al., 2015).

In a study a corpus-driven approach is utilized to identify the most common multi-word patterns in conversation and academic writing, and to investigate the differing pattern types in the two registers. The linguistic characteristics of two types of multi-word sequences: 'multi-word lexical collocations' (combinations of content words) versus 'multi-word formulaic sequences' (incorporating both function words and content words) were introduced (Douglas, 2016).

A comprehensive and broad-coverage framework for representing diverse MWEs in corpora, without requiring a lexicon was introduced (Nathan, 2014). The approach to MWEs in context is not restricted to a particular lexical or even syntactic inventory of candidates. Included are the full spectrum of MWE classes - ranging from the most fixed (proper names, nominal compounds, connectives like as well as, idioms like by and large) to the most flexible (especially verb phrase expressions subject to internal modification or other syntactic processes affecting word order and/or contiguity). For example, the expression whose citation form is pay attention to could be instantiated as paid no attention to or attention was paid to, both of which contain gaps between the lexicalized parts of the expression. Further, the object of the preposition is not part of the MWE, so the MWE is not a complete constituent by a standard syntactic analysis.

A corpus-based system is presented to expand multi-word index terms using a part-of-speech tagger and a full- fledged derivational morphological system, combined with a shallow parser (Klavans and Jacquemin. 1997). The unique contribution of that research was in using these linguistically based tools with filters in order to avoid the problems of semantic degradation typically associated with derivational analysis. The expansion and subsequent conflation of terms increases indexing coverage up to 30%, with precision of nearly 90% for correct identification of related terms. The system core is language independent and provides a uniform platform on which to build multilingual applications. Language specific modules have been developed for English and French.

Analysis of the production of multi-word units present in English argumentative texts written by non-native speakers of the language is presented (Peromingo, 2010). The aim of this study is to examine the potential influence of the mother tongue on learners' production of both correct and incorrect multi-word units that are typically used in English for creating textual cohesion: lexical bundles, in particular, linking adverbials.

Both standard and focused retrieval tasks were addressed based on comprehensible language models and interactive query expansion (IQE) (Eric and Fidelia, 2010). Query topics were expanded using an initial set of Multi Word Terms (MWTs) selected from top n

ranked documents. MWTs are special text units that represent domain concepts and objects. As such, they can better represent query topics than ordinary phrases or ngrams.
The Manawi system (Liling and Santanu, 2014) included two innovations: (i) the usage of outputs from NLP tools, viz. billingual multi-word expression extractor and named-entity recognizer to improve SMT quality and (ii) the introduction of a novel filter method based on sentence-alignment features. The Manawi system showed the potential of improving translation quality by incorporating multiple NLP tools within the MT pipeline.

A corpus-driven approach (Annelen and Steyer, 2007) is used to the study of multi-word expressions, which constitute a significant part of language. Using the largest available collection of written German which has approximately two billion word tokens and is located at the Institute for the German Language (IDS). A strongly usage-based approach to multi-word expressions was used. To find multi-word expressions, researchers were guided by corpus data.

Statistical methods were applied to perform automatic extraction of Hungarian collocations from corpora (Lexemes et al., 2004). Due to the complexity of Hungarian morphology, a complex resource preparation tool chain has been developed. This tool chain implements a reusable and, in principle, language independent framework.

A list of Arabic multiword expressions (MWE) had been collected from various dictionaries (Abdelati et al., 2012). The MWEs are grouped based on their syntactic type. Every constituent word in the expressions was manually annotated with its full context-sensitive morphological analysis. Some of the expressions contain semantic variables as place holders for words that play the same semantic role.

To extract multi-word terms from corpora (Boulaknadel et al., 2008): first, the linguistic specification of MWTs for Arabic language was defined. Then, a term extraction program was developed and several statistical measures were evaluated in order to filter the extracted term-like units for keeping the most representative of domain specific corpus.

An investigation of the degree of usefulness of dictionaries when rendering English and Arabic multi-word items, such as idioms, collocations, phrasal/prepositional verbs, and compounds/iḍāfas was made (Alnaser, 2010). The multi-word items are known for their metaphorical meanings and fixed structures, as both characteristics cause confusion to the translator/foreign language learner. The usefulness of the translation dictionaries was determined based on two criteria. First, by evaluating the use of these dictionaries for the rendering of the multi-word items in undergraduate translation and lexicography classes. Second, by assessing the lexicographical documentation and treatment of these items in those dictionaries. It has been concluded that the percentages of dictionary use in advanced classes of translation were higher, which indicates the awareness of the importance of dictionaries in these classes. In addition, students of Arabic-English translation classes used dictionaries less than the English-Arabic classes since they dealt with texts of their native language and that English multi-word items were more difficult to render than the Arabic ones. Moreover, findings showed that Arabic multi-word items were treated better than the English multi-word items in their respective dictionaries even though the English-Arabic dictionaries document more than the Arabic-English dictionaries.

2. QURAN CORPUS

The Quranic Arabic Corpus (Corpus, 2011) as mentioned above is a collaboratively constructed linguistic resource initiated at the University of Leeds, with multiple layers of annotation including part-of-speech tagging, morphological segmentation and syntactic analysis using dependency grammar. The motivation behind that work was to produce a resource that enables further analysis of the Quran. The project contrasts with other Arabic treebanks by providing a deep linguistic model based on the historical traditional grammar known as i'rāb (إعراب). A new approach to linguistic annotation of an Arabic corpus was introduced : online supervised collaboration using a multi-stage approach. The different stages include automatic rule-based tagging, initial manual verification, and online supervised collaborative proofreading. A popular website attracting thousands of visitors per day, the Quranic Arabic Corpus has approximately 100 unpaid volunteer annotators each suggesting corrections to existing linguistic tagging. To ensure a high-quality resource, a small number of expert annotators were promoted to a supervisory role, allowing them to review or veto suggestions made by other collaborators. The Quran also benefits from a large body of existing historical grammatical analysis, which may be leveraged during that review. In this paper we evaluate and report on the effectiveness of the chosen annotation methodology. Discussions of the unique challenges of annotating Quranic Arabic online described the custom linguistic software used to aid collaborative annotation (Dukes et al., 2011).

3. THE LEXICON OF MULTIWORDS IN THE HOLLY QURAN

The most famous lexicon of words of Quran is based on words roots (Mohammed Fouad, 2008). Roots are sorted alphabetically. Another lexicon was compiled and made available on the internet is the one by the author (Khedher, 2001) with two parts. Multiword lexicon for Quran was unavailable before the lexicon by the author based on multiword sorted according to corresponding roots of words of the multiword and was published in 2002 (Khedher, 2002). The lexicon is sequenced in an alphabetical order of the roots i.e. the sequence of the root of the first word in the multiword, second root of the second word, etc. When the multiwords are of the same words, the sequence used to start with the one first occurring in the sequence of the whole Quran. But when tashkeel is regarded in the sequence then the "fatha" is taken first, followed by "DHamma" followed by "kasra"

The text of Quran used was as near to the othmanic style of writing as possible. The style was chosen according to "Mushaf Al-Madinah". The only exception for that is to convert the unwritten Alef (known as Alef khanjariyah) to a written one except for the cases widely used otherwise. As an example the words which are in othmanic style: "ملك العلمين الكتب" were used as "مالك العالمين الكتاب " while the common words: " ذلك هذا الرحمن" were kept as they are as these are commonly written like that.. In cases for some words which are written differently in different locations, they were kept as they are e.g. " رحمت رحمة". The first verse of Chapter 1 was considered as "بسم الله الرحمن الرحيم" and not the first verse of all chapters of the Quran, as there are some difference of opinions in this respect.

Number of verses of Quran according to "Mushaf Al-Mdinah" is 6236 verses. Number of words is 77479 words. It was found that the number of unrepeated words was 18841 words according to the othmanic style of writing. While when the above conversion of "Alef " was followed the number becomes 18232 word. There was another conversion, that is the

"shaddah" on the beginning of some words if the preceding word ending with "tanween". When cancelling this "shaddah", the number is reduced to 17884 words. Words like يعلمون ، سيعلمون ، عالمون ، العلماء ، علام ، عليم ، علَّم ، معلَّم are all considered as of the root: "علم" and the verbs استقام وأقام وقام have the root: " قوم "

Roots of the words were taken from references (Mohammed Fouad, 2008) and (Khedher, 2001). When dealing with roots of the words as a pronoun "مِن" and "مَن" as a preposition, it was necessary to distinguish between them, so that an extension to the first character was added for the first one to become "مـن" while the second one was kept without extension "من".

Most of the Arabic words (nouns or verbs) refer to a 3-characters root. Some words refers to 4-characters root (nouns or verbs). For names of persons, جبريل وميكال وعيسى وموسى ونوح وآدم locations, مكة ومصر ويثرب etc., te same names are considered as roots. Special Characters which are at the beginning of some chapters of the Quran were considered as the same as their roots. Prepositions and other words which are neither nouns nor verbs, which have no root and are near to each other were considered of the same origin and have a supposed common root, e.g. "لم" و"لن" و"لا" or"أن" و"إن"

Some words which refer to the same thing may have been written differently in different locations e.g. "إبراهيم " و " إبراهم" which refer to Prophet Abraham. The total number of perfect roots and supposed roots (accordintg to the above remarks) becomes 1768. All words whose root appears once in the whole Quran can be considered as stop words between multiwords. Multiwords considered were not only in the same verse but even between successive verses were considered as continuation with the sign '#" in between. The longest multiwords found in the Quran contain 26 words which are of similar roots are the verses 5-8 Chapter 23 and 29-32 Chapter 70

No continuation between multiwords of different chapters. A multiword is considered only if it consists of at least two consecutive words whose roots are common with roots of other consecutive words. Such multiwords occure at least twice in the whole Quran.

The procedure for finding the target multiwords is as follows (Khedher, 2002):

1. Starting from the text of the Quran associated with corresponding roots. Text includes signs showing chapters and verses separations.
2. Processing the text so as to consider the multiword length of 30 words at the beginning.
3. Comparing all the roots series of the 30 words multiwords to find out of there are any similar multiword with same corresponding roots series.
4. It was found that there are no such similar multiwords. Hence the number is reduced by one word to 29 and the same test is repeated and so on in reducing the number of words.
5. It was found that the first similar multiwords were of 26 words lengths. They are : from chapter 23 (Al muminoun)

وَالَّذِينَ هُمْ لِفُرُوجِهِمْ حَافِظُونَ (5) إِلاَّ عَلَى أَزْوَاجِهِمْ أَوْ مَا مَلَكَتْ أَيْمَانُهُمْ فَإِنَّهُمْ غَيْرُ مَلُومِينَ (6) فَمَنِ ابْتَغَى وَرَاءَ ذَلِكَ فَأُولَئِكَ هُمُ الْعَادُونَ (7) وَالَّذِينَ هُمْ لأَمَانَاتِهِمْ وَعَهْدِهِمْ رَاعُونَ (8)

And from chapter 70 (Al maarij)

وَالَّذِينَ هُمْ لِفُرُوجِهِمْ حَافِظُونَ (29) إِلاَّ عَلَى أَزْوَاجِهِمْ أَوْ مَا مَلَكَتْ أَيْمَانُهُمْ فَإِنَّهُمْ غَيْرُ مَلُومِينَ (30) فَمَنِ ابْتَغَى وَرَاءَ ذَلِكَ فَأُولَئِكَ هُمُ الْعَادُونَ (31)

وَالَّذِينَ هُمْ لأَمَانَاتِهِمْ وَعَهْدِهِمْ رَاعُونَ (32)

6. The same procedure is repeated with reduction of multiword's length till a length of 2 words is reached.
7. All multiwords were associated with their chapters and verses numbers. If the multiword is spread over more than one verse, the beginning verse number and the ending one are separated by (-) and the words sequence and roots sequence by (*)
8. To group multiwords in different groups, all multiwords of same roots and same words are grouped together with sequence according to sequence of chapters and verses of the Quran
9. This is followed by groups of the same roots sequence but different words
10. There are two possibilities in presentation of the multiwords: the first is according to alphabetical sequence of roots. This was the form found in the reference book (Khedher, 2002). The second possibility is to present the multiwords in a sequence which agrees with the appearance in the Quran of the first multiword in the group. The first time the multiword appears, all associated multiwords are presented. In the locations of these multiwords, a reference to the first location is given.
11. Long multiwords may have associated multiwords. However part of these long multiwords when considered as a new multiword may have associated multiwords more than what the long multiword has. Hence these short multiwords were considered as new multiwords and are presented with their associated multiwords after the long multiwords. This action was done by a special program.
12. Number of repetition of multiwords in each case is shown in the book (Khedher, 2002). Name of the chapter and the verse number are included.

Manual work in some stages of the development of the final form of information was inevitable due to the nature of the text of Quran. It was impossible to follow a fixed routine which is so general to include all cases.

Figure (1) shows the flow chart for getting the multiword from the database of the Quran.

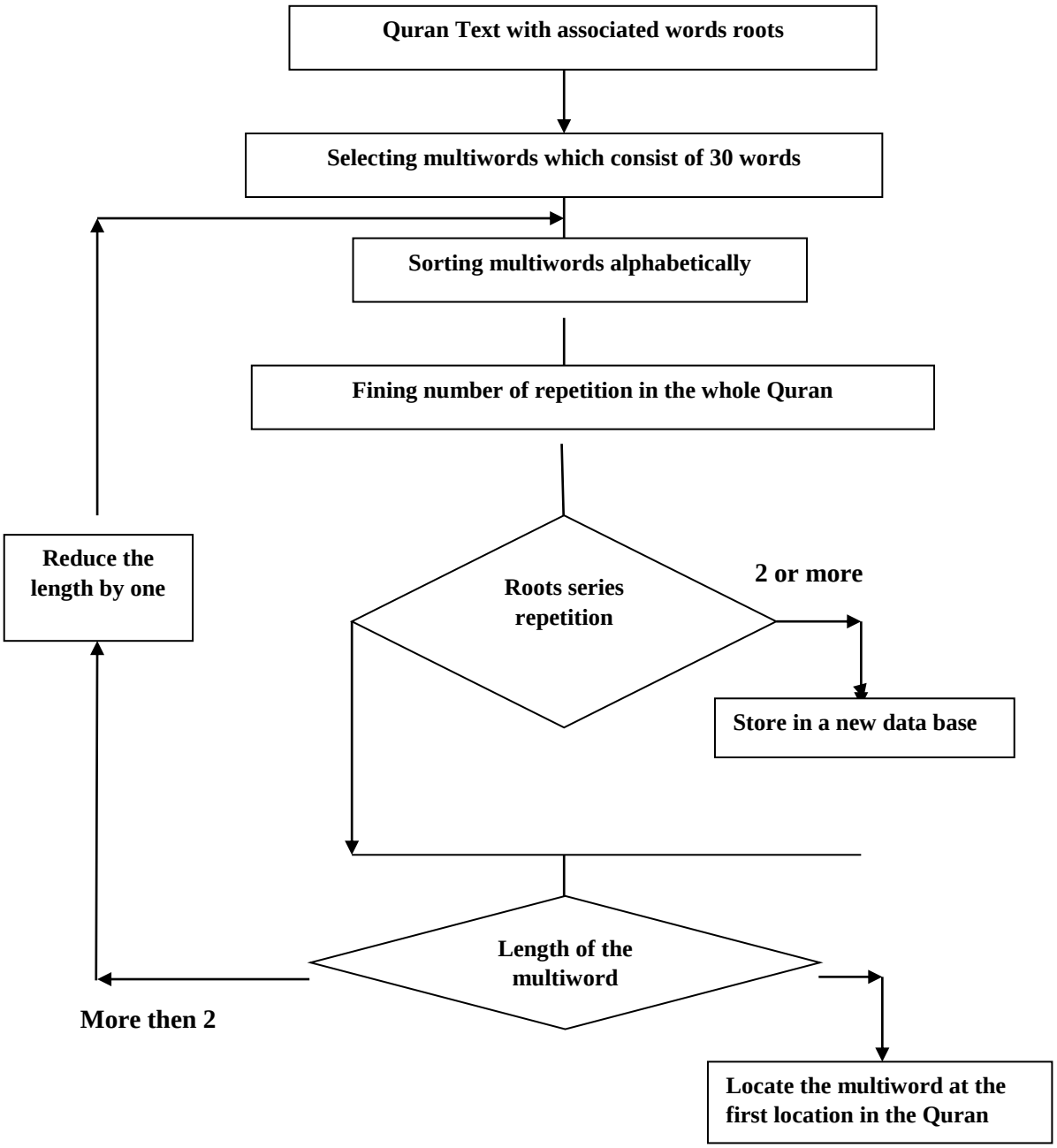

Figure 1: Flow chart of finding repeated multiword

4. MULTIWORD CORPUS OF QURAN

The above information about multiword lexicon are redesigned and put in a corpus form as follows:

1. It has been put in the order of appearance in the sequence of Quran chapters and verses. When the multiword appears, all other multiword with the same roots are given with it. First those with the exact wording followed by those with different wording but same roots. Each group contains multiword in sequence of appearance in the Quran chapters and verses.
2. Each entry will contain the information about it and those of the first multiword in the group.
3. The entry is repeated in its place (denoted by RPH) in the chapter and the verse it appears, with reference to its first place so as to know the group and its contents.
4. The information for each entry consists of 6 items (denoted by RF) for the entry itself, 6 items for its reference (denoted by RPH), text of the multiword (denoted by PH) and the text of the corresponding roots (denoted by RT) i.e.

RF[A : B : C : D : E : F] RPH[A1 : B1 : C1 : D1 : E1 : F1] PH[multiword] RT[roots of multiword]

Where:

A: Serial Number of the Chapter (Sura): 1= الفاتحة…114= الناس

B: Serial Number of the Verse (Aya) in the Chapter: 1

C: Serial Number of Group A in the whole Quran

D: Serial Number of Group B and C whether from same A or not in the whole Quran

E: Serial Number of Groups B and C in same A. Group B is given Number 1 and C from 2 and above

F: Serial Number of multiwords inside Number 5 above

Note that:

A1, B1, C1, D1, E1, F1 are same as A, B, C1, D1, E1, F for the first multiword in the group of same roots

And: C1, D1 and E1 are common between RF and RPH groups

As an example, the following are the first few entries in the corpus

RF[1 : 1 : 2 : 2 : 1 : 1] RPH[1 : 1 : 2 : 2 : 1 : 1] PH [بِسْمِ اللَّهِ]RT[سمى ءله]

RF[11 : 41 : 2 : 2 : 1 : 2] RPH[1 : 1 : 2 : 2 : 1 : 1] PH [بِسْمِ اللَّهِ]RT[سمى ءله]

RF[27 : 30 : 2 : 2 : 1 : 3] RPH[1 : 1 : 2 : 2 : 1 : 1] PH [بِسْمِ اللَّهِ]RT[سمى ءله]

RF[5 : 4 : 2 : 3 : 2 : 1] RPH[1 : 1 : 2 : 3 : 2 : 1] PH [اسْمَ اللَّهِ]RT[سمى ءله]

RF[6 : 138 : 2 : 3 : 2 : 2] RPH[1 : 1 : 2 : 3 : 2 : 1] PH [اسْمَ اللَّهِ]RT[سمى ءله]

RF[22 : 28 : 2 : 3 : 2 : 3] RPH[1 : 1 : 2 : 3 : 2 : 1] PH [اسْمَ اللَّهِ]RT[سمى ءله]

RF[22 : 34 : 2 : 3 : 2 : 4] RPH[1 : 1 : 2 : 3 : 2 : 1] PH [اسْمَ اللَّهِ]RT[سمى ءله]

RF[22 : 36 : 2 : 3 : 2 : 5] RPH[1 : 1 : 2 : 3 : 2 : 1] PH [اسْمَ اللَّهِ]RT[سمى ءله]

RF[6 : 118 : 2 : 4 : 3 : 1] RPH[1 : 1 : 2 : 4 : 3 : 1] PH [اسْمُ اللَّهِ]RT[سمى ءله]

RF[6 : 119 : 2 : 4 : 3 : 2] RPH[1 : 1 : 2 : 4 : 3 : 1] PH [اسْمُ اللَّهِ]RT[سمى ءله]

RF[6 : 121 : 2 : 4 : 3 : 3] RPH[1 : 1 : 2 : 4 : 3 : 1] PH [اسْمُ اللَّهِ]RT[سمى ءله]

RF[22 : 40 : 2 : 4 : 3 : 4] RPH[1 : 1 : 2 : 4 : 3 : 1] PH [اسْمُ اللَّهِ]RT[سمى ءله]

A second example is given below in Figure (2).

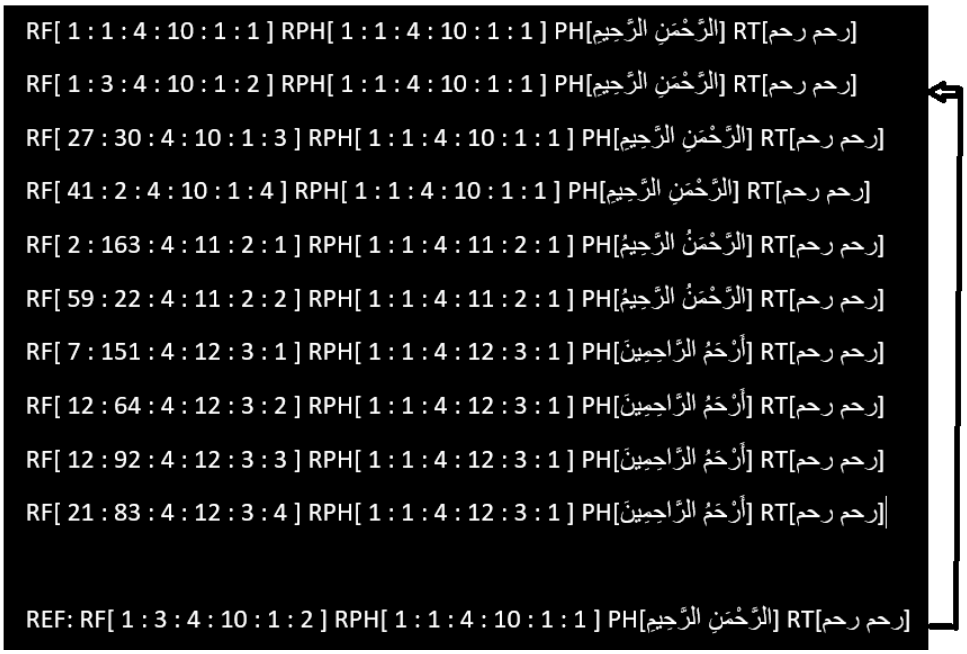

Figure 2: An example for multiword and its references.

The first line shows the following:

RF is for the information related to the multiword: [الرَّحْمَنِ الرَّحِيمِ] which is:

1: Chapter 1 (AL-Fatiha)

1: Verse No. 1

4: Group A related to different roots [رحم رحم]

10: Group B Multiword No. 10 among different multiwords of different words.

1: This is the first multiword in the whole group of same roots.

1: The first multiword in the group of exact words.

RPH[1 : 1 : 4 : 10 : 1 : 1] refers to the first multiword in the group of same roots. Here it is the same as RF

It is to be noted that the Group B related (No. 10 above) is repeated in RPH as in RF

PH[الرَّحْمَنِ الرَّحِيمِ]: The exact multiword with Tashkeel

RT[رحم رحم]: The corresponding series of roots related to this multiword.

In the second line:

RF[1 : 3 : 4 : 10 : 1 : 2] RPH[1 : 1 : 4 : 10 : 1 : 1] PH[الرَّحْمَنِ الرَّحِيمِ] RT[رحم رحم]

It is related to the same multiword which is in the third verse of Chapter 1. That was the only difference from the first shown line. Notice that the

RPH[1 : 1 : 4 : 10 : 1 : 1] is the same as the first multiword to indicate that it refers to the first line.

After many lines, when reaching multiwords of verse 3 of chapter 1:

REF: RF[1 : 3 : 4 : 10 : 1 : 2] RPH[1 : 1 : 4 : 10 : 1 : 1] PH[الرَّحْمَنِ الرَّحِيمِ] RT[رحم رحم]

This line refers to line 2 above. Which contains same information. This is to show when reaching this verse, one has to go back to verse 1 to find the accompanied multiword.

Table 1 show some statistics related to word and multiword in the corpus. Of course not all words or verses in the Quran appear in the corpus, since any word or phrase which appears once in the Quran was not included in the corpus.

Table 1 Some Statistics of words and multiword in the Holly Quran

Description	Number
Verses	6236
Total number of words	77479
No. of words in Othmani style	18841
No. of words in modern Arabic style	18232
No. of words with no shaddah on first letter	17884
No. of words with no shaddah on first letter and no Tashkeel at the end	15263
No. of Roots including assumed roots (e.g. Huroof) for words with no roots in Arabic Lexicons	1768
No. of multiword roots	18743
No. of multiwords	45637
No. of entries in the multiword corpus	110803

5. CONCLUSIONS

Multiword corpus is important for natural language processing specially for machine translation. This paper introduces the multiword corpus for the holy Quran based on the roots of the contained words. Multiwords were introduced in groups of the same roots. Subgroups in these groups may contain multiwords of the same roots but different words. Words with different Tashkeel are considered different words. Each entry of the corpus contains information about the entry itself and its reference entry which appears in the Quran for the first time. The corpus is available on: mw.quran.corpus.al-mishkat.com

6. REFERENCES

Abdelati Hawwari, Kfir Bar, Mona Diab, Building an Arabic Multiword Expressions Repository https://aclweb.org/anthology/W/W12/W12-3403.pdf

Alnaser, Mohammad, Multi-word Items in Dictionaries from a Translator's Perspective, Ph.D Thesis Durham University,2010

Annelen Brunner and Kathrin Steyer, Corpus-driven study of multi-word expressions based on collocations from a very large corpus, http://www1.ids-mannheim.de/fileadmin/lexik/uwv/dateien/BrunnerSteyerBirmingham2007.pdf

Boulaknadel, Siham, Beatrice Daille, Driss Aboutajdine, A multi-word term extraction program for Arabic language, LINA FRE CNRS 2729 University of Nantes, GSCM LRIT University Med V. 2 rue la Houssinire BP 92208 44322 Nantes, B.P. 2008.

Corpus 2011 http://corpus.quran.com

Douglas Biber, A corpus-driven approach to formulaic language in english: Multi-word patterns in speech and writing,

Dukes, Kais, Eric Atwell, Nizar Habash, Supervised Collaboration for Syntactic Annotation of Quranic Arabic, Language Resources and Evaluation Journal, 2011

Eric SanJuan and Fidelia Ibekwe-SanJuan, Multi Word Term queries for focused Information Retrieval, http://dev.termwatch.es/cv/pub/ibesan10.pdf

Greaves, Chris and Martin Warren, What can a corpus tell us about multi-word units? The Routledge Handbook of Corpus Linguistics, March 2010, ISBN: 9780415464895,

Khedher, 2001

محمد زكي خضر ، معجم كلمات القرآن – www.al-mishkat.com/words

Khedher, 2002

محمد زكي خضر ، المعجم المفهرس للتراكيب المتشابهة لفظًا في القرآن الكريم – دار عمار – الأردن 2002

Khedher, 2004

محمد زكي محمد خضر ، الجوانب البرمجية في إعداد المعجم المفهرس للتراكيب المتشابهة لفظًا في القرآن الكريم – دراسات – الجامعة الأردنية – المجلد 31 العدد 1, 2004

Klavans, Judith and Christian Jacquemin. 1997. A natural language approach to multi-word term conflation. In Proceedings, DELOS Workshop on Cross-Language Information Retrieval, ETHZ, Zurich. ERCIM: European Consortium for Informatics and Mathematics.

Lexemes Kis, Balázs; Villada, Begoña; Bouma, Gosse; Ugray, Gábor; Bíró, Tamás; Pohl, Gábor; Nerbonne, John Rijksuniversiteit Groningen, A New Approach to the Corpus-based Statistical Investigation of Hungarian Multi-word, Proceeding of the 4th International Conference on Language resources and Evaluation, Lisbon, Portogal, 2004

Liling Tan and Santanu Pal, Manawi: Using Multi-Word Expressions and Named Entities to Improve Machine Translation, http://statmt.org/wmt14/pdf/W14-3323.pdf

Meghawry, Samah, Abeer Elkorany, Akram Salah, and Tarek Elghazaly, Semantic Extraction of Arabic Multiword Expressions", Computer Science & Information Technology (CS & IT), 2015.

Mohammed Fouad, 2008

محمد فؤاد عبد الباقي: المعجم المفهرس لألفاظ القرآن الكريم- دار إحياء التراث العربي

Nathan Schneider, "Lexical Semantic Analysis in Natural Language Text", Carnegie Mellon University, PhD thesis, 2014. http://www.cs.cmu.edu/~nschneid/thesis/thesis-print.pdf

Peromingo, Juan Pedro Rica, Corpus analysis and phraseology: Transfer of multi-word units, LINGUISTICS AND THE HUMAN SCIENCES, VOL 6, NO 1-3 (2010)

Schneider, Nathan and Noah Smith, "A Corpus and Model Integrating Multiword Expressions and Supersenses", 2015. http://people.cs.georgetown.edu/nschneid/p/sst-slides.pdf

Yoon, H., & Hirvela, A. (2004). ESL Student Attitudes toward Corpus Use in L2 Writing. Journal Of Second Language Writing, 13(4), 257–283. Retrieved 21 March 2012.