Morphological Annotation of Quranic Arabic

Kais Dukes[1] and Nizar Habash[2]

[1] School of Computing, University of Leeds, LS2 9JT, United Kingdom
[2] Center for Computational Learning Systems, Columbia University, New York, USA
E-mail: sckd@leeds.ac.uk, habash@ccls.columbia.edu

Abstract

The Quranic Arabic Corpus (http://corpus.quran.com) is an annotated linguistic resource with multiple layers of annotation including morphological segmentation, part-of-speech tagging, and syntactic analysis using dependency grammar. The motivation behind this work is to produce a resource that enables further analysis of the Quran, the 1,400 year old central religious text of Islam. This paper describes a new approach to morphological annotation of Quranic Arabic, a genre difficult to compare with other forms of Arabic. Processing Quranic Arabic is a unique challenge from a computational point of view, since the vocabulary and spelling differ from Modern Standard Arabic. The Quranic Arabic Corpus differs from other Arabic computational resources in adopting a tagset that closely follows traditional Arabic grammar. We made this decision in order to leverage a large body of existing historical grammatical analysis, and to encourage online collaborative annotation. In this paper, we discuss how the unique challenge of morphological annotation of Quranic Arabic is solved using a multi-stage approach. The different stages include automatic morphological tagging using diacritic edit-distance, two-pass manual verification, and online collaborative annotation. This process is evaluated to validate the appropriateness of the chosen methodology.

1. Introduction

The Quranic Arabic Corpus (http://corpus.quran.com) is an on-line annotated linguistic resource with multiple layers of annotation including morphological segmentation, part-of-speech tagging, syntactic analysis using dependency grammar (إعراب القرآن الكريم) and a semantic ontology. The motivation behind this work is to produce a resource that enables further analysis of the Quran, the 1,400 year old central religious text of Islam. The 77,430 words of the Quran form a distinct genre difficult to compare to other texts of Arabic. Processing Quranic Arabic is a unique challenge from a computational point of view, since it differs significantly from Modern Standard Arabic (MSA).

In this paper, we focus on the morphological annotation in the Quranic Arabic Corpus. We describe a new multi-stage approach to this component. Given the importance of the Quran, special care has been taken to ensure a high level of accuracy for the final part-of-speech tagging and morphological annotation. An initial tagging was performed using the Buckwalter Arabic Morphological Analyzer (Buckwalter, 2002), which was adapted to work with Quranic Arabic. This initial stage of annotation was reasonably accurate, but to produce a more reliable research resource, manual correction was required. The annotated corpus was then put online to allow for collaborative annotation. This proved quite effective - over an initial period of six months, 2,000 words (2.6%) were revised as a result of supervised volunteer correction. Website visitors have since grown steadily to 1,500 users per day, while the number of online corrections has reduced over time. This would suggest the current part-of-speech tagged Quran has become more stable and that the corpus annotation is progressing towards further accuracy.

The popular online site of the Quranic Arabic Corpus has been used by numerous Quranic scholars, language researchers and students of Arabic, many of whom have shown particular interest in the morphological annotation described in this paper. Specifically, part-of-speech tags and inflectional features have been found to be a useful aid to students of the Quran, who wish to get as close as possible to the original Arabic text and understand its intended meanings through grammatical analysis.

This paper is organized as follows: Section 2 discusses the challenges of morphological annotation and related work; Sections 3 and 4 outline the tagging scheme and annotation process; Section 5 evaluates the chosen methodology; and Section 6 concludes.

2. Morphological Annotation of Arabic

2.1 Processing Quranic Arabic

Processing Quranic Arabic is a unique challenge from a computational point of view, since the vocabulary and spelling differ from MSA. However, Quranic Arabic comes with the advantage of being fully diacritized, unlike most other Arabic texts. Each word of the Quran contains detailed diacritics (marks) over all letters describing its exact vowelization (see Figure 1). Using this information gives an advantage to automatic annotation when compared to other forms of Arabic.

Figure 1 below shows an example word in the Quranic Arabic Corpus, as displayed to website users viewing the morphological annotation. The three numbers at the top of the figure give the chapter number, verse number and word number. The Quran is split into 114 chapters. Each chapter contains a sequence of numbered verses. In this example, the annotation corresponds to the fourth word of chapter 21, verse 70. The next line in Figure 1 is a segment-for-segment interlinear translation, followed by a full translation and a phonetic transcription. The pronunciation shown is derived automatically from the morphological annotation and diacritics already present in the text.

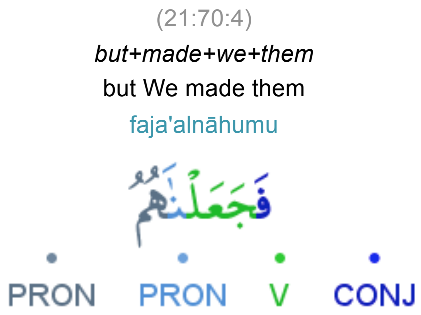


Figure 1: Morphological segmentation of a fully diacritized Arabic word in the Quranic Arabic Corpus

Arabic is read from right to left. In Figure 1, the word is split into four morphological segments: a prefixing conjunction, the main stem (a verb), and two suffixes - an attached subject pronoun and an attached object pronoun. This is typical of Quranic Arabic, where most white-space delimited words are in fact composed of multiple fused morphological segments, with prefixes and suffixes attached to a stem (Jones, 2005). Morphological annotation involves segmenting each word, and assigning a part-of-speech tag (e.g. noun, verb, preposition or pronoun) and inflectional features (such as person, gender and number) to each segment.

2.2 Annotated Arabic Corpora

Related annotated Arabic corpora may be compared to the Quranic Arabic Corpus in terms of their size (number of words before morphological segmentation), the depth of annotation provided, and the source of the original text. Another dimension of critical importance is whether automatically generated annotations were manually verified or not. Human manual verification provides a level of confidence in the annotated data when used for further research, or in the case of the Quranic Corpus, when used for educational purposes.

Figure 2 provides a comparison summary of related annotated Arabic corpora. The first three corpora are the Penn Arabic Treebank (PATB) (Maamouri et al., 2004), the Prague Arabic Dependency Treebank (PADT) (Smrž and Hajič, 2006) and the Columbia Arabic Treebank (CATiB) (Habash and Roth, 2009). All three corpora contain both morphological annotation of individual words, and syntactic parses of sentences in either constituency phrase structure grammar or dependency grammar. They are similar in that they consist of MSA text sampled mostly from newswire articles. The KDATD corpus developed at King Abdulaziz City for Science and Technology enriches sampled text by manually adding diacritics (Elshafei, 2005). The fifth annotation project listed is a study conducted at the University of Haifa. It is the most similar to the Quranic Arabic Corpus, given that it also attempts to provide a morphological analysis of the Quran. However, their automatic annotation was not manually verified (Dror et al, 2004). We compare our resource to theirs in further detail in the next section.

Corpus	Words	Annotation	Verified
PATB 1, 2, 3	613K	constituency	Y
PADT 1.0	114k	dependency	Y
CATiB 1.0	228K	dependency	Y
KDATD	-	diacritics	Y
Haifa	77k	morphological	N
Quranic*	77k	morphological	Y

Figure 2: Comparison of related annotated Arabic corpora to the Quranic Arabic Corpus (Quranic*)

Each of the corpora listed above differs by the type of annotation it adds to the original text. The aim of the PATB, is to produce a 1 million word annotated corpus, consisting of parsed constituency syntax trees. The first 3 released parts add to a total of 613K words before clitic segmentation, and are sampled from MSA newswire text (Maamouri et al., 2004). This corpus contrasts with the PADT. Although the text is also newswire articles, the syntactic analysis shows functional relations between words using dependency grammar (Smrž and Hajič, 2006). CATiB avoids annotating linguistic information that can be determined automatically, and follows a linguistic representation inspired by traditional Arabic grammar (Habash and Roth, 2009). Both PADT and CATiB include additional trees (not reported above) that are automatically converted from the PATB.

2.3 Previous Annotation of the Quran

The Quranic Arabic Corpus can be compared directly with the morphological analysis of the Quran conducted at the University of Haifa. To the best of our knowledge, this is the only other study to produce a morphologically analyzed part-of-speech tagged Quran encoded as a structured linguistic database. As the authors of the Haifa study note, computer analysis of the Quran is an intriguing but largely unexplored field. The Quranic Arabic Corpus differs from the Haifa effort in two important respects: (a) higher accuracy through manual verification, and (b) a more easily accessible annotation style that adopts traditional Arabic grammar notations.

The annotation in the Haifa Quranic Corpus was produced automatically using a rule-based morphological tagger, following a generative approach. Using a published concordance of the Quran, a list of base word forms was selected. An inflectional generator was then encoded into a finite state machine (FSM), consisting of 50 morphological rules for nouns, and 300 rules for verbs (Wintner, 2008). To perform the final annotation, the list of base word forms was fed into the FSM, which generated all possible conjugations and declensions using the predefined morphological rules. The large generated set of forms was then intersected with the set of original words in the Quran. The published dataset which resulted from the study is a list of possible analyses for each Arabic word in the Quran. The authors of the study report that about 70% of the words in the Quran received a unique morphological analysis, with the remaining words having several possible analyses each. Although full manual verification of the annotations did not take place, the authors used a representative sample to estimate the overall accuracy of the annotations in the corpus at an F-measure of 86% (Dror et al, 2004).

3. The Quranic Arabic Corpus Tagset

The Quranic Arabic Corpus differs from other related annotated Arabic corpora by adopting historical traditional Arabic grammar. Known as *i'rāb* (إعراب), this standardized grammar of the Quran has been developed and documented in detail for over 1,000 years – far longer than corresponding grammars for most other languages (Akesson, 2001; Fischer and Rodgers 2002; Haywood and Nahmad, 2005). In fact, traditional Arabic grammar is widely recognized as one of the origins of modern dependency grammar (Kruijff, 2006; Owens, 1988). Adopting this approach leads to morphological annotation which uses familiar terminology, and enables anyone who is already experienced with Quranic syntax to immediately participate in the online annotation effort. Using traditional grammar along with its standardized terminology also enables the morphological annotation to be verified against the many existing books and publications on Quranic grammar (Muhammad 2007; Nadwi 2006; Omar 2005; Rafai 1998; Siddiqui 2008). Traditional Arabic grammar defines a detailed part-of-speech hierarchy which applies to both words and morphological segments. Fundamentally, a word may be classified as a verb, nominal, or a particle. The set of nominals include nouns, proper nouns, adjectives, subject pronouns and object pronouns. The particles include prepositions, conjunctions and interrogatives, as well as many others.

In the Quranic Arabic Corpus, initial automatic tagging was carried out using a modified version of the Buckwalter Arabic Morphological Analyzer (BAMA), adapted to the unique language of the Quran (details in the next section). This was then followed by several stages of manual correction. BAMA defines its own tagset and segmentation scheme suitable for MSA (Buckwalter 2002). Since BAMA was used to perform the initial automatic analysis in the corpus, a mapping was required to convert to the desired Quranic tagset. For the vast majority of words, this was a one-to-one process. However, in few cases, the Quranic tagset was more detailed. For these words (such as the several types of particles), manual disambiguation was required.

Cat*	Tag	Arabic	Description
1	N	اسم	Noun
	PN	اسم علم	Proper noun
	IMPN	اسم فعل أمر	Imperative verbal noun
2	PRON	ضمير	Personal pronoun
	DEM	اسم اشارة	Demonstrative pronoun
	REL	اسم موصول	Relative pronoun
3	ADJ	صفة	Adjective
	NUM	رقم	Number
4	T	ظرف زمان	Time adverb
	LOC	ظرف مكان	Location adverb
5	V	فعل	Verb
6	P	حرف جر	Preposition
7	EMPH	لام التوكيد	Emphatic *lām* prefix
	IMPV	لام الامر	Imperative *lām* prefix
	PRP	لام التعليل	Purpose *lām* prefix
8	CONJ	حرف عطف	Coordinating conjunction
	SUB	حرف مصدري	Subordinating conjunction
9	ACC	حرف نصب	Accusative particle
	AMD	حرف استدراك	Amendment particle
	ANS	حرف جواب	Answer particle
	AVR	حرف ردع	Aversion particle
	CAUS	حرف سببية	Particle of cause
	CERT	حرف تحقيق	Particle of certainty
	COND	حرف شرط	Conditional particle
	EQ	حرف تسوية	Equalization particle
	EXH	حرف تحضيض	Exhortation particle
	EXL	حرف تفصيل	Explanation particle
	EXP	أداة استثناء	Exceptive particle
	FUT	حرف استقبال	Future particle
	INC	حرف ابتداء	Inceptive particle
	INTG	حرف استفهام	Interrogative particle
	NEG	حرف نفي	Negative particle
	PREV	حرف كاف	Preventive particle
	PRO	حرف نهي	Prohibition particle
	REM	حرف استئنافية	Resumption particle
	RES	أداة حصر	Restriction particle
	RET	حرف اضراب	Retraction particle
	SUP	حرف زائد	Supplemental particle
	SUR	حرف فجاءة	Surprise particle
	VOC	حرف نداء	Vocative particle
10	INL	حروف مقطعة	Quranic initials

***Categories**: 1=Nouns, 2=Pronouns, 3=Nominals, 4=Adverbs, 5=Verbs, 6=Prepositions, 7=*lām* prefixes, 8=Conjunctions, 9=Particles, 10=Disconnected letters.

Figure 3: Part-of-speech Tagset

Features	Tags / Descriptions
prefix features	**Al+** (determiner *al*) **bi+** (preposition *bi*) **ka+** (preposition *ka*) **ta+** (preposition *ta*) **sa+** (future particle *sa*) **ya+** (vocative particle *yā*) **ha+** (vocative particle *ha*)
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 (**كان واخواتها**). 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)
verb 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*)

Figure 4: Morphological Feature Tags

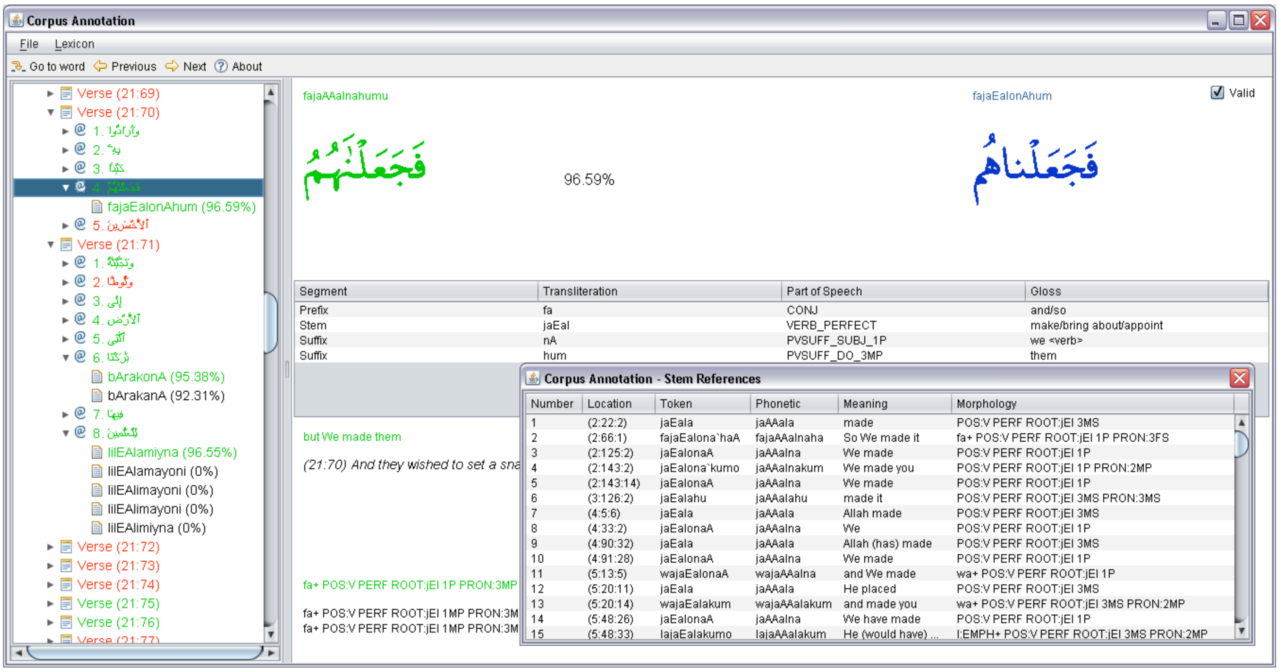
Figure 3 shows the part-of-speech tags used for morphological annotation of the Quran. Note that unlike a language such as English, where a single part-of-speech tag is typically assigned to each word, in a morphologically rich and highly cliticizing language such as Arabic, tags are applied to individual morphological segments (Habash, 2007). One white-space delimited word may consist of several segments in the corpus (a stem, together with fused suffixes and prefixes). In addition to part-of-speech tags, each morphological segment is annotated using a feature-value matrix of inflectional attributes (Figure 4) (Habash, 2007). These include person (first, second or third), number (singular, dual or plural) and gender (masculine or feminine). Specific features also apply to verbs and nouns. Verbs can be conjugated according to different aspects (perfect, imperfect and imperative) as well as moods of the imperfect (indicative, subjunctive, jussive and energetic) (Ryding, 2008). Nouns inflect for different grammatical cases (nominative, accusative and genitive) and may be definite or indefinite. According to traditional grammar, nouns may be derived from verbs as with the active participle, passive participle and the verbal noun. In the Quranic Arabic Corpus, this derivational morphology is manually tagged using a feature named "derivation type" (Wightwick and Gaafar 2008).

Quranic Arabic also requires some genre-specific tags. These include tags for Quranic initials which are sequences of disconnected letters (such as *alif lām mīm*) that occur at the start of certain chapters. There are numerous suggestions as to the mystical meaning of these letters and so they are given their own part-of-speech tag. In addition, the Quran contains a unique vocative suffix not found in MSA. The online corpus annotation guidelines provide detailed documentation for the tags and inflectional features.

4. Annotating the Quran

4.1 Automatic Annotation

Morphological analysis of the Quran involved automatic annotation using BAMA followed by manual verification. Developing such a rich set of annotated data would not have been possible without leveraging the existing tools and resources used to construct other corpora. It is useful to compare how the three largest Arabic treebanks were annotated. The PATB uses BAMA to provide morphological annotation. BAMA uses a comprehensive morphological lexicon of MSA, which groups conjugated word-forms together by lemma. Given an input word, the BAMA algorithm will suggest multiple possible morphological analyses. During the annotation of the PATB, annotators would use BAMA to select the correct analysis for each word (Maamouri et al., 2004). The PADT uses its own morphological analyzer, Elixir-FM, although it should be noted that this analyzer builds on the existing BAMA lexicon (Smrž and Hajič, 2006).



Segment	Transliteration	Part of Speech	Gloss
Prefix	fa	CONJ	and/so
Stem	jaEal	VERB_PERFECT	make/bring about/appoint
Suffix	nA	PVSUFF_SUBJ_1P	we <verb>
Suffix	hum	PVSUFF_DO_3MP	them

Number	Location	Token	Phonetic	Meaning	Morphology
1	(2:22:2)	jaEala	jaAAala	made	POS:V PERF ROOT:jEl 3MS
2	(2:66:1)	fajaEalona`haA	fajaAAalnaha	So We made it	fa+ POS:V PERF ROOT:jEl 1P PRON:3FS
3	(2:125:2)	jaEalonaA	jaAAalna	We made	POS:V PERF ROOT:jEl 1P
4	(2:143:2)	jaEalona`kumo	jaAAalnakum	We made you	POS:V PERF ROOT:jEl 1P PRON:2MP
5	(2:143:14)	jaEalonaA	jaAAalna	We made	POS:V PERF ROOT:jEl 1P
6	(3:126:2)	jaEalahu	jaAAalahu	made it	POS:V PERF ROOT:jEl 3MS PRON:3MS
7	(4:5:6)	jaEala	jaAAala	Allah made	POS:V PERF ROOT:jEl 3MS
8	(4:33:2)	jaEalonaA	jaAAalna	We	POS:V PERF ROOT:jEl 1P
9	(4:90:32)	jaEala	jaAAala	Allah (has) made	POS:V PERF ROOT:jEl 3MS
10	(4:91:28)	jaEalonaA	jaAAalna	We made	POS:V PERF ROOT:jEl 1P
11	(5:13:5)	wajaEalonaA	wajaAAalna	and We made	wa+ POS:V PERF ROOT:jEl 1P
12	(5:20:11)	jaEala	jaAAala	He placed	POS:V PERF ROOT:jEl 3MS
13	(5:20:14)	wajaEalakum	wajaAAalakum	and made you	wa+ POS:V PERF ROOT:jEl 3MS PRON:2MP
14	(5:48:26)	jaEalonaA	jaAAalna	We have made	POS:V PERF ROOT:jEl 1P
15	(5:48:33)	lajaEalakumo	lajaAAalakum	He (would have) ...	l:EMPH+ POS:V PERF ROOT:jEl 3MS PRON:2MP

Figure 5: Custom Java Application used for Manual Annotation in the Quranic Arabic Corpus

Similarly, CATiB uses its own analyzer and segmentation tools (MADA+TOKAN), which use the BAMA lexicon, but have a different analysis engine (Habash and Rambow, 2005; Habash, 2007).

Given the success of applying the BAMA lexicon to existing Arabic corpora, we decided to use BAMA for the morphological analysis of the Quranic Arabic Corpus, although we recognize that the tool would require significant adaptation given that the Quran is written in classical Arabic, as opposed to MSA. Another point is that the BAMA analyzer is in the public domain. It was essential that any tools used to produce the Quranic Arabic Corpus do not have any copyright concerns on the resulting data, since we wanted the corpus to be freely available.

BAMA uses its own detailed lexicon of Arabic to identify possible choices for segmentation and tagging of each word. A custom annotation tool was used to display the closest matching word in the lexicon, and allowed annotators to select or override from a list of possible diacritized analyses. For other Arabic corpora, the analyzer is typically run against text without diacritics. In the case of the Quran, the text is already diacritized and this additional information was leveraged to develop a modified analyzer.

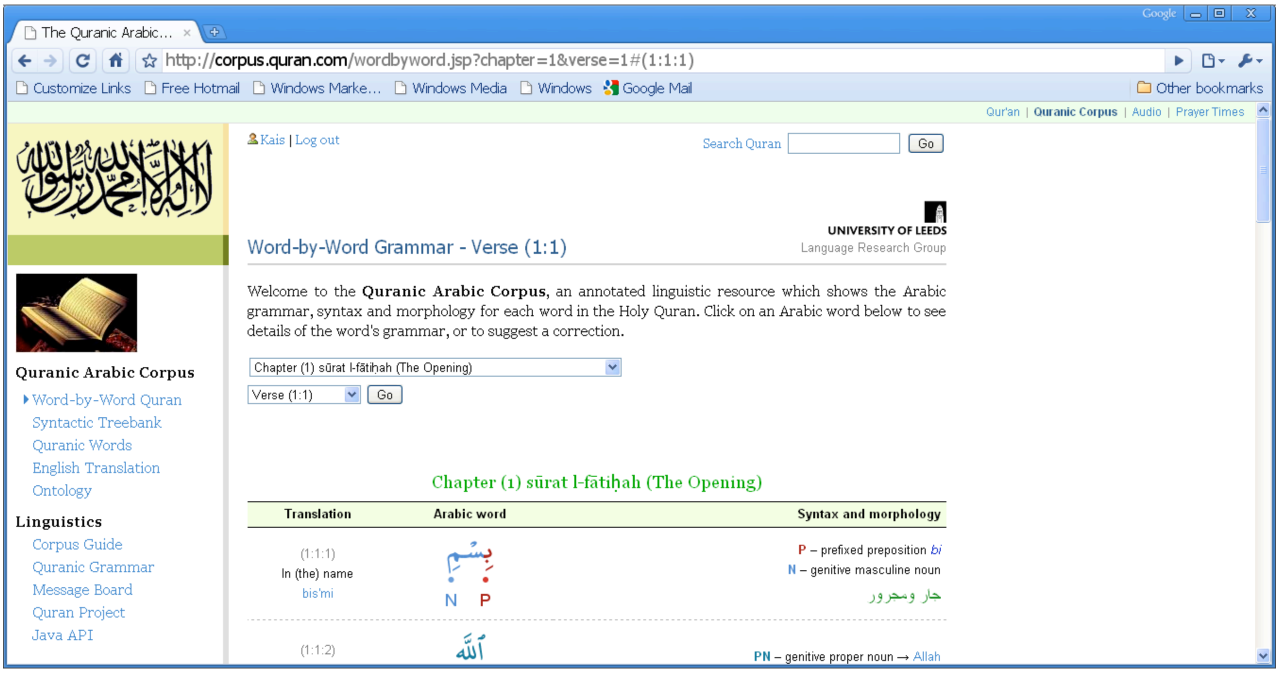
Three extensions were made to BAMA in order to allow processing of Quranic Arabic: spelling, ranking and filtering. Running an unmodified analyzer against the Quran produces low accuracy for part-of-speech tagging because the spelling of the Quran differs from MSA. Most of the differences involve orthographic variation of the Arabic *hamza* (glottal stop) and the dagger *alif* (a diacritic used for the long vowel ā). BAMA was extended to account for these differences. The different diacritized analyses are ranked in terms of their edit-distance from the Quranic diacritization, with the closer matches ranked higher. The BAMA analysis with the highest rank is then chosen as the unique part-of-speech for that word.

We discuss the performance of our extended BAMA analyzer in Section 5.

4.2 Manual Annotation

After applying the automatic annotation algorithm to the corpus, two annotators manually verified the results in turn, with the second annotator reviewing the text after the initial set of corrections by the first annotator. A custom Java annotation tool was used for this stage of manual annotation (see Figure 5). The depth of morphological analysis planned for the corpus exceeded that provided by BAMA. Although the analyzer produced most of the planned features, certain key parts of the morphological analysis could only be produced manually. This included missing verb voice (active/passive), the energetic mood for verbs, the interrogative *alif* prefix, identifying participles, verb forms, and disambiguating *lām* prefixes.

Although each of these features had to be added by hand, most do not occur very often, and the analyzer nearly always correctly identified the remaining set of features. BAMA produces stems not roots, an important distinction. It was possible to automatically annotate the root for each word. These roots were imported from the open source Zekr Quran browser (http://zekr.org), which contains an accurate verified root list, used to support the search feature in that software.



Word-by-Word Grammar - Verse (1:1)

Welcome to the **Quranic Arabic Corpus**, an annotated linguistic resource which shows the Arabic grammar, syntax and morphology for each word in the Holy Quran. Click on an Arabic word below to see details of the word's grammar, or to suggest a correction.

Chapter (1) sūrat l-fātiḥah (The Opening)

Translation	Arabic word	Syntax and morphology
(1:1:1) In (the) name bis'mi	بِسْمِ N P	P – prefixed preposition *bi* N – genitive masculine noun جار ومجرور
(1:1:2)	ٱللَّهِ	PN – genitive proper noun → Allah

Figure 6: Word-by-word morphological annotation of the Quran at http://corpus.quran.com

```
bi+ POS:N LEM:{som ROOT:smw M GEN
POS:PN LEM:{ll~ah GEN
Al+ POS:ADJ LEM:r~aHoma`n ROOT:rHm MS GEN
Al+ POS:ADJ LEM:r~aHiym ROOT:rHm MS GEN
Al+ POS:N LEM:Hamod ROOT:Hmd M NOM
l:P+ POS:PN LEM:{ll~ah GEN
POS:N LEM:rab~ ROOT:rbb M GEN
Al+ POS:N LEM:Ea`lamiyn ROOT:Elm MP GEN
Al+ POS:ADJ LEM:r~aHoma`n ROOT:rHm MS GEN
Al+ POS:ADJ LEM:r~aHiym ROOT:rHm MS GEN
```

Figure 7: Part-of-speech and morphological feature tags for the first 10 words of the Quran (one line per word).

4.3 The Quranic Arabic Corpus Online

The Quranic Arabic Corpus is available online at http://corpus.quran.com. For each word in the Quran, a morphological analysis is shown, with a visual representation of its morphological segments (see Figure 6). The website includes a search feature which allows keywords to be used in either Arabic or English. A word-by-word contextual English translation is displayed along with a phonetic transcription developed using an algorithm driven by diacritics and annotations. The morphological tags used are explained for each word using a plain English description, as shown in the column in the right in the above screenshot. To ease reading and correction, the Arabic text is shown with different colors for each segment's part-of-speech, such as blue for nominals and purple for adjectives. The fully annotated corpus can be browsed online, and all data is made freely available for download, encoded in both XML and plain text format (Figure 7).

To further increase the accuracy of annotations, an online website was setup, allowing interested volunteers to submit corrections, effectively turning annotation of the Quran into a community effort. Accuracy can be discussed by posting comments. This is a simple forum feature, where corrections to the morphology for each word can be suggested. Suggestions are reviewed by an Arabic linguist before being incorporated into the morphological annotations in the corpus.

The current status of the corpus is that corrections are still being made, although they are being submitted less frequently, giving confidence that the accuracy of annotation has increased over time. The methodology used to produce the corpus has been one of incremental annotation. Starting with an automatic algorithm using full diacritics, two annotators, and then online corrections, it is now hoped that the annotations are of a high accuracy.

5. Evaluation

To evaluate the accuracy of automatic versus manual annotation, we consider the number of words that required revision to their annotation at each stage of correction. The automatic algorithm outlined above produced an analysis for 67,516 out of 77,430 words (87% unchecked recall). Complete coverage was not possible due to out-of-vocabulary errors in the BAMA lexicon. Following automatic analysis, the morphological annotation was reviewed in stages by several annotators. A paid native speaker of Arabic reviewed every word in the Quran working full-time over a three-month period. At this stage, corrections were made to 21,550 words (28%). This included the 9,914 words not analyzed by the automatic algorithm (13% of all words), as well as 11,636 corrections to existing analyses (15% of all words). This allows us to measure the performance of automatic annotation at 72% (recall), 83% (precision) and 77% (F-measure). Recall and accuracy are identical in this case since every word received only one analysis (or no analysis). The automatic algorithm correctly analyzed approximately 3/4 of all words. Without using BAMA, it is likely to have taken a single annotator far more than three months to manually tag each word in the corpus.

Annotation stage	Words revised	% of Quran
automatic algorithm	67516	87.19
annotator #1	21550	27.83
annotator #2	1014	1.3
online corrections	2,000	2.6

Figure 8: Number of modifications at each stage of morphological annotation

A second annotator – a trained Arabic linguist – then reviewed the corpus again, including the first annotator's corrections, and made changes to 1,014 words (1.3% of all words). Finally, the corpus was put online for community volunteer correction. This has resulted in over 2000 approved corrections to words by users of the website over several months. This suggests that our current morphological annotation of the Quran is approaching very high accuracy, and should continue to do so over time.

6. Conclusion and Future Work

In this paper, we presented the Quranic Arabic Corpus, which is the first manually verified and computationally analyzed morphological Quran corpus. Building on this work, our next goal is a complete syntactically parsed corpus of the Quran using the morphological analysis presented here (Dukes et al., 2010; Dukes and Buckwalter, 2010). We believe that the methodology of incremental community corrections presented in this work is applicable to other corpora such as classical texts or other important central works.

Acknowledgments

We would like to thank Eric Atwell and Tim Buckwalter for helpful discussion and advice. We also would like to thank all the volunteers who helped make this project a success.

References

Joyce Akesson (2001). Arabic Morphology and Phonology: Based on the Marah Al-Arwah by Ahmad b. 'Ali Mas'ud. Brill.

Tim Buckwalter (2002). Buckwalter Arabic Morphological Analyzer version 1.0. Linguistic Data Consortium, University of Pennsylvania.

Judith Dror, Dudu Shaharabani, Rafi Talmon and Shuly Wintner (2004). Morphological Analysis of the Qur'an. Literary and Linguistic Computing, 19(4):431-452, 2004.

Kais Dukes, Eric Atwell and Abdul-Baquee M. Sharaf (2010). Syntactic Annotation Guidelines for the Quranic Arabic Treebank. In Proceedings of the Language Resources and Evaluation Conference (LREC 2010). Valletta, Malta.

Kais Dukes and Tim Buckwalter (2010). A Dependency Treebank of the Quran using Traditional Arabic Grammar. In Proceedings of the 7th international conference on Informatics and Systems. Cairo, Egypt.

Geert-Jan Kruijff (2006). Dependency grammar. The Encyclopedia of Language and Linguistics 2nd edition, Elsevier Publishers.

Moustafa Elshafei, Husni Al-Muhtaseb, and Mansour Alghamdi (2006). Machine Generation of Arabic Diacritical Marks. In Proceedings of the International Conference on Machine Learning; Models, Technologies & Applications.

Wolfdietrich Fischer and Jonathan Rodgers (2002). A Grammar of Classical Arabic: Third Revised Edition. Yale University Press.

Nizar Habash. (2007) "Arabic Morphological Representations for Machine Translation." Book Chapter. In Arabic Computational Morphology: Knowledge-based and Empirical Methods. Editors Antal van den Bosch and Abdelhadi Soudi.

Nizar Habash and Owen Rambow. (2005) Arabic Tokenization, Morphological Analysis, and Part-of-Speech Tagging in One Fell Swoop. In Proceedings of the Conference of the Association for Computational Linguistics (ACL'05), Ann Arbor, MI.

Nizar Habash and Ryan Roth (2009). CATiB: the Columbia Arabic Treebank. In Proceedings of (ACL'09), Suntec, Singapore.

John A. Haywood and H. M. Nahmad (2005). A New Arabic Grammar of the Written Language. Lund Humphries Publishers.

Alan Jones (2005). Arabic Through the Qur'an. Islamic Texts Society.

Mohamed Maamouri, Ann Bies and Tim Buckwalter (2004). The Penn Arabic Treebank: Building a large-scale annotated Arabic corpus. In NEMLAR Conference on Arabic Language Resources and Tools, Cairo, Egypt.

Ebrahim Muhammad (2007). From the Treasures of Arabic Morphology. Zam Zam Publishers.

Abdullah Abbas Nadwi (2006). Vocabulary of the Holy Quran. Millat Book Centre.

Abdul Mannan Omar (2005). Dictionary of the Holy Quran. Noor Foundation International.

Jamal-Un-Nisa Bint Rafai (1998). Basic Quranic Arabic Grammar. Ta-Ha Publishers Ltd.

Jonathan Owens (1988) The Foundations of Grammar: An Introduction to Medieval Arabic Grammatical Theory. John Benjamins Publishers.

Karin C. Ryding (2008). A reference grammar of Modern Standard Arabic. Cambridge University Press.

Abdur Rashid Siddiqui (2008). Quranic Keywords: A Reference Guide. The Islamic Foundation.

Otakar Smrž and Jan Hajič (2006). The Other Arabic Treebank: Prague Dependencies and Functions. In Arabic Computational Linguistics: Current Implementations, CSLI Publications.

Jane Wightwick and Mahmoud Gaafar (2008). Arabic Verbs and Essentials of Grammar. McGraw-Hill.

Shuly Wintner (2008). Strengths and weaknesses of finite-state technology: a case study in morphological grammar development. Natural Language Engineering 14(4):457-469.