Arab World English Journal
INTERNATIONAL PEER REVIEWED JOURNAL ISSN: 2229-9327
مجلة اللغة الانكليزية في العالم العربي

Arab World English Journal (AWEJ) Volume. 8 Number 4 December 2017 Pp 101-120
DOI: https://dx.doi.org/10.24093/awej/vol8no4.7

Morphological Analysis of the Glorious Qur'an: A Comparative Survey of Three Corpora

Yasser Muhammad Naguib Sabtan
Department of Languages and Translation
Dhofar University, Oman
&
Faculty of Languages and Translation
Al-Azhar University, Egypt

Abstract

Some attempts have been made in the academic community to carry out an automatic morphological analysis of the Qur'anic text. Among the well-known endeavors in this regard is the morphological annotation of the Quranic Arabic Corpus (QAC) which was carried out in Leeds University, UK. In addition, researchers in the University of Haifa had previously implemented a computational system for the morphological analysis of the Qur'an. More recently, a new Quranic corpus has been built in Mohammed I University in Morocco. To the best of our knowledge, these are the only three studies to produce a morphologically analyzed part-of-speech tagged Qur'an encoded as a structured linguistic database. This paper surveys the morphological analysis in the above-mentioned annotation projects and compares between them to test the quality of their analysis using five criteria related to display of the text in the corpus, word segmentation, morphological disambiguation, part of speech (POS) tag set and manual verification. The paper concludes that the QAC of Leeds and the Quranic corpus of Morocco surpass the Quranic corpus of Haifa with regard to most of these criteria. Furthermore, some additional POS tags for derivative nouns are suggested in a step to reach a more fine-grained tag set that could be proposed for POS tagging of Qur'anic Arabic.

Keywords: Arabic morphological analysis, Arabic POS tagging, corpus annotation, corpus linguistics, the Glorious Qur'an

Cite as: Sabtan, Y. M. N. (2017). Morphological Analysis of the Glorious Qur'an: A Comparative Survey of Three Corpora. *Arab World English Journal, 8* (4). DOI: https://dx.doi.org/10.24093/awej/vol8no4.7

1. Introduction

Arabic is known for its rich and complex morphology, where words are explicitly marked for case, gender, number, definiteness, mood, person, voice, tense and other features (Maamouri et al. 2006). The Arabic morphological system is generally considered to be of the non-concatenative (or non-linear) type where morphemes are not combined sequentially, but root letters are interdigitated with patterns to form stems. A root is a sequence of mostly three or four consonants which are called radicals. The pattern, on the other hand, is represented by inserting a template of vowels in the slot within the root's consonants. This combination of root, pattern and vocalism is normally referred to as templatic morphemes. Thus, an Arabic word is constructed by first creating a word stem from templatic morphemes to which affixational morphemes are then added (Habash, 2007). Thus, a word in Arabic may contain up to five morphemes (i.e. a stem with a number of concatenated affixes). All elements are optional except the stem. This morphological nature of Arabic will be made clear in section 2 below.

Arabic has a number of varieties that are spoken across the Arab world. Two main varieties are widely used among the Arab countries and are understood by all Arabs. The first one is Classical Arabic (CA), which is the language of the Qur'an and Sunna (prophetic traditions). CA is normally written with diacritic marks above the consonants. This was basically done to help people to read such Arabic texts perfectly. The second variety is Modern Standard Arabic (MSA), which is the contemporary language that is used in newspapers, magazines, academic books, novels, TV shows, etc. MSA is written without diacritics on the consonants. Besides these two main varieties, there are other varieties that are classified as colloquial language or dialects. These dialects differ from one country to another and even from one part to another inside the same country. In fact, the situation in the Arab world is one of diglossia (Farghaly & Shaalan, 2009, Mahmoud, 2013), where two or more varieties of the same language are used by a speech community and each variety is used for a specific purpose and in a distinct situation (Ferguson, 1959). Thus, CA is the language of religion (the Qur'an and Sunna) and is used by Arabic speakers in their daily prayers while MSA is used in formal settings such as the media, the news, and the classroom. As for the regional dialects, they are confined to every day communication and are first acquired by Arabic children.

The Qur'an, written in CA, consists of 114 *surahs* (roughly 'chapters'). These chapters comprise 6236 verses that contain roughly 77,800 words. The Qur'anic text is written with diacritic marks above the letters. This was basically done to help people to read it correctly. It should be noted that the Qur'an is a type of text that is difficult to compare with other forms of Arabic, since the vocabulary and the spelling differs from Modern Standard Arabic. In addition, the Qur'anic text is characterized by unique linguistic or rather rhetorical features, which should pose special interests and challenges for computational linguistics solutions (Sharaf & Atwell, 2009). The linguistic style of the Qur'an makes extensive use of many rhetorical devices such as foregrounding and backgrounding, grammatical shift, metaphors and figurative language, idiomatic expressions, culture-bound items, and lexically compressed items where lengthy details of semantic features are compressed and encapsulated in a single word (Abdul-Raof, 2001).

Arabic morphological analyzers aim to identify and separate affixes (prefixes and suffixes) and clitics from the surface word and recover the root or the stem that may have undergone morphophonemic changes (Farghaly, 2010). Those analyzers also specify the grammatical categories (parts of speech) of words. Arab grammarians traditionally classify Arabic words into three main grammatical categories, namely noun, verb and particle. These categories could be classified into further sub-classes which collectively cover the whole of the Arabic language (Haywood & Nahmad, 1965). In addition, words are then morphologically analyzed with regard to those linguistic features such as number (singular, dual or plural), gender (masculine or feminine), case (nominative, accusative or genitive), definiteness (definite or indefinite), …etc. Thus, morphological analyzers classify words with their part of speech (POS) along with their morpho-syntactic features.

In this paper the author reviews the morphological analysis of the Qur'anic words in three corpora: (i) the Quranic Arabic Corpus (QAC)[1] which is a collaborative web-based project carried out at Leeds University (ii) the Haifa Quranic Corpus (henceforth HQC)[2] which was conducted at the University of Haifa and (iii) the latest annotated Corpus of the Qur'an which was built at Mohammed I University in Morocco[3]. This corpus will be referred to as (MQC) which stands for "Mohammed I Qur'anic Corpus". Throughout the survey the strengths and weaknesses of the analysis in these corpora are discussed.

The remainder of this paper is organized as follows: in section two the complexity of Arabic morphology and the challenge of POS tagging are discussed. Section three sheds light on different computational efforts for the morphological analysis of the Glorious Qur'an. In section four a comparison is made between the three corpora: QAC, HQC and MQC. Section 5 shows the results of the comparison process. Finally, a conclusion of the paper is presented in section 6.

2. Arabic Morphological Analysis

Since Arabic is a highly inflected language with a complex morphological system, a word in Arabic may contain up to five parts as follows:

1. **Proclitics**, which occur at the beginning of a word, (e.g. conjunctions such as و "and", ف "then", prepositions such as ب "with" or "by", ل "to").
2. **Prefixes**, such as the prefix of the imperfective verb ي, the future marker س "will" and the definite article ال "the".
3. **A stem**, which can be represented in terms of a 'root' and a 'pattern', as described above.
4. **Suffixes**, such as verb endings, nominal cases, nominal feminine ending, plural markers …etc.
5. **Enclitics**, which occur at the end of a word, are complement pronouns.

For example, the Arabic word ليكتبونها *lyktbwnhA*[4] "to write it" contains the previous components as shown in table 1.

ISSN: 2229-9327

Table 1. ***An example showing Arabic word structure***

Proclitic	Prefix	Stem	Suffix	Enclitic
ل	ي	كتب	ون	ها

As the previous table shows, the Arabic word ليكتبونها *lyktbwnhA* contains a number of affixes and clitics that have corresponding words in English. This applies to both vowelized and non-vowelized words. It should be noted that all concatenations and inflections are optional except the stem which is the obligatory element.

Arabic morphological analysis is a tough and complicated process due to the complex nature of Arabic morphology. This complex nature is most vivid in such cases where a single Arabic word could stand as a complete sentence, particularly in Qur'anic Arabic. For instance, فأسقيناكموه *fa>asoqayonaAkumuwhu* "then we gave it to you to drink", which is composed of a stem along with a number of affixes and clitics, gives the meaning of a complete sentence

As pointed out earlier, one of the main functions of a morphological analyzer is to specify the POS category for each word. Generally, morphological analyzers are designed to generate all possible analyses of the analyzed words, indicating the potential POS categories for such words out of their context. But when context is taken into account, the process of determining POS categories is called POS tagging, word-class tagging or sometimes morpho-syntactic tagging. In this regard, a distinction is often made between morphological analysis problems (which are handled by a morphological analyzer) and morphological disambiguation problems (Habash et al. 2009). A number of POS taggers use morphological analyzers as one of their components. In other words, the morphological analyzer proposes a number of potential POS categories for input words and then the POS tagger chooses the right POS category for each word in its context. MADA[5] (Habash et al. 2009) is one of such taggers. Some other taggers do not use a morphological analyzer and use other techniques to carry out POS tagging. For example, Ramsay and Sabtan (2009) produced a lexicon-free maximum-likelihood tagger which makes use of very simple clues based on the initial and final characters of a word along with transition probabilities between tags, and then uses transformation-based learning (TBL) to patch the errors in this initial assignment. As regards the morphological analysis in the three corpora under study, the QAC and MQC provide morpho-syntactic tagging for each Qur'anic word in its contextual verse. But Haifa analyzer is incapable of performing context-dependent morphological disambiguation, and sometimes provides multiple analyses for each word, especially in case of verbs (Talmon & Wintner 2003; Dror et al. 2004).

Arabic POS tagging is not an easy task due to the complicated nature of Arabic word structure and the high degree of lexical ambiguity which is particularly pervasive in non-vowelized (or undiacritized) texts. This ambiguity occurs when a single written form may correspond to a number of different lexemes which may have a number of different senses as well as POS categories. In this regard, we will discuss two important reasons that represent a challenge for Arabic POS tagging.

1. Homographs: These are words that have the same orthographic form but different pronunciations and meanings (Jackson, 1988). This phenomenon is widespread in non-vowelized Arabic, as shown in the examples in table 2 below.
2. Internal word structure ambiguity: a complex Arabic word could be segmented in different ways (Farghaly & Shaalan, 2009). In such cases a POS tagger has to determine the boundaries between segments or tokens to give each token its proper POS tag. 'Segmentation' is a method to determine the boundaries between all the word parts. This word segmentation ambiguity is sometimes termed 'coincidental identity'. This occurs when clitics accidentally produce a word-form that is homographic with another full form word (Kamir et al., 2002; Attia, 2006). Examples for such cases are given in table 3.

Table 2. ***Arabic homographs***

Word	**Meanings**	**POS Category**	**Gloss**
ذهب *hb	ذَهَبَ *ahaba	verb	to go
	ذَهَبٌ *ahabN	noun	gold
قدم qdm	قَدِمَ qadima	verb	to arrive from
	قَدَّمَ qad~ama	verb	to introduce
	قَدَمْ qadamo	noun	foot

Table 3. ***Arabic words with different segmentations***

Complex Word	**Possible Tokens**			**POS Category**	**Gloss**
ولي wly	وَ *wa*	لِ *li*	ي *y*	conj. + prep. + pronoun	and for me
	وَلِي *waliy*			noun	a pious person favored by God
كمال kmAl	كَ *ka*	مَال *maAl*		prep. + noun	as money
	كَمَال *kamaAl*			noun	perfection
	كَمَال *kamaAl*			proper noun	a person's name

ISSN: 2229-9327

As noted earlier, the Qur'anic text is vowelized or diacritized where diacritics are placed on letters to indicate the pronunciation of words. The complicated nature of Arabic morphology is noticeable in both vowelized and non-vowelized texts. As for the degree of lexical ambiguity, which is so vivid in non-vowelized texts, it is also apparent in vowelized text but in a lesser degree.

3. Morphological Analysis of Qur'anic Arabic

Morphological analysis has been carried out for analyzing Classical Qur'anic Arabic. Some studies have focused on stemming the Qur'anic text to obtain the stem after removing all affixes and clitics. Thabet (2004) proposed a light stemming approach that uses a transliterated version of the Qur'an in western script. Thabet's main objective for stemming the Qur'an was to prepare the text as data for multivariate analysis of the lexical semantics of the Qur'an. In addition, Yusof et al. (2010) developed a rule-based stemming algorithm to stem the Qur'an through identifying the various word patterns. Their approach, which deals only with triliteral roots, was tested on the 30th chapter of the Glorious Qur'an. More recently, Sabtan (2012) presented a light stemmer for Arabic, using a corpus-based approach. The stemmer, which was tested on the Qur'anic text in its non-vowelized form, groups morphological variants of words in the corpus based on letter-sequence similarity, before stripping off their affixes to produce their common stem. The aim of developing such a stemmer was to investigate the effectiveness of using word stems for extracting bilingual equivalents from an Arabic-English parallel corpus. The Qur'anic Arabic text with an English translation was used as the parallel corpus. Nonetheless, all the previously mentioned attempts do not provide a morphologically analyzed part-of-speech tagged Qur'an encoded as a structured linguistic database.

Other research efforts have worked on providing a full morphological analysis of the Glorious Qur'an encoded as a structured linguistic database. Within this framework a study was conducted at the University of Haifa to present a computational system for morphological analysis and annotation of the Qur'an for research and teaching purposes. The system consists in a set of finite-state based rules using Finite State Machines technology to annotate the Arabic morphology of the Qur'an. However, the automatic annotation was not manually verified. The accuracy of the system is estimated at 86% (Dror et al. 2004).

In addition, the Qur'anic text has been linguistically annotated at Leeds University, UK. The Quranic Arabic Corpus (QAC) is a newly available linguistic resource enriched with multiple layers of analysis including morphological annotation and POS tagging, syntactic analysis using dependency grammar and a semantic ontology (Dukes & Habash, 2010; Dukes & Buckwalter, 2010; Dukes, 2013; Dukes et al. 2010, 2013). In this paper the author focuses only on the layer of morphological analysis. Other layers, i.e. syntactic and semantic analysis, are outside the scope of this paper. The motivation behind the QAC work is to produce a resource that enables researchers interested in the Qur'an to get as close as possible to the original Arabic text and understand its intended meanings through grammatical analysis. Buckwalter's Arabic Morphological Analyzer (BAMA) (Buckwalter, 2002) was used to generate the initial tagging in the morphological annotation of the Quranic Arabic Corpus. The analyzer was adapted to work with Quranic Arabic text. It was necessary to convert from MSA BAMA tag set to the desired Quranic tag set. Then, manual correction was carried out online

through collaborative annotation to produce a more reliable research resource (Dukes & Habash, 2010).

In a new attempt at Mohammed I University to produce a morphologically analyzed corpus of the Qur'anic text, Zeroual & Lakhouaja (2014) presented a new Quranic Corpus rich in morphological information. But this corpus is not yet available online. They used a semi-automatic technique, which consists in using the morphsyntactic analysis system of MSA words "AlKhalil Morpho System" followed by manual verification. Each word in this corpus is associated with the following morphological information: stem, POS tag, lemma, root and pattern. It is worth noting that a lemma is the uncliticized perfective third person masculine singular form in case of verbs. For nouns, it is the uncliticized singular indefinite masculine or feminine form (Saleh & Habash, 2009).

4. Comparative Description of the Corpora

A number of studies have focused on comparing between morphological analyzers from different aspects. For instance, Attia (2006) compares between BAMA (Buckwalter, 2002), Xerox Arabic Morphological Analysis and Generation (Beesley, 2001) and Attia's Arabic Morphological Transducer with respect to ambiguity. Sawalha & Atwell (2008) conduct a comparative evaluation of a number of morphological analyzers and stemmers to test their accuracy with regard to stemming or root extraction. In this paper the author focuses on another aspect in the morphological analysis process. In particular, he compares between three morphologically analyzed corpora of Qur'anic words with regard to the way the annotation is displayed, the actual morphological processing and the manual verification of the analysis. He also discusses the tag sets that are available in the QAC and MQC and proposes more tags that could be added so as to reach a more fine-grained tag set that could be used for tagging the Qur'an.

Hamada (2009) proposes a number of standards for evaluating Arabic morphological analyzers. These standards are concerned with the entire process of morphological analysis, including the ability to specify the root, affixes and POS category of a given word (whether it is vowelized or non-vowelized). Due to the fact that the tag sets in the three corpora are not the same, a large-scale automatic evaluation is not possible. In this regard, Hamada (2009) points out that the automatic evaluation of morphological analyzers requires a unified tag set. Therefore, the author conducted a small-scale manual evaluation experiment to discuss the differences between the three corpora under study with respect to their morphological analysis.

The HQC is automatically morphologically analyzed without any manual correction. According to Dror et al. (2004), the accuracy of the system scores 86%. The morphological annotation of the QAC, on the other hand, was manually verified. However, it should be noted that the manually built morphological analysis is not error-free. Dukes et al. (2013) point out that the current estimated accuracy of morphological annotation in the QAC is measured at 98.7%, using the approach of supervised collaboration. As for the MQC, Zeroual & Lakhouaja (2014) indicate that AlKhalil morphological analyzer, which was used to analyze the words in the corpus, has analyzed 94% of the Qur'anic words. Thus, they had to manually add the remaining 6% to the output results. In addition, some of the given results are not correct. They

point out that 22% of input words have multiple output analysis. So, the correct analysis for each word has been selected, if it exists. If no correct result exists, they have added them manually. As for those words that have one output analysis, but it is not correct in the context, they have added them manually. This is because AlKhalil system made morphological analysis out of context.

There are a number of differences between the three corpora (HQC, the QAC and MQC) with regard to the way the Qur'anic text is encoded in these corpora. The HQC does not use the standard Arabic transcript but uses a phonemic transcription of the text. The transcription is based on pure ASCII notations, largely with single-symbol equivalents of the Arabic graphemes, and double letters expressing long vowels. Also, hyphenation is used to isolate nominal and verbal bases from the various affixes, e.g. *wa-kaana* "and was" (Talmon & Wintner 2003; Dror et al. 2004).

The QAC, in contrast, uses the Arabic script along with a phonetic transcription, word-for-word translation and location reference based on (Chapter: Verse) standard besides three levels of analysis: morphological annotation, a syntactic treebank and a semantic ontology. Moreover, a single complex word in the QAC is divided into multiple morphological segments with a POS tag assigned to each segment. What is also more interesting is that the QAC linguistic annotation is color-coded and is thus easy to read. This, in turn, facilitates the deep understanding of the Glorious Qur'an. As a result, over a million visitors use the QAC website per year (Atwell, 2012).

As for MQC, the morphological annotation of this corpus is not yet available online and there are only sample examples cited by the authors in their paper (Zeroual & Lakhouaja, 2014), which will be used in the comparative analysis. The corpus contains the Arabic script along with Buckwalter transliteration. However, the current stage of the corpus does not contain word-for-word translation or location reference.

With regard to the QAC annotation the author only displays the morphological level of linguistic annotation, as other levels of linguistic analysis are outside the scope of this paper. Besides the color-coded linguistic annotation on the QAC website, (which shows not only morphological analysis but syntactic and semantic analysis as well), there is an available file on the website that contains only the morphological analysis of the QAC as a resource for linguistic investigation. The data in this file will be used in the discussion of the morphological analysis of the QAC.

It is time now to throw light on the morphological analysis in the three corpora under study. The differences between the morphological annotations in the three corpora are made clear through the Qur'anic word الحمد *AlHamodu* "praise".

First, figure 1 shows the morphological analysis of this word in the HQC. Then the morphological analysis of the same word in the QAC and MQC is illustrated in tables 4 and 5 respectively.

l-Hamd-u	Def+Hmd+fa&l+Noun+Masc+Sg+Nom

Figure 1. **Morphological analysis of a Qur'anic word in HQC**

The previous figure shows the morphological analysis of a Qur'anic word in the HQC. The analysis of a word contains a number of morphological components which are described as follows:

- The Qur'anic word: A hyphen is used in the word الحمد (*l-Hamd-u*) to isolate the nominal stem (*Hamd*) from the prefix (the definite article *l*) and the nominative case marker (*u*). It should be noted that the definite article *Al* is shortened to just *l* in pronunciation after being connected with the preceding word in the previous verse.
- The morphological analysis: The following figure shows the description of the morphological information in the HQC.

Def: definiteness	POS: Noun	Case: Nom (nominative)
Root: *Hmd*	Gender: Masc (masculine)	
Pattern: *fa&l*	Number: Sg (singular)	

Figure 2. **Abbreviated tags in the morphological analysis of a Qur'anic word in HQC**

The current example الحمد *AlHmd* "praise" is a singular masculine noun in the nominative case preceded by the definite article. The root of this noun is *Hmd* whose pattern is *fa&l* (i.e. فعل).

Table 4. ***Morphological analysis of a Qur'anic word in the QAC***

Location	**Word**	**Morphological Segments**	**Tag**	**Morphological Features**
(1:2:1)	ٱلْحَمْدُ AloHamodu	Alo	DET	PREFIX|Al+
		Hamodu	N	STEM|POS:N|LEM:Hamod| ROOT:Hmd|M|NOM

Table 4 shows the morphological analysis of a Qur'anic word in the Quranic Arabic Corpus (QAC). The word is first segmented by separating the definite article prefix [Al] that is tagged as [DET] (i.e. determiner) from the stem [Hamodu] which is then tagged as [N], i.e. noun. The POS tag is followed by the lemma [Hamod] and then the root [Hmd]. Finally, the morphological feature of gender [M], i.e. masculine as well as the case marking [Nom], i.e. nominative are shown at the end of the analysis.

Since the morphological analysis in the QAC is used in preparation for syntactic annotation, suffixes and enclitics that have syntactic functions within a word are annotated with their

ISSN: 2229-9327

grammatical relations. For instance, in figure 3 below the masculine plural suffix has the syntactic function of subject.

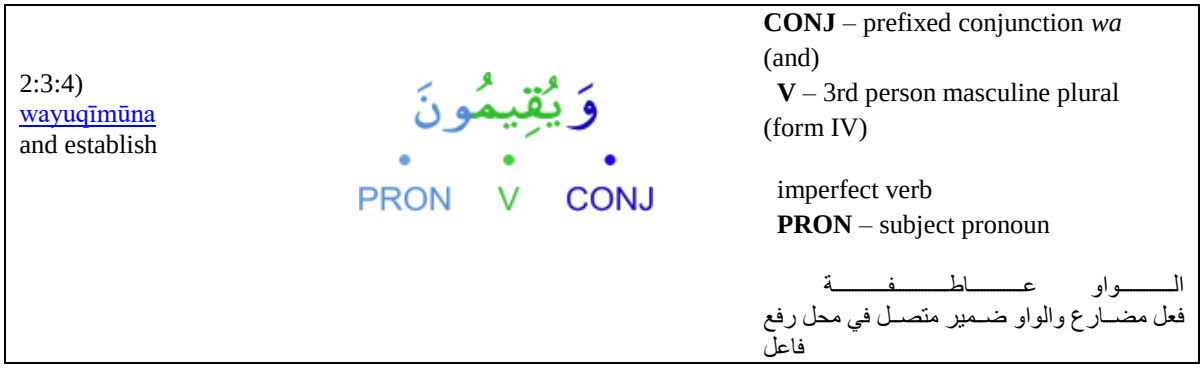

***Figure 3*. A tagged word in the QAC showing a syntactic function for the plural suffix**

The previous figure shows that the word is composed of the proclitic conjunction [و (*wa*)], the imperfect verb [يقيم (*yqym*)] and the 3rd person masculine plural suffix [ون (*wn*)]. Notably, the plural suffix functions as a subject pronoun.

As for the morphological analysis in MQC, the following table sheds light on the analysis of the example word.

Table 5. *Morphological analysis of a Qur'anic word in MQC*

Word	Stem	Stem Pattern	POS tag	Lemma	Lemma Pattern	Root
الْحَمْدُ AloHamodu	حمد Hmd	فَعْلُ faEolu	مأ VN	حَمْد Hamod	فَعْل faEol	حمد Hmd

As Table 5 shows, the example word in the MQC is associated with a number of morphological information: the stem, part-of-speech tag, lemma, root, and the vocalized patterns for each of the stem and lemma. The Arabic tag مأ is used to refer to the POS category مصدر أصلي and its English tag "VN" means 'Verbal Noun'. A sample of the tag set used in MQC is shown in Appendix C, based on Zeroual & Lakhouaja's paper (2014). It is obvious that the analysis does not contain information for the Arabic definite article "Al" which is a prefix attached to the beginning of the word. This has been observed in other nominal and verbal affixes. So, the MQC analyzes the main component of the word, i.e. the stem and leaves out affixes and clitics. However, the MQC, unlike the QAC and HQC, specifies the pattern of the stem and lemma.

5. Results and Discussion

The morphologically analyzed corpora of both the QAC and HQC are available online, but the MQC corpus is not yet available online and therefore the author could only discuss the examples cited by Zeroual and Lakhouaja (2014) in their paper. The author compares between the three corpora with regard to the following morphological information: stem, affixes, POS tags,

morphological features, lemma and root. Table 6 shows the morphological analysis of a number of words in the three corpora under investigation. The example words constitute verse 2 in Chapter 1 (*Sūrat Al-Fātiha* "The Opening").

الْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِينَ ﴿2﴾ Praise be to Allah, Lord of the Worlds.[6]

We will discuss the analysis for each corpus to uncover the differences between the three corpora. The 0 symbol is used when a certain morphological feature is not relevant for the word under analysis. But the # symbol is used for the non-existing information.

Table 6. ***Morphological analysis information in the three corpora***

Corpus	Word	Affix	POS Tag	Stem	POS Tag (with Morphological Features)	Lemma	Root
HQC (Haifa)	الْحَمْدُ *AloHamodu*	الْ *Alo*	Definite Article	حَمْدُ *Hamodu*	Noun + Masc. Sing. Nominative	#	حمد *Hmd*
	لِلَّهِ *lil~ahi*	لِ *li*	Preposition	لله *l~ahi*	Proper Name + Genitive	#	#
	رَبِّ *rab~i*	0	0	رَبِّ *rab~i*	Noun + First Person Masc. Sing. Dependent Pronoun	#	ربب *rbb*
	رَبِّ *rab~i*	0	0	رَبِّ *rab~i*	Noun + Masc. Sing. Genitive	#	ربب *rbb*
	الْعَالَمِينَ *AloEaAlamiyna*	الْ *Alo*	Definite Article	عَالَمِينَ *EaAlamiyna*	Noun + Masc. Plural	#	علم *Elm*
QAC (Leeds)	الْحَمْدُ *AloHamodu*	الْ *Alo*	Determiner	حَمْدُ *Hamodu*	Noun + Masc. Nominative	حَمْد *Hamod*	حمد *Hmd*
	لِلَّهِ *lil~ahi*	لِ *li*	Preposition	لله *l~ahi*	Proper Noun + Genitive	الله *All~ah*	الله *Alh*
	رَبِّ *rab~i*	0	0	رَبِّ *rab~i*	Noun + Masc. Genitive	رَب *rab~*	ربب *rbb*
	الْعَالَمِينَ *AloEaAlamiyna*	الْ *Alo*	Determiner	عَالَمِينَ *EaAlamiyna*	Noun + Masc. Plural Genitive	عَالَمِين *EaAlamiyn*	علم *Elm*
MQC (Morocco)	الْحَمْدُ *AloHamodu*	#	#	حمد *Hmd*	Verbal Noun مصدر أصلي	حَمْد *Hamod*	حمد *Hmd*

ISSN: 2229-9327

	لِلَّهِ *lil~ahi*	#	#	لِلَّه *l~ahi*	Noun اسم الجلالة	الله *All~ah*	الله *Alh*
	رَبِّ *rab~i*	0	0	رَبِّ *rab~i*	Noun اسم جامد	رَب *rab~*	ربب *rbb*
	الْعَالَمِينَ *AloEaAlamiyna*	#	#	عالمين *EaAlamiyn*	Noun اسم جامد	عَالَم *EaAlam*	علم *Elm*

It is obvious in table 6 that there are differences between the three corpora with respect to their morphological analysis. These differences can be shown as follows:

- While the QAC and MQC provide a unique analysis for each word, the HQC contains multiple analyses for many words, as the case with the word *rab~i* "Lord". It has two analyses: the first analysis refers to the basic stem "lord" with the addition of the first person singular masculine dependent pronoun "ي" meaning "my lord". The second analysis, however, refers to the stem only, i.e. "lord".
- Both the QAC and MQC provide the lemma and root as part of the morphological analysis. The HQC, on the other hand, does not provide the lemma for words. In case of roots, some words do not have the root as part of their analysis.
- Both the QAC and MQC segment the stem from other affixes. However, the QAC assigns POS tags for each segment, while the MQC assigns POS tags to stems only. As for the HQC, it provides POS tags for all components of the word without word segmentation as shown in figure 1 above.
- The QAC and HQC contain more information about the morphological features of gender, number, and case.
- The analysis in the HQC does not contain information about the lemmas of words, while lemmas are given in the QAC and MQC.
- The lemma of the word الْعَالَمِينَ *AloEaAlamiyna* "worlds" in the QAC analysis is given as *EaAlamiyna* which is not actually the lemma but the stem. According to Zeroual and Lakhouaja (2014), many lemmas in the QAC are in fact stems. The MQC, in contrast, provides the correct lemma of the word, which is عَالَم *EaAlam* "world".

Based on the previous discussions, the key differences between the three corpora can be illustrated using five main criteria, as shown in table 7.

Table 7. *Comparison of the Three Qur'anic Corpora*

Haifa Quranic Corpus (Haifa University, 2004)	
Display of the text	It uses only a phonemic transcription of the Arabic text which makes it difficult to search for a specific root.
Word segmentation	Words are POS-tagged without being tokenized into morphological segments.

Morphological disambiguation	It contains multiple analyses for many words.
POS tag set	The authors did not publish a well-defined annotation scheme, including the POS tag set that was used to annotate the corpus.
Manual verification	It was not manually verified and authors reported 86% accuracy.
The Quranic Arabic Corpus (Leeds University, 2010)	
Display of the text	It uses the Arabic script, a phonetic transcription, word-for-word translation and location reference.
Word segmentation	Words are divided into morphological segments with POS tags assigned to each segment.
Morphological disambiguation	It provides a unique analysis for each word in its contextual verse.
POS tag set	The QAC has a well-defined annotation scheme. The POS tag set and morphological feature tags are published in Dukes & Habash (2010).
Manual verification	It involved automatic annotation using BAMA followed by manual verification. 98.7% accuracy using the approach of supervised collaboration was reported.
Mohammed I Quranic Corpus (Mohammed I University in Morocco, 2014)	
Display of the text	It uses the Arabic script along with Buckwalter transliteration, but does not include word-for-word translation or location reference.
Word segmentation	It segments the stem from other affixes and clitics but assigns a POS tag for the stem only.
Morphological disambiguation	Initially 22% of input words have multiple analyses. Then, the correct analysis for each word has been selected.
POS tag set	It uses a fine-grained POS tag set. A sample of the tag set is described in the authors' paper about the corpus.
Manual verification	A semi-automatic method was used to annotate the Quranic text by means of using AlKhalil morphological analyzer followed by a manual treatment.

It is noticeable that the morphological analysis in the Quranic Arabic Corpus (QAC) and Mohammed I Quranic Corpus (MQC) has a number of advantages that are lacking in Haifa Quranic Corpus (HQC). These advantages are concerned with the points that have been discussed in the previous table.

Though the QAC employs a fine-grained POS tag set as shown in appendices A and B, it uses a less fine-grained tag set with regard to nouns. The morphological feature tags for derivative nouns include only three categories: active participle, passive participle and verbal noun.

Based on POS sub-classification of Arabic nouns, more derivative types could be added to the three derivative nouns in the QAC. Table 8 shows the tags for the three derivative nouns in the QAC along with some additional POS tags for other derivative nouns. It should be noted that some of these types are included in the MQC tag set as shown in Appendix C.

Table 8. ***Proposed tags for derivative nouns***

Tags	**Description**	**Arabic Equivalent**	**Example**
ACT PCPL	Active participle	اسم فاعل	مَالِك *maAlik* (owner)
PASS PCPL	Passive participle	اسم مفعول	مُطَهَّرة *muTah~arap* (purified)
VN	Verbal noun	مصدر أصلي	إيمَان <*iymaAn* (faith)
VNM	Verbal noun with initial mīm	مصدر ميمي	مَّغْفِرَة *magofirap* “forgiveness”
INSTM	Noun of instrument	اسم آلة	مِيزَان *miyzaAn* (balance)
DIM	Diminutive	اسم تصغير	بُنَيَّ *bunaY~a* (my son)
ELTV	Elative noun	اسم تفضيل	أَظْلَم >*aZolam* (more unjust)
TM	Noun of time	اسم زمان	مَوْعِد *mawoEid* (appointment)
PLC	Noun of place	اسم مكان	ٱلْمَشْرِق *Alma$oriq* “the east”
INSTN	Noun of instance	اسم مرة	رَجْفَة *r~ajofap* “earthquake”
MNR	Noun of manner	اسم هيئة	وِجْهَة *wijohap* "direction"
REL	Relative noun	اسم منسوب	عَرَبِيّ *Earabiy~* “Arab”
FEXG	Form of exaggeration	صيغة مبالغة	علّام *Eal~aAm* “All-Knower”

6. Conclusion

In this paper the author aimed to make a survey of the morphological analysis in three corpora of the Glorious Qur’an. These three annotation projects are the morphological tagging in Haifa Quranic Corpus (HQC) in 2004, the morphological annotation of the Quranic Arabic Corpus (QAC) in Leeds University (2010) and the newly constructed and morphologically analyzed Qur'anic corpus in Mohammed I University (MQC) in 2014. In the survey a comparative

evaluation of the three corpora was conducted with regard to five main criteria, namely display of the text in the corpus, word segmentation, morphological disambiguation, POS tag set and manual verification. The evaluation shows that the morphological analysis in both of the QAC and MQC has a number of advantages that are lacking in the HQC. Most importantly, the QAC and MQC provide a unique analysis for each word in its contextual verse. In addition, the automatic analysis has been manually verified in both corpora. The HQC, on the other hand, contains multiple analyses for many words and remains manually unverified. The QAC is advantageous in another aspect, namely the way the annotation of the text is displayed. The QAC uses the Arabic script along with a phonetic transcription, word-for-word translation and location reference. As for the POS tag sets, the HQC authors did not publish a well-defined scheme concerning the POS tag set that was used to annotate the corpus. As for the QAC and MQC, both corpora use a fine-grained POS tag set, though some noun tags are underspecified in the QAC. Underspecification means that the POS tag in question does not provide a full description of the mopho-syntactic features of a given word. In this regard, some nouns are not subcategorized into its derivative types. Therefore, additional tags for subcategories of nouns have been proposed to be used along with the existing tagsets for potential POS tagging of the Qur'anic text.

Notes

1. http://corpus.quran.com
2. http://cl.haifa.ac.il/projects/quran/index.shtml
3. This corpus has not been made available online till the time of writing this paper.
4. The author uses the standard Buckwalter transliteration scheme for converting Arabic script to the Roman alphabet.
5. MADA stands for "Morphological Analysis and Disambiguation for Arabic".
6. The translation of the verse is rendered by Pickthall, as shown on the Quranic Arabic Corpus website.

About the Author:

Dr. Yasser Sabtan earned his PhD in Computational Linguistics from the University of Manchester, UK in 2011. He is currently an Assistant Professor at the Department of Languages and Translation, Dhofar University, Oman. Dr. Sabtan is also affiliated to the Department of English, Faculty of Languages and Translation, Al-Azhar University, Egypt. His research interests focus on Arabic computational linguistics, corpus linguistics, machine translation, audiovisual translation and pragmatics.

References

Abdul-Raof, H. (2001). *Qur'an Translation: Discourse, Texture and Exegesis*. London and New York: Routledge.

Attia, M. A. (2006). An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks. In *Proceedings of the Challenge of Arabic for NLP/MT Conference*, October 2006. The British Computer Society, London, UK, pp. 48-67.

Atwell, E. (2012). Corpus resources for learning Arabic to understand the Quran. In *Higher Education Academy workshop on "The Role of Corpora in LSP (Language for Specific Purposes) Learning and Teaching"*, Parkinson B08, University of Leeds, UK.

Beesley, K. R. (2001). Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001. In *Proceedings of the ACL 2001 Workshop on Arabic Language Processing: Status and Prospects*, Toulouse, France.

Buckwalter, T. (2002). Buckwalter Arabic Morphological Analyzer Version 1.0. In *Linguistic Data Consortium Catalog number LDC2002L49, and ISBN 1-58563-257-0*.

Dror, J. Shaharabani, D., Talmon, R. & Wintner, S. (2004). Morphological Analysis of the Qur'an. In *Literary and Linguistic Computing*, 19 (4), 431-452.

Dukes, K. (2013). *Statistical Parsing by Machine Learning from a Classical Arabic Treebank*. Ph.D. Thesis, School of Computing, University of Leeds, UK.

Dukes, K. Atwell, E., & Habash, N. (2013). Supervised Collaboration for Syntactic Annotation of Quranic Arabic. In *Language Resources and Evaluation Journal (LREJ)*. Special Issue on Collaboratively Constructed Language Resources, pp. 1-30.

Dukes, K., Atwell, E., & Sharaf, A. (2010). Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank. In *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, pp. 1822-1827.

Dukes, K. & Buckwalter, T. (2010). A Dependency Treebank of the Quran using Traditional Arabic Grammar. In *Proceedings of the 7th International Conference on Informatics and Systems (INFOS 2010)*, Cairo, Egypt, pp. 1-7.

Dukes, K. & Habash, N. (2010). Morphological Annotation of Quranic Arabic. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, Valletta, Malta.

Farghaly, A. (2010). Arabic Machine Translation: A Developmental Perspective. In *International Journal on Information and Communication Technologies*, *3* (3), 3-10.

Farghaly, A. & Shaalan, K. (2009) Arabic Natural Language Processing: Challenges and Solutions. In *ACM Transactions on Asian Language Information Processing (TALIP)*, *8* (4), pp. 1-22.

Ferguson, C (1959). Diglossia. In *Word*, *15*, 325-340.

Habash, N. (2007). Arabic Morphological Representations for Machine Translation. In Soudi, A., Van den Bosch, A. and Neumann, G. (Eds.), *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, Springer, pp. 263-285.

Habash, N., Rambow, O., & Roth, R. (2009). MADA+TOKAN: a toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, pp. 102–109.

Hamada, S. (2009). مقترح لمعايير وضـوابط تقييم المحللات الصـرفية. [Proposed Criteria for Evaluation of Morphological Analyzers]. *In Proceedings of the 9th Conference on Language Engineering (ESOLEC'2009)*, pp. 101-124.

Haywood, J.A. & Nahmad, H.M. (1965). *A new Arabic grammar of the written language*. 2nd edn. London: Lund Humphries.

Jackson, H. (1988). *Words and their Meaning*. London: Longman.

Kamir, D., Soreq, N., & Neeman, Y. (2002). A Comprehensive NLP System for Modern Standard Arabic and Modern Hebrew. In *Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages*, Philadelphia, PA, USA, pp. 1-9.

Maamouri, M., Bies, A., & Kulick, S. (2006) Diacritization: A Challenge to Arabic Treebank Annotation and Parsing. In *Proceedings of the Challenge of Arabic for NLP/MT Conference*, The British Computer Society, London, UK, pp. 35-47.

Mahmoud, A. (2013). A linguistic perspective of the effect of English on MSA: Manifestations and ramifications. In *Journal of King Saud University – Languages and Translation*, *25*, 35–43.

Ramsay, A. & Sabtan, Y. (2009). Bootstrapping a lexicon-free tagger for Arabic. In *Proceedings of the 9th Conference on Language Engineering (ESOLEC 2009)*, Cairo, Egypt, pp. 202–215.

Sabtan, Y. (2012). Arabic Stemming: A Corpus-Based Approach. In *Proceedings of the 12th Conference on Language Engineering (ESOLEC'2012)*, Cairo, Egypt, pp. 92-101.

Saleh, I. M. & Habash, N. (2009) Automatic Extraction of Lemma-based Bilingual Dictionaries for Morphologically Rich Languages. In *Proceedings of the 3rd Workshop on Computational Approaches to Arabic Script-based languages at the MT Summit XII*, Ottawa, Ontario, Canada.

Sawalha, M. & Atwell, E. (2008). Comparative Evaluation of Arabic Language Morphological Analysers and Stemmers. In *Proceedings of COLING 2008 poster session*, Manchester, UK.

Sharaf, A. & Atwell, E. (2009). A Corpus-based Computational Model for Knowledge Representation of the Quran. In *Proceedings of CL2009 International Conference on Corpus Linguistics*, Liverpool, England.

Talmon, R. & Wintner, S. (2003). Morphological Tagging of the Qur'an. In *Proceedings of the Workshop on Finite-State Methods in Natural Language Processing, an EACL'03 Workshop*, Budapest, Hungary.

Thabet, N. (2004) Stemming the Qur'an. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-Based Languages*, COLING-04, Geneva, Switzerland, 2004, pp. 85-88.

Yusof, R., Zainuddin, R., Baba, M. S., & Yusoff, Z. M. (2010). Qur'anic Words Stemming. In *The Arabian Journal for Science and Engineering, 35* (2C), 37-49.

Zeroual, I. & Lakhouaja, A. (2014). A New Quranic Corpus rich in Morphological Information. In *Proceedings of 5th International Conference on Arabic Language Processing (CITALA 2014)*, November 26th – 27th 2014, Oujda, Morocco, pp. 134-138.

Appendices

Appendix A: Part-of-Speech tag set in the Quranic Arabic Corpus (QAC)

TAG	Description	Arabic Equivalent
N	Noun	اسم
PN	Proper noun	اسم علم
IMPN	Imperative verbal noun	اسم فعل أمر
PRON	Personal pronoun	ضمير
DEM	Demonstrative pronoun	اسم اشارة
REL	Relative pronoun	اسم موصول
ADJ	Adjective	صفة
NUM	Number	رقم
T	Time adverb	ظرف زمان
LOC	Location adverb	ظرف مكان

V	Verb	فعل
P	Preposition	حرف جر
EMPH	Emphatic lām prefix	لام التوكيد
IMPV	Imperative lām prefix	لام الأمر
PRP	Purpose lām prefix	لام التعليل
CONJ	Coordinating conjunction	حرف عطف
SUB	Subordinating conjunction	حرف مصدري
ACC	Accusative particle	حرف نصب
AMD	Amendment particle	حرف استدراك
ANS	Answer particle	حرف جواب
AVR	Aversion particle	حرف ردع
CAUS	Particle of cause	حرف سببية
CERT	Particle of certainty	حرف تحقيق
COND	Conditional particle	حرف شرط
EQ	Equalization particle	حرف تسوية
EXH	Exhortation particle	حرف تحضيض
ACC	Accusative particle	حرف نصب
EXL	Explanation particle	حرف تفصيل
EXP	Exceptive particle	أداة استثناء
FUT	Future particle	حرف استقبال
INC	Inceptive particle	حرف ابتداء
INTG	Interrogative particle	حرف استفهام
NEG	Negative particle	حرف نفي
PREV	Preventive particle	حرف كاف
PRO	Prohibition particle	حرف نهي
REM	Resumption particle	حرف استئنافية
RES	Restriction particle	أداة حصر
RET	Retraction particle	حرف اضراب
SUP	Supplemental particle	حرف زائد
SUR	Surprise particle	حرف فجاءة
VOC	Vocative particle	حرف نداء
INL	Quranic initials	حروف مقطعة

Appendix B: Morphological feature tags in the Quranic Arabic Corpus (QAC)

Features	**Tags / Descriptions**
prefix features	**Al+** (determiner *Al*) **bi+** (preposition *bi*) **ka+** (preposition *ka*) **ta+** (preposition *ta*) **sa+** (future particle *sa*) **yā+** (vocative particle *yā*) **hā+** (vocative particle *hā*)

letter *alif* as a prefixed particle	**A: INTG+** (interrogative *alif*) **A: EQ+** (equalization *alif*)
letter *wāw* as a prefixed particle	**wa+** (conjunction *wāw*) **w:P+** (preposition *wāw* – used as a particle of oath)
letter *fa* as a prefixed particle	**f:CONJ+** (conjunction *fa*) **f:REM+** (resumption *fa*) **f:CAUS+** (cause *fa*)
letter *lām* as a prefixed particle	**l:P+** (preposition *lām*) **l:EMPH+** (emphasis *lām*) **l:PRP+** (purpose *lām*) **l:IMPV+** (imperative *lām*)
root	**ROOT**: (uses Buckwalter transliteration)
lemma	**LEM**: (uses Buckwalter transliteration)
special	**SP**: (used if the word belongs to a special group such as (كـان وأخواتـهـا). Certain words in the corpus are tagged this way where this is relevant for syntactic function, and not easily determined by lemma or part-of-speech; for example, the particle *mā* (ما) in a negative sense can behave like the verb *laysa* (ليس) and place a predicate into the accusative case)
person	**1** (first person), **2** (second person), **3** (third person)
gender	**M** (masculine), **F** (feminine)
number	**S** (singular), **D** (dual), **P** (plural)
aspect	**PERF** (perfect), **IMPF** (imperfect), **IMPV** (imperative)
mood	**IND** (indicative), **SUBJ** (subjunctive), **JUS** (jussive), **ENG** (energetic)
voice	**ACT** (active), **PASS** (passive)
verbal form	**I** to **XII**
derivation	**ACT PCPL** (active participle) **PASS PCPL** (passive participle) **VN** (verbal noun)
state	**DEF** (definite) **INDEF** (indefinite)
case	**NOM** (nominative) **ACC** (accusative) **GEN** (genitive)
suffix features	**PRON**: (attached pronoun, compound feature with person, gender and number) **+VOC** (vocative suffix for *Allāhumma* (اللهم)

Appendix C: Sample of Part-of-Speech tags in Mohammed I University's Quranic Corpus (MQC)

Category	Tag	English Equivalent
Particles	حرف جر	Preposition
	أداة استفهام	Interrogative particle
	أداة استِثناء	Exceptive particle
	أداة نَفي	Negative particle
	أداة شرط	Conditional particle
	حَرف ابتِداء	Inceptive particle
Nouns	اسم علم	Proper noun
	صفة مشبهة	Adjective
	اسم تفضيل	Elative noun
	اسم فاعل	Active participle
	مصدر أصلي	Gerund / Verbal noun
	مصدر ميمي	Gerund/ verbal noun with initial mīm
	اسم آلة	Instrumental noun
Verbs	فعل ماض مبني للمعلوم	Perfect verb (Active voice)
	فعل ماض مبني للمجهول	Perfect verb (Passive voice)
	فعل مضارع مبني للمعلوم	Imperfect verb (Active voice)
	فعل مضارع مبني للمجهول	Imperfect verb (Passive voice)
	فعل أمر	Imperative verb
	فِعل ما ض جامِد	Perfect verb non-conjugated