UNIVERSITY OF LEEDS

This is a repository copy of *Supervised collaboration for syntactic annotation of Quranic Arabic*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/81650/

Version: Accepted Version

Article:
Dukes, K, Atwell, ES and Habash, N (2013) Supervised collaboration for syntactic annotation of Quranic Arabic. Language Resources and Evaluation, 47 (1). 33 - 62. ISSN 1574-020X

https://doi.org/10.1007/s10579-011-9167-7

Reuse
Unless indicated otherwise, fulltext items are protected by copyright with all rights reserved. The copyright exception in section 29 of the Copyright, Designs and Patents Act 1988 allows the making of a single copy solely for the purpose of non-commercial research or private study within the limits of fair dealing. The publisher or other rights-holder may allow further reproduction and re-use of this version - refer to the White Rose Research Online record for this item. Where records identify the publisher as the copyright holder, users can verify any specific terms of use on the publisher's website.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

White Rose university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Supervised Collaboration for Syntactic Annotation of Quranic Arabic

Kais Dukes[1] **· Eric Atwell · Nizar Habash**

Revised submission: 25 March 2011

Abstract The Quranic Arabic Corpus (http://corpus.quran.com) is a collaboratively constructed linguistic resource initiated at the University of Leeds, with multiple layers of annotation including part-of-speech tagging, morphological segmentation (Dukes & Habash, 2010) and syntactic analysis using dependency grammar (Dukes & Buckwalter, 2010). 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 project contrasts with other Arabic treebanks by providing a deep linguistic model based on the historical traditional grammar known as i'r*ā*b (إعراب). By adapting this well-known canon of Quranic grammar into a familiar tagset, it is possible to encourage online annotation by Arabic linguists and Quranic experts.

This article presents a new approach to linguistic annotation of an Arabic corpus: 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 are 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 this review. In this paper we evaluate and report on the effectiveness of the chosen annotation methodology. We also discuss the unique challenges of annotating Quranic Arabic online and describe the custom linguistic software used to aid collaborative annotation.

Keywords Collaborative Annotation · Arabic · Treebank · Quran · Corpus

[1] K. Dukes (✉) · E. Atwell
University of Leeds, United Kingdom
e-mail: sckd@leeds.ac.uk

N. Habash
Columbia University, New York, USA

1 Introduction

Online collaborative annotation (Chamberlain et al., 2009) has recently emerged as an increasingly viable alternative to more conventional approaches for developing tagged corpora. In a linguistic setting, online collaboration has been used for a wide variety of tasks ranging from syntactic annotation of Latin and Ancient Greek texts (Bamman et. al., 2009) to named-entity resolution of international hotels (Su et al., 2007). This article discusses the collaborative aspects of constructing the multilingual Quranic Arabic Corpus (Dukes et al., 2010), an annotated linguistic resource in English and Arabic, developed through online volunteer contributions. The Quran is Islam's central religious text, written 1,400 years ago in a form of Arabic that is no longer native. However, the Quran is sometimes studied in depth from childhood, so effectively this type of Arabic is a form of 'second language'.

Although the focus is Arabic, the collaborative methods presented in this article are sufficiently general to be of wider interest to other annotation efforts. The main task that online annotators are asked to perform is to proofread morphological and syntactic tagging. Annotators verify this tagging against 'gold standard' analyses from Arabic reference books of grammatical analysis of the Quran. This reference material contains equivalent grammatical information, but in an unstructured form (see figure 1). Accuracy is important when analyzing a religious text such as the Quran, especially as the annotated data in the Quranic corpus is distributed and used by several other related Arabic language projects. To ensure consistent, high-quality annotation across the corpus, a small number of expert annotators are promoted to a supervisory role, reviewing and discussing the work of others by comparing against the large body of historical analysis of canonical Quranic grammar.

(A)

استمع: فعل ماض مبني على الفتح
نفر: فاعل مرفوع بالضمة
من الجن: جار ومجرور متعلق بصفة محذوفة من نفر

(B)

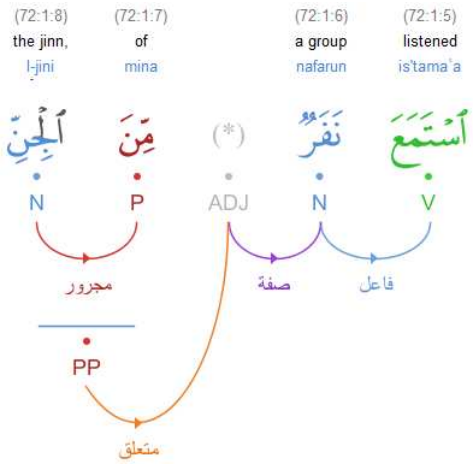

Figure 1 (A) An example from one of the 'gold standard' traditional Arabic grammar textbooks. (B) The corresponding visual representation of dependency grammar, from the Quranic Arabic Corpus.

An observation underlying the chosen approach to annotation is that adopting an intuitive annotation procedure allows annotators to spend more time focusing on making key linguistic decisions. Providing a well-documented and easy-to-use interface is essential for an online effort where volunteer contributors may not have the motivation or time to learn how to use a non-intuitive annotation tool.

The approach of supervised collaborative annotation used to develop the Quranic corpus can be compared to recent work that evaluates the use of crowdsourcing, such as Amazon Mechanical Turk, for either simpler tasks or tasks more usually performed by trained annotators. In (Snow et al., 2008; Nowak & Rüger, 2010) it is shown that for certain tasks, a majority vote applied to the annotation of a single item can be used to filter out the noisy judgements of non-experts. We argue that supervised collaboration may be more suitable for a sensitive corpus such as the Quran, Islam's central religious text, and is likely to lead to more accurate results when the number of non-experts outweighs more experienced contributors. In this article's evaluation section, we report on a small-scale experiment using Amazon Mechanical Turk to attempt linguistic annotation of Quranic Arabic, with lower resulting accuracy compared to the methodology of supervised collaboration.

The Quranic corpus can also be compared to more conventional approaches to annotating Arabic corpora. Four recently developed Arabic treebanks (Maamouri et al., 2004; Smrž & Hajič, 2006; Habash & Roth, 2009; Al-Saif & Markert, 2010) use a small number of paid annotators. Quality is ensured by providing a well-documented set of guidelines, by following a training process, and by having different annotators make multiple passes of the same text. In a collaborative setting, annotation guidelines still apply, but training and quality control need to be handled more carefully. When constructing the Quranic Arabic Corpus, it was found that making the annotation process as intuitive as possible led to greater accuracy and consistency, more rapid annotation, and attracted a larger number of expert linguists and Quranic scholars, who are willing to spend more time volunteering contributions. In contrast with previous Arabic annotation efforts, the Quranic corpus directly uses the historical traditional Arabic grammar of i'r*ā*b (إعراب), which is immediately familiar to the majority of linguists trained in the Arab world, as well as to formal students of the unique language of the Quran. The corpus also includes several layers of annotation which are not necessarily part of the collaborative effort, but are derived from other trusted sources or generated by software tools. These serve to make the website a more attractive and useful resource generally, and help to attract and motivate volunteer collaborators. These additional resources include verse-aligned audio recordings, an automatically generated phonetic transcription with prosodic phrasing, seven alternative verse-aligned English translations from authoritative published sources, a chapter-verse-word number indexing system to simplify cross-referencing, and a word-for-word aligned English translation.

In order to further simplify the annotation process, natural language generation is applied to generate concise and easily readable descriptions of morphological and syntactic tagging for each word in the corpus. In addition, an evolving set of annotation guidelines are updated whenever difficult syntactic constructions are encountered for the first time during volunteer annotation. A message board is used as an online forum to promote open discussion between annotators and users of the corpus, who are typically Arabic students or Quranic researchers. These users have

an interest in, and sometimes challenge, the existing annotation in the corpus developed by collaborators.

Although the corpus is a successful and useful annotated resource as suggested by user feedback, organizing online collaborative analysis of Quranic Arabic is particularly challenging (http://corpus.quran.com/feedback.jsp). Nearly all annotators are in agreement over the most important grammatical features for each word, such as part-of-speech and grammatical case. However, encouraging a large number of volunteers to contribute to annotation through linguistic discussion can lead to differences of opinion that are often hard, if not impossible, to resolve definitively for a small proportion of words in the corpus. Despite not being a key focus of the linguistic tagging tasks, much inter-annotator disagreement revolves around the issues of the most appropriate interlinear Arabic-to-English translation and the subtly different uses of gender in Quranic Arabic. We discuss these and other related cases further in section 4, as well as describe the resolution procedure used, where possible, to guide collaborators towards agreement in difficult cases.

This article is organized as follows. Section 2 provides an overview of the custom linguistic software and tools used to aid online collaborative annotation. Section 3 compares the resource to previous tagged Arabic corpora and discusses the special challenge of linguistic annotation for Quranic Arabic in a collaborative setting. We highlight key parts of Quranic grammar, and describe the online morphological and syntactic annotation tasks for the corpus. Section 4 describes the annotation methodology, which involves a multi-stage approach of automatic rule-based tagging, initial manual verification, followed by online collaborative proofreading. In section 5, we evaluate the chosen methodology by deriving combined accuracy measures over the two classes of non-expert and supervisory annotators. We report that introducing a supervisory role later in the project boosted the accuracy of suggestions by 22%, as measured by comparing to gold standard reference works of canonical Quranic grammar. We also consider the reliability of the data as compared to more conventional forms of annotation for Arabic corpora. Section 6 describes current applications of the annotated resource, and concludes with a discussion of planned collaborative tasks for further linguistic annotation.

2 Collaborative Annotation Tools and Resources

The Quranic corpus has been developed by following the principles of supervised collaboration with inter-annotator discussion (described further in section 4) and ease-of-use. The second principle of usability is essential when online volunteers may not have the motivation or time to follow a non-intuitive annotation process. In this section we describe the online interface used by collaborators to review linguistic analyses, shown in figure 2. The website displays annotations generated from morphological and syntactic tags stored in a linguistic database. To provide an easily usable online resource, a drill-down interface (Böhm & Daub, 2008) is used to ‘zoom’ into annotations, summarizing linguistic tagging at different levels of detail. This type of interface is not usually applied to tagged corpora, but is especially useful for a rich, layered dataset such as the Quranic corpus.

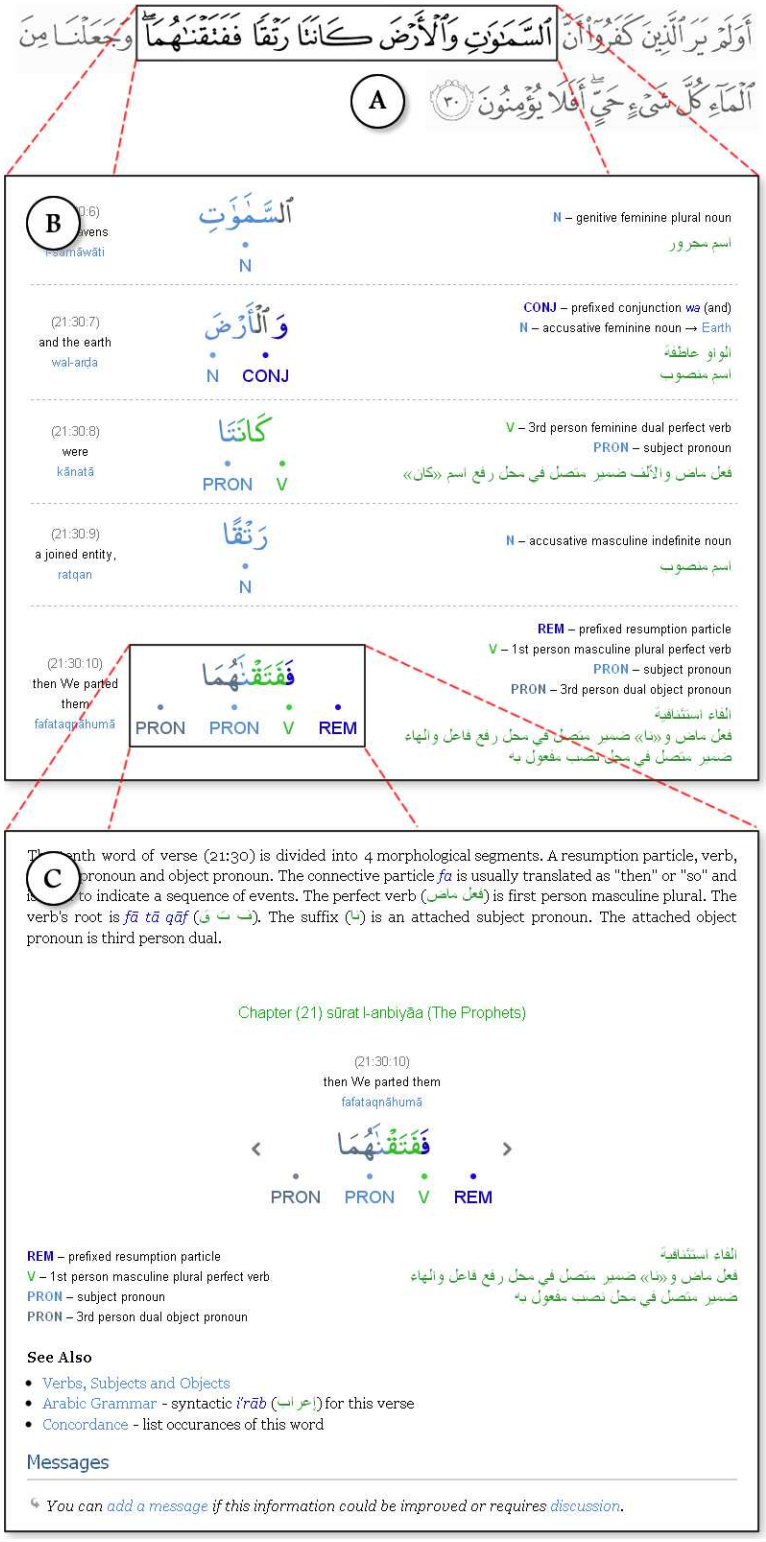

Figure 2
Drill-down interface.

To encourage volunteer collaborators to assist with annotation, suggesting corrections online is designed to be a subtle and non-intrusive process. The majority of the website's monthly 50,000 visitors use the resource to understand the original Arabic of the Quran through grammatical analysis, and are not involved in online discussion. A small but active minority of collaborators participate in proofreading and work towards improving the resource. Most users who decide to volunteer begin by using the website to research a particular section of the Quran that they have an interest in, and then go on suggest corrections to annotation that could be improved.

For each verse in the Quran, the original Arabic script (figure 2a) is displayed online alongside seven parallel translations into English. Clicking on the Arabic script displays the website's most used feature, the interlinear format (figure 2b) (Bow et. al., 2003; Pietersma; 2002). This shows a running word-by-word summary of annotation for each verse, alongside an algorithmically generated phonetic transcription, and a word-aligned interlinear translation into English. Color-coding is used to highlight morphological segmentation of the Arabic script, with corresponding grammatical summaries displayed in both Arabic and English. Collaborators can view further detail for an individual word by clicking through to the analysis web page, where template-based Natural Language Generation (NLG) is used to generate a more detailed grammatical summary for each word in the corpus (figure 2c). The analysis page allows collaborators to review all relevant tags for each word in the corpus using a textual summary that describes morphological segmentation, part-of-speech tagging, and syntactic dependency analysis in English and Arabic. For example, a typical Quranic word such as fafataqn*āhumā* (ففتقناهما), translated as then we parted them both, will have a detailed grammatical description generated automatically using the tags stored in the linguistic database:

> The tenth word of verse (21:30) is divided into 4 morphological segments. A resumption particle, verb, subject pronoun and object pronoun. The connective particle fa is usually translated as "then" or "so" and is used to indicate a sequence of events (الفاء استئنافية). The perfect verb (فعل ماض) is first person masculine plural. The verb's root is *fā tā qāf* (ت ف ق). The suffix (نا) is an attached subject pronoun. The attached object pronoun is third person dual.

Based on observing inter-annotator discussion, the majority of collaborators usually prefer to proofread morphological and syntactic analysis in this textual format, instead of reviewing lists of abbreviated tags, features and syntactic relations. The benefit of this approach is that since the grammatical information is equivalent, the underlying tags in the database are indirectly reviewed in parallel. At the same time, a textual format is more easily comparable to the linguistic analyses in gold standard reference works of canonical Quranic grammar. Collaborators are invited to review and suggest corrections to this information online (this methodology is described further in section 4). An 'add message' button on the analysis page allows collaborators to start a new discussion thread, with comments for a specific word shown alongside annotations:

> You can add a message if this information could be improved or requires discussion.

To simplify the proofreading process, the analysis page includes a 'See Also' section that provides a set of contextual hyperlinks that are used by annotators to directly access related resources and tools (figure 2c). This additional usability feature allows online collaborators to spend more time making key linguistic decisions. Quick and easy 'one click' access to relevant information provides the ability to see the choices and decisions made previously by other collaborators for related words in the corpus. This compares with other annotation projects for tagging Arabic offline that require annotators to spend time searching through guidelines and other documentation, often without direct access to the work of others who may be working in isolation on the same annotated text. The contextual hyperlinks in the 'See Also' section are generated dynamically according to the type of word under analysis, depending on part-of-speech, syntactic role and morphology. For example, for the previously discussed Arabic word *fafataqnāhumā* in verse (21:30), hyperlinks provide quick access to the relevant section in the annotation guidelines for verbs, subject and objects. Additional contextual links provide a graphical visualization of syntax using dependency graphs, as well as further links to other online grammatical analyses for the verse at related Arabic grammar and Quran websites.

Two other popular resources provided alongside corpus annotations are the Quranic dictionary and morphological search. Both these resources are based around root, lemma and stem, which in Arabic linguistics are distinct concepts. Roots are an abstract grouping of words, and lemmas are a further subdivision. The root of an Arabic word is not a word itself, but a sequence of three or four letters, known as radicals, from which most words can be derived through the Arabic template-pattern system. A lemma is a real representative word that groups together other related words that differ by inflection, and is used as entry headers in standard Arabic dictionaries. The simplest non-inflected form of a word is chosen as the lemma: third person masculine for verbs and singular for nouns. Stems arise in morphological segmentation and are not necessarily actual words. After removing clitics from a compound word-form, the stem will remain.

The online morphological search tool acts as a powerful concordance, allowing annotators to find related words by searching on part-of-speech, stem, lemma, root and other annotated morphological features (http://corpus.quran.com/searchhelp.jsp). The Quranic dictionary organizes words first by root then further by lemma, and provides a contextual translation into English. Both these online tools allow collaborators to quickly find related words, so that comparing against previous annotations and related analyses is made easier. The Quranic corpus project also includes a mailing list with hundreds of subscribers, including active annotators and interested Arabic linguists and Quranic experts worldwide.[1] The message board allows collaborators to discuss annotations for a particular word, while the mailing list is an extended forum for more general topics such as refinements to the tagset, enhancements to annotation guidelines and general project discussion. All of these additional tools and resources help to make the online annotation effort as simple as possible, and gives proofreading collaborators access to further related information and more detailed context when needed. We considered using an existing open-

[1] comp-quran mail archive: http://www.mail-archive.com/comp-quran@comp.leeds.ac.uk

source Wiki platform to host the discussion forum; but we concluded it was essential to integrate the search and feedback mechanisms into a tailor-made architecture.

3 The Challenge of Collaborative Annotation for Quranic Arabic

In this section, we discuss the some of the challenges faced when performing linguistic annotation for Quranic Arabic in a collaborative setting. In 3.1 we contrast our approach to other recent annotated Arabic corpora, and also consider similarly constructed collaborative resources for other languages such as ancient Greek and Latin. In 3.2 we discuss the nature of the gold standard reference material used by annotators to assist with online proofreading. This helps to overcome the difficultly of working with annotations for an ancient text in language that is no longer natively spoken. In 3.3 we describe the collaborative morphological and syntactic annotation tasks performed by annotators, and outline our choice of tagset and syntactic representation.

3.1 Related Annotated Corpora

Developing a successful machine-readable annotated language resource depends both on the quality of the data, as well as on the choice of computational linguistic representation. Processing a highly inflected and morphologically rich language such as Arabic presents a unique set of challenges, as noted by (Soudi et al., 2007):

> The morphology of Arabic poses special challenges to computational natural language processing systems. The exceptional degree of ambiguity in the writing system, the rich morphology, and the highly complex word formation process of roots and patterns all contribute to making computational approaches to Arabic very challenging.

Quranic Arabic, the unique form of the Arabic language used in the Quran, is not spoken today except in restricted liturgical contexts (Jones, 2005), but it is the direct ancestor language of Modern Standard Arabic (MSA). Annotating the Quran presents a different set of challenges compared to MSA due to the fact that the text is over 1,400 years old. The Quranic script is more varied than modern Arabic in terms of orthography, spelling and inflection. For example, the same word can be spelt different ways in separate chapters. Fortunately, it is possible to build on previous experience in annotating Arabic using more conventional approaches, when considering how best to annotate the Quranic corpus in an online collaborative setting. In addition, the Quran is fully diacritized which reduces its ambiguity.

The Quranic Arabic Corpus is not the first attempt to produce a machine-readable linguistically annotated resource for the Quran. Previous related work includes the offline morphological analysis performed at the University of Haifa (Dror et al., 2004). The authors note that in comparison to MSA, Quranic Arabic remains relatively unexplored in the context of computational linguistic analysis and annotation:

> Except for isolated efforts, little has been done with computer-assisted analysis of the text. Thus, for the present, computer-assisted analysis of the Quran remains an intriguing but unexplored field.

In comparison to the Quranic Arabic Corpus, this previous automatic processing of the Quranic text was not completed, and remains manually unverified with multiple possible analyses for each word in the final published dataset. Based on considering a random sample, the authors of the Haifa analysis estimate the final accuracy of annotations using an F-measure of 86% (Dror et al., 2004).

Previous work for annotating Modern Standard Arabic includes the three major Arabic treebanks that have been recently developed: the Penn Arabic Treebank (Bies & Maamouri, 2003; Maamouri et. al., 2004), the Prague Arabic Dependency Treebank (PADT) (Hajič et. al., 2004; Smrž & Hajič, 2006) and the Columbia Arabic Treebank (CATiB) (Habash & Roth, 2009; Habash et. al., 2009). These corpora were tagged offline through the more conventional approach of using a small number of paid trained annotators. Each of the resulting treebanks has a different scope and aim, and each has its own form of representation for modeling Arabic morphology and syntax. Typically these tagging schemes were reused from previous annotation projects for other languages, such as English, and adjusted to fit Arabic. The primary use of these existing treebanks is as a resource to train statistical parsers of Arabic, and to provide empirical evidence for the frequency of Arabic linguistic constructions. Figure 3 below compares these tagged Arabic corpora to the annotations in the Quranic corpus. A more detailed comparison of linguistic tagging schemes is provided in (Atwell, 2008), and for Quranic Arabic in particular see (Sawalha & Atwell, 2010).

Figure 3 Comparison of syntactically annotated Arabic corpora.

Corpus	Dependency	Features	Traditional
Penn	no	yes	no
PADT	yes	yes	no
CATiB	yes	no	yes (subset)
Quranic	yes (hybrid)	yes	yes

The second column in figure 3 indicates if the resource has been syntactically annotated using dependency grammar or if constituent phrase structure is used. The next column indicates if morphological feature tagging is included in the mark up, which involves annotating each word segment with additional linguistic information, such as person, number, gender, lemma, noun cases and verb moods. The last column specifies if traditional Arabic grammar is used, which simplifies the annotation process for Arabic in an online collaborative setting. Both the Penn and the Prague Arabic treebanks use models of syntactic representation which are not immediately intuitive to native speakers of Arabic, often requiring training to be able to participate in the annotation effort. In contrast, the Columbia Treebank (CATiB) uses a subset of traditional grammar which is sufficient for further development of statistical parsing. Moving towards historical traditional grammar for annotating Arabic corpora has been shown to allow for more rapid annotation with minimal user training, due to the use of familiar standardized terminology (Habash et. al., 2009). For a more detailed comparison of MSA treebanks see (Habash, 2010).

Related work also includes corpus linguistics and text analytics applied to other sacred books and historic collections of text aside from the Quran. The AIBI

conferences on the Bible and the Computer have introduced many related and applicable concepts such as the interlinear format for online word-by-word display of annotated texts (Pietersma, 2002) although applying similar concepts to the Quranic text has not yet been done computationally. Work related to concepts and ontologies for other sacred texts include (Wilson, 2000) who provided a conceptual glossary and index to the Vulgate translation of the Gospel according to John. The syntactic annotation in the Quranic Arabic Corpus carried out as part of this research project can also be compared to treebanks for other morphologically rich languages, such as dependency annotation for Latin and Ancient Greek (Bamman et. al., 2009) although unfortunately this does not cover any religious texts such as the Greek new testament. While not equipped with rich visual interfaces, the annotations in the Latin and Ancient Greek project are made by specialized experts.

3.2 Traditional Arabic Grammar (إعراب)

In the Arabic-speaking world, there is a long tradition of understanding the Quran through grammatical analysis, and over the centuries this knowledge has accumulated in a grammatical framework known as *i'rāb* (إعراب). One motivation for the historic development of traditional Arabic grammar has been to understand functional inflection. In Arabic, nouns can be found in one of three cases (the nominative, genitive or accusative case). Each of these grammatical cases is realized through a different morphemic case-ending, which results in the noun being pronounced slightly differently, and written using different vowelized diacritics. Similarly, imperfect verbs (فعل مضارع) are found in three main moods (the indicative, subjunctive or jussive). Automatic prediction of case and case-endings has been one focus of recent computational research for Arabic (Habash et. al. 2007; Zitouni et. al. 2006; Habash & Rambow, 2007).

A fundamental aim of historical traditional Arabic grammar is to explain the reason for the inflection of each noun and verb in a sentence, based on syntactic function. For example, when a noun is the subject of a verb it is found in the nominative case, yet when it is the object of a verb it is found in the accusative case (Mace, 2007; Muhammad, 2007). To relate inflection to syntactic function for the entire Arabic language requires a sophisticated grammatical framework. A well-defined and thought out grammatical theory, *i'rāb* is capable of handling multiple parts-of-speech, and a wide variety of linguistic constructions and grammatical dependencies. A key insight for online collaborative annotation of Quranic Arabic is to extend the approach of using traditional syntax in the CATiB treebank (Habash & Roth, 2009) by attempting to represent as much of traditional Arabic grammar as possible. This leads to morphological and syntactic annotation which uses familiar terminology, and enables anyone who is already experienced with Quranic syntax to immediately participate in the annotation effort. This is especially important for online volunteer annotators who may not have the time or motivation to undergo a lengthy training process in order to understand a non-standard form of syntactic representation, as required for other recently developed tagged Arabic corpora.

In addition to being a more familiar grammatical framework, adopting i'r*ā*b allows accuracy to be measured and improved by cross-referencing against a large published body of works on canonical Quranic grammar. Given the importance of

the Quran to the Islamic faith, any syntactic annotation needs to be carefully considered since alternative parses for a sentence can suggest alternative meanings for the scripture for certain verses. Fortunately, the unique form of Arabic in which the Quran has been inscribed has been studied in detail for over 1,000 years (Jones, 2005; Ansari, 2000). This is far longer than corresponding grammars for most other languages, and in fact traditional Arabic grammar is considered to be one of the historic origins of modern dependency grammar (Kruijff, 2006; Owens, 1988).

In more conventional approaches to corpus annotation, the accuracy of annotated data is usually inferred from the value of an inter-annotator agreement coefficient such as the κ-statistic (Carletta, 1996). For annotating Quranic Arabic, it is possible to use a collection of certain key reference works as a form of gold standard to measure accuracy (Dukes et. al., 2010). The primary reference for the annotation effort is (Salih, 2007) al-i'r*ā*b al-mufassal li-kit*ā*b all*ā*h al-murattal ("A Detailed Grammatical Analysis of the Recited Quran using i'r*ā*b"), a carefully researched work that collates and builds on centuries of grammatical analysis. This 12-volume authoritative reference of canonical Quranic grammar spans over 10,000 pages.[1] For each of the 77,430 words in the Quranic text, a detailed description is given of not only part-of-speech and morphology, but also a contextual syntactic analysis using dependency grammar. Although most of the annotation in the Quranic corpus can be cross-checked against al-i'r*ā*b al-mufassal, this work does not cover several morphological features which are tagged using online collaboration. For verifying the annotation of derived Arabic verb forms and roots, as well as for grammatical gender, Lane's Lexicon (Lane, 1992) and Wright's reference grammar (Wright, 2007) are used. Both of these are widely considered to be highly authoritative reference works on classical Arabic grammar, and for the Quran in particular.

Producing a machine-readable annotated resource backed by these existing gold standard analyses is not simply a matter of scanning in the material and applying automatic character recognition. The Quranic Arabic Corpus is designed to be an open source resource, and any material used must be free of copyright. Even if this was not a concern, character recognition for printed Arabic texts such as Salih's al-i'r*ā*b al-mufassal is still an active area of research (Amara & Bouslama, 2005). A further obstacle to automatic extraction is that the grammatical analyses in these reference works are not encoded as a series of easily machine-readable tags or tables. Instead the syntactic dependencies and morphological analyses are described in free text, often using detailed technical linguistic language. The approach to syntactic annotation for the Quranic corpus, described further in section 4, is to apply automatic rule-based parsing to the original Arabic text, which is then manually verified online. This is cross-checked using these gold standard reference works of *i'rāb*, which provide existing analyses for each word in the Quran. The Quranic Arabic Corpus provides for the first time an accurate structured database based on this wealth of existing traditional analysis, with annotation developed using online collaboration.

[1] Available online: http://www.archive.org/download/imkam12

3.3 Collaborative Morphological and Syntactic Annotation Tasks

Collaborative annotation for the Quran is particularly challenging due to the rich morphology of the language, and the depth of information requiring review. Annotation tasks involve proofreading morphological segmentation, part-of-speech tags and inflection features, as well as reviewing syntax graphs using dependency grammar. The rule-based morphological analyzer, used for initial offline tagging of Quranic Arabic, produces automatic annotation with an F-measure accuracy score of 77% (Dukes & Habash, 2010). This initial data is inserted into a structured linguistic database and then proofread by volunteer contributors using online discussion. A second level of annotation involves using the resulting manually verified morphological annotations for further automatic syntactic parsing which is in turn proofread online.

Figure 4 Multilingual natural language generation is used to simplify collaborative annotation.

The first of these proofreading tasks is segmentation. In contrast to English, for a highly inflected language such as Arabic, morphological segmentation is an important prerequisite task before full syntactic analysis (Habash et. al., 2009; Bies & Maamouri, 2003). In the Penn English Treebank, verbal contractions such as *weren't* are split into separate segments (were and *n't*), each with a different part-of-speech (Bies et. al., 1995). These segments form individual units in syntactic analysis, each being separate leaf nodes in a syntax tree. The situation for Quranic Arabic is similar, but as many as 54% of the Quran's 77,430 words require segmentation, resulting in 127,806 morphological segments. A typical 'word' in the Quran consists of multiple segments fused into a single whitespace-delimited word form, as shown in figure 4 above. This example (read from right-to-left) shows a prefixed conjunction, a verb, and two suffixed pronoun clitics, where segmentation has been performed according to traditional *i'rāb*.

Online volunteers correcting morphological annotation review this choice of segmentation together with the set of features annotated for each segment. These include person, gender, number, grammatical case and verb moods. As shown in figure 4, natural language generation is used to derive concise summaries in both

Arabic and English based on the tags and morphological features annotated in the corpus. This online presentation of data differs from that used during initial offline processing, where for each word in the Quran a rule-based morphological analyzer builds a feature-value matrix (Habash, 2007; Soudi et al., 2007; Smrž, 2007). Figure 5 shows the underlying matrix for the same Arabic word in the corpus corresponding to figure 4:

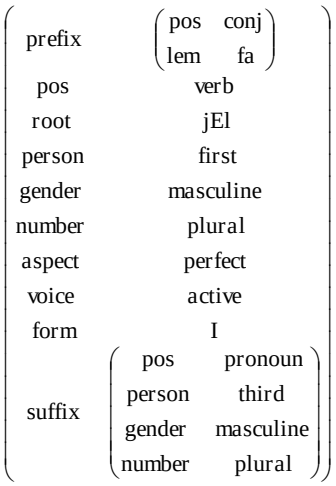

Figure 5 Morphological feature-value matrix produced by the offline rule-based tagger.

These features are encoded using a sequence of machine-readable morphological tags. The database record corresponding to above feature-value matrix is stored as:

```
f:CONJ+ POS:V ROOT:jEl 1MP PERF ACT (I) PRON:3MP
```

Although machine-readable and well documented, this format is not easily understandable by online volunteer annotators. For other Arabic corpora tagged offline, annotation is encoded using an analogous set of abbreviated tags. For example, the MADA system uses a similar set of features and a related tagset for MSA (Habash, Rambow & Roth, 2009). In contrast, for the collaboratively developed Quranic Arabic Corpus, it is necessary to translate these tags into an online format more easily understandable by users of the resource and by annotators (figure 4). This expanded representation is crucial in order to attract skilled linguists and Quranic experts to the project, without requiring technical training. The online corpus guidelines (Dukes et. al., 2010) provide detailed documentation for the annotation scheme. In this article, a summary of key part-of-speech tags is shown in figure 6.

Category	Tag	Arabic	Description
Nouns	N	اسم	Noun
	PN	اسم علم	Proper noun
	IMPN	اسم فعل أمر	Imperative verbal noun
Pronouns	PRON	ضمير	Personal pronoun
	DEM	اسم اشارة	Demonstrative pronoun
	REL	اسم موصول	Relative pronoun
Nominals	ADJ	صفة	Adjective
	NUM	رقم	Number
Adverbs	T	ظرف زمان	Time adverb
	LOC	ظرف مكان	Location adverb
Verbs	V	فعل	Verb
Prepositions	P	حرف جر	Preposition
lām prefixes	EMPH	لام التوكيد	Emphatic *lām* prefix
	IMPV	لام الامر	Imperative *lām* prefix
	PRP	لام التعليل	Purpose *lām* prefix
Conjunctions	CONJ	حرف عطف	Coordinating conjunction
	SUB	حرف مصدري	Subordinating conjunction
Particles	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
Disconnected letters	INL	حروف مقطعة	Quranic initials

Figure 6 Part-of-speech tagset for morphological segments (prefixes, stems and suffixes).

The syntactic annotation task involves proofreading dependency tagging. For each numbered verse in the Quran (*āy*ah), a syntax graph is initially generated offline by a rule-based parser, using the previously proofread morphological analysis and part-of-speech tags. In contrast to other syntactically annotated Arabic corpora, the Quranic corpus does not show only bracketed structures or flat lists of relations. To simplify

online collaboration, a novel visualization of Arabic dependency grammar using non-terminal phrase nodes is generated by automated graph-layout algorithms, based on the annotations in the database (figure 7). This online representation shows dependency relations, a phonetic transcription and an interlinear translation into English. This new graphical scheme for Arabic syntax is also being adapted for other related Semitic languages such as Amharic (Gasser, 2010). Figure 7 below shows a dependency graph that describes the syntax of verse (99:1) of the Quran. Read from right-to-left, this visualization describes the Quranic grammar of *i'rāb* using dependency relations, and is used to simplify online collaboration. The Arabic syntactic constructions annotated include a temporal conditional clause (شرط), a passive verb subject representative (نائب فاعل), and a cognate accusative (مفعول مطلق). The dependency on the left is a possessive construction (مضاف إليه) also known as the genitive construction. A more detailed description of this online visual representation of Quranic Arabic syntax is given in (Dukes & Buckwalter, 2010).

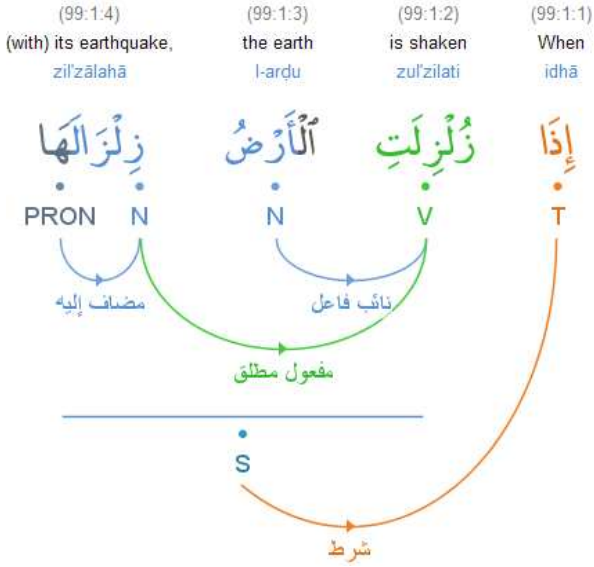

Figure 7 Dependency graph for verse (99:1) of the Quran.

4 Annotation Methodology

4.1 Motivation, Communication and Role-Based Collaboration

The annotation methodology for the Quranic corpus contrasts with other recently developed collaborative linguistic resources, by using role-based differentiation and open communication between contributors. Recent work has shown that the use of crowdsourcing, such as Amazon Mechanical Turk (AMT), can be effective for labelling tasks that can be clearly understood by non-experts. In (Snow et al., 2008), tasks are restricted to multiple-choice responses, and 10 independent annotators label each item. The resulting data is noisier than expert data, but aggregation leads to reliability. In the AMT model of annotation, workers are motivated by small

monetary rewards for each task, but there is typically no explicit communication or collaboration between annotators. Other forms of motivation for non-expert collaboration include enjoyment and interest. Phrase Detectives (Chamberlain et al., 2009) models annotation as a game with a purpose, and uses a suitably entertaining system to capture anaphora resolution. Players review decisions made previously by other contributors, leading to a large reliable natural language resource.

Perhaps the best example of a fully collaborative resource is Wikipedia, constructed entirely by unpaid volunteer editors who are motivated by the interest they share in the articles being developed. Recent research has consistently shown that the effectiveness of Wikipedia depends not only on incremental edits to improve quality, but also crucially on open communication and discussion between editors to resolve issues, and to promote common understanding (Kittur & Kraut, 2010). This compares with the Quranic corpus, where unpaid collaborators are motivated by a common interest in chapters of the Quran, and reach consensus through discussion. The annotation methodology adopted is multi-stage, so that the annotated resource is increasingly accurate at each stage of review. Following initial automatic tagging and offline manual correction by two experts (Dukes & Habash, 2010), the annotated corpus is put online for further collaborative error-detection. Volunteer annotators have an interest in finding mistakes in the existing tagging by comparing against gold standard reference works of i'r*ā*b.

In the Wikipedia model of collaboration, editors work directly on a shared article that continuously improves through the use of incremental edits. For the Quranic corpus, a message board is used to gather suggested corrections.[1] The message board is open to the general public, so that anyone who shares an interest in the annotation effort can participate after a simple registration procedure. There are three different collaboration roles: contributors, editors and supervisors. New users who have recently registered will be general contributors who have read-only access to the annotations, but can post suggested corrections online. Editors are project organizers, and have both read and write access to the linguistic database. They typically do not suggest new corrections, focusing on the more time-consuming task of comparing against gold standard references. When a suggestion is a genuine correction, the corpus annotations are updated, resulting in incremental improvement of accuracy.

Online annotation progressed initially with multiple volunteer contributors providing suggestions, but with only 2 editors reviewing these and making edits (phase A). During a later second stage, the supervisor role was introduced by promoting a small number of contributors to this status (phase B). Supervisors retain their read-only access to annotations, but are differentiated by their ability to veto incorrect suggestions made by other contributors. These trusted experts are chosen if they consistently provide high-quality corrections and have suitable academic credentials. Supervisors typically contribute to multiple discussion threads and provide guidance to less experienced annotators. The effect of introducing a supervisory role is evaluated in section 5, where we report that the accuracy of suggestions considered for edits in phase B increased by 22%. This is due to supervisors filtering out incorrect comments from non-experts, allowing editors to focus on considering suggestions that are more likely to be genuine corrections.

[1] http://corpus.quran.com/messageboard.jsp

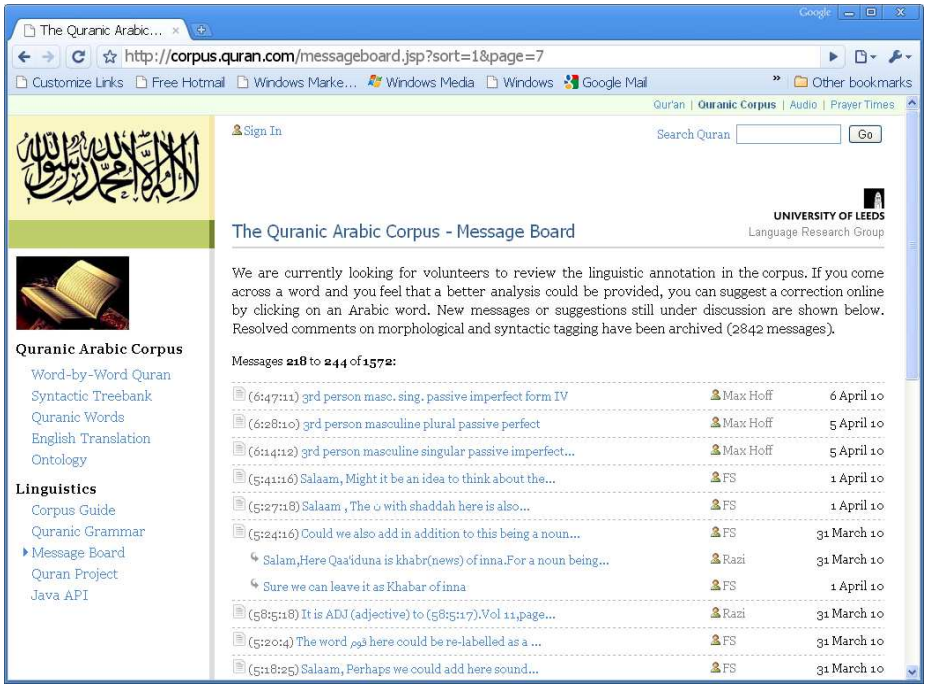

Figure 8 The message board promotes communication between annotators and users of the resource.

Given the rich morphology and syntax of Quranic Arabic, and the depth of annotation, collaborators participate using free text entry as opposed to restricted multiple-choice responses. This more natural form of expression promotes crucial communication between annotators and users of the website. Messages are organized into threads that discuss the correct morphological and syntactic tagging for each individual word in the corpus. An online summary page lists the most recent active threads under discussion (figure 8). A simple but common case of collaboration is a thread in which a contributor suggests a correction that is reviewed by a supervisor:

> 20th April, 2010
> FS: Is this not a LOC - accusative location adverb as opposed to a noun?
> AR: Yes, it is indeed zarf makaan mansoob.
> FS: Thank you.

In this example, a contributor with online pseudonym FS suggests a correction to a word's part-of-speech tag using English terminology. The reviewed suggestion along with others periodically results in annotations being updated by editors. It is revealing that a supervisor responds using equivalent transliterated Arabic terminology. This is possible due to a common understanding that traditional i'r*ā*b is being used as the grammatical framework. In the following related example, a word has an incorrect automatic morphological analysis of 3rd person feminine singular. A contributor is interested in participating in order to highlight incorrect tagging as well as to clarify their own understanding of Arabic grammar:

> 24th April, 2010
> TH: I am a beginner grammar student. I thought this word is 2nd person masculine singular. Please help me understand.
> AR: You are right. The verb is indeed 2nd person masculine singular. This needs to be corrected.

As well as confirming corrections and providing useful educational feedback to contributors, supervisors veto incorrect suggestions made by non-experts. The thread below discusses the correct dependency graph for verse (5:24) of the Quran, using terminology from traditional i'r*ā*b. The issue is if the syntactic role of h*ā*l is applicable, also known as the circumstantial accusative[1]:

> 31st March, 2010
> FS: Could we also add in addition to this being a noun that it is hal?
> RZ: For a noun to be hal it must be mansoob but here noun is marfoo', so it is not hal. Vol 3, page 45. Thanks.
> FS: Sure. We can leave it as khabar of inna.

A supervisor vetoes this suggestion while providing useful feedback. The reason that the suggestion is incorrect is explained: circumstantial accusatives are always found in the accusative case and never in the nominative. As justification, the supervisor provides a reference to volume 3, page 45 of (Salih, 2007), which provides detailed syntactic analysis for the relevant verse, and is available as a link through the website. The guidelines cite this as the chosen primary work for verifying syntactic annotation; hence the common understanding that an abbreviated reference is sufficient in order to cite the gold standard. As shown by these examples, the dual nature of the message board involves common understanding to incrementally improve the accuracy of a shared resource, but is also an open forum for researchers to engage with subject experts.

4.2 Resolving Inter-annotator Disagreement

The public threads archived on the Quranic message board are an interesting case study in collaborative annotation. For the one year period to May 2010, the combined threads total nearly 5,000 messages (100,000 words) of linguistic discussion between 100 participants. This secondary 'annotator corpus' of recorded interactions most often involves mutual understanding and exchanges between collaborators and supervisors, but also contains cases of disagreement before reaching consensus. A full linguistic analysis of the many varied discussions covered online is beyond the scope of this article. In this section, we focus on examples of the methods used to resolve different types of inter-annotator disagreement, when resolution is possible at all.

[1] This functional role from traditional *i'rāb*, along with related syntactic dependencies, is described further in the online annotation guidelines: http://corpus.quran.com/documentation/circumstantialaccusative.jsp

A type of disagreement that is hard or potentially impossible to resolve definitively is translation. Although the Quranic corpus focuses on morphological and syntactic tagging, the website provides a word-by-word interlinear translation into English shown alongside annotations, as a guide for researchers. To aid annotation and to put difficult verses into context, parallel renditions into English are sourced from seven widely used and respected translations, collated by (Abbas, 2009). Despite not being an encouraged collaborative task, many volunteers are interested in translation, often using previously annotated analyses to discuss the relative merits of their arguments:

> 11th February, 2010
> MN: This is in the nominative case. In common parlance "refuge" means a sheltered or protected state. In consideration of being in hell, an appropriate word depicting this may be used, like "abode, place of return".
> AR: The literal meaning of ma'waa is indeed "a place of refuge, a resort, a sheltered and protected place". Its use here is obviously ironical. There are many places in the Quran where the rhetorical device of irony, using a word to mean its exact opposite, is employed to convey a sarcastic, mocking tone. See also 4:138:1-6, 9:3:28-32. The gender of the noun should be feminine because it carries the feminine ending.
> MN: Out of the 7 parallel translations on this website only two have used "refuge". For the same word in 3:151 only Sahih has used "refuge". Not at a single place of its 12 occurrences has it been used ironically.
> AR: Not only Salih, but Arberry also uses "refuge". Pickthall uses "retreat" which the Oxford Dictionary defines as a "place of shelter or seclusion". Alright, the majority use "abode", but the fact remains that "refuge" IS the literal meaning in English of the word ma'waa.

This type of thread can be long-running without a definitive resolution. A consequence of using unrestricted free text to suggest online corrections is that collaborators occasionally engage in fringe discussion. Although interesting from a linguistic perspective, corpus editors ignore off-topic suggestions such as translation, and focus on reviewing corrections to annotations and tagging. If necessary, collaborators are encouraged to continue unresolved off-topic conversations using the project's mailing list, keeping these off the public message board.

Aside from translation, most cases of disagreement involving annotation tasks are usually resolved. Consensus is achieved through following an escalating resolution procedure. The most common method for resolution is to refer to the annotation guidelines, which specify preferred annotation style. If the type of disagreement is more fundamental, such as differing parts-of-speech, or if the annotation guidelines require enhancing, annotators are challenged to each cite references to the literature that justify their analyses. In the situation that both annotators provide justifications for differing analyses, the analysis from the gold standard reference texts is adopted as definitive (Salih, 2007; Lane, 1992). After a difficult linguistic construction is encountered for the first time by annotators and agreement is reached, the evolving set of online annotation guidelines are improved. Consensus is also achieved through refining and clarifying the definition of the annotation tasks, following feedback

from expert collaborators. As an example, the part-of-speech tagset (figure 6) has been refined and better documented by introducing more granular parts-of-speech for Quranic particles, so that the chosen tags more closely align with the linguistic analyses in the gold standard reference works.

An interesting case of disagreement between annotators that highlights the resolution process is the gender of angels according to the Quran. The historical context for this inter-annotator discussion is a belief in pre-Islamic Arabia that angels were the daughters of God (Al-Mubarakpuri, 2003). A common theme in the Quran is that God has no offspring. It is also generally accepted in Islam that angels are not feminine creatures, as indicated by verse (43:19) which refers to pre-Islamic beliefs:

وَجَعَلُوا۟ ٱلْمَلَٰٓئِكَةَ ٱلَّذِينَ هُمْ عِبَٰدُ ٱلرَّحْمَٰنِ إِنَٰثًا ۚ أَشَهِدُوا۟ خَلْقَهُمْ ۚ سَتُكْتَبُ شَهَٰدَتُهُمْ وَيُسْـَٔلُونَ ﴿١٩﴾

(43:19) And they have made the angels, who are servants of the Most Merciful, females. Did they witness their creation? Their testimony will be recorded, and they will be questioned.

Gender in Quranic Arabic is a detailed and sensitive issue, as highlighted by the following example which discusses the correct grammatical analysis for a related verse. According to traditional Quranic exegesis, the noun mu*'aqqibātun* (literally, the successive ones) in verse (13:11) refers to angels. In Arabic linguistics, the concept of gender may refer to semantic, morphemic or grammatical gender. A word can have different values for these three attributes, as gender can differ across meaning, form and syntactic function. In the Quranic corpus, grammatical gender is tagged, since this determines how words function syntactically, and allows gender agreement to be considered through relations in dependency graphs. The noun mu*'aqqibātun* (معقبات) has a feminine-sounding morphemic ending, but acts as grammatically masculine. This noun was initially incorrectly tagged as feminine by the offline rule-based analyzer. At the time of the online discussion below, the guidelines required enhancing, and should have clarified that grammatical gender is being tagged as opposed to morphemic or semantic gender. The thread begins with an annotator challenging the incorrect automatic tagging of feminine on semantic and exegetic grounds. An example is provided of a related Arabic word that is semantically masculine, but morphemically feminine-sounding ("Caliph" or khalifa):

> 17th November, 2009
> MN: The word "angels" does not go with feminine, since the Quran states that only disbelievers describe angels as feminine. Can't mu*'aqqibātun* be considered masculine like khalifatan?

A second annotator suggests that grammatical as opposed to semantic gender should be tagged, but unfortunately provides an incorrect analysis of grammatically feminine. The annotator indicates that consensus can be reached by verifying the differing analyses against gold standard references. This is the next step towards resolution after consulting the annotation guidelines:

> KD: The full grammatical analysis for this word is feminine plural, active participle from ʿaqqaba, form II of ʿaqiba. This word is a grammatical feminine. This does not mean that angels are feminine.
> MN: How can one accept a grammatical analysis for this word as feminine plural?
> KD: Can you please cite a reference for your own grammatical analysis?

Inter-annotator discussion is a crucial part of the Quranic Arabic Corpus and leads to more accurate analyses through common understanding. This is highlighted by a third annotator who suggests that although the word has a feminine-sounding morphemic surface ending, it may be important to tag the word online as grammatically and semantically masculine, especially as the corpus website is used as a study resource for researchers:

> HS: I think, and I may be mistaken, that I might agree with MN with the fact that this word should be treated as masculine. As far as my knowledge of Arabic is concerned, the plural of some masculine words might sound feminine. One example is when we say: rijal*ā*t instead of rij*ā*l and it is originally a masculine word. The issue here is sensitive, since tagging this word as feminine might lead the readers to think that the word "angels" in Arabic is feminine, hence angels are feminine creatures. We understand that the word per se, both grammatically and semantically speaking, is masculine.

The original collaborator agrees with the sensitive nature of the analysis, but stresses that the word should not be tagged as feminine on semantic grounds:

> MN: I feel concerned because of verse (53:27): Indeed, those who do not believe in the hereafter name the angels female names.

A fourth annotator contributes to the gender tagging thread for the noun mu*ʾaqqibātun* by using the website's built-in concordance tool. The referent word mal*ā*ekah (angels) is shown to act as either grammatically masculine or feminine in related verses, according to context. The use of the lemma-based concordance provides annotators with quick and easy access to examples of tagging for previous related words:

> AB: I took a corpus linguistics approach and looked at the concordance lines for the 54 occurrences of malaekah. Of these, 32 occurrences used pronouns to refer to the angels in the same verse, and showed that 21 used masculine and 11 used feminine pronouns. One verse (47:27) used both masculine and feminine pronouns. So, in reality angels are not female (based on 43:19 and other verses). But grammatically the majority of the time they are referred to as males and sometimes as females.
> KD: It's great the Arabic concordance functionality for the Quran on this website can be used in such a way. It is interesting that both feminine and masculine pronouns are used, purely in the sense of grammatical (not real semantic or physical) gender.

For this thread and for related examples, consensus between annotators is reached through discussion. In this particular verse, the word mu*'aqqibātun* although feminine in form, is masculine in meaning as well as in grammatical function. The thread concludes with the next stage of the resolution procedure. The analysis is confirmed by the original collaborator who verifies against the gold standard reference, in this case (Lane, 1992) for Quranic Arabic gender annotation:

> MN: I got this information from the Lane's Lexicon entry for this word: While feminine in form, grammatically this is masculine. This is a double plural, and so is masculine in the same way.
> KD: It looks like your reference from Lane's Lexicon sums this up. This reference does suggest that we change this word to masculine.

The conclusion to this discussion thread is that the annotation guidelines required enhancing to specify that grammatical gender is being tagged, as opposed to morphemic or semantic gender. Following the above discussion, the guidelines have since been extended to explain the different types of gender in Quranic Arabic.[1] This resolution process and annotation methodology contrasts with recent collaborative efforts that use a majority vote to filter out the noisy judgements of non-experts. For a sensitive corpus such as the Quran, Islam's central religious text, inter-annotator discussion is crucial for accurate results when the number of non-experts generally outweighs more experienced contributors. Experts proofreading Quranic annotations typically cite references and take the time and effort to pursue and justify their analyses. Through discussion and communication between collaborators, consensus can be reached even for linguistically challenging topics such as the correct gender tagging for angels in the Quran, as well as for related issues.

We did consider the possibility of an undecidable ambiguity: Atwell (2008) notes that some corpus tagging schemes allow for two part-of-speech tags for rare occasions when the part-of-speech is genuinely ambiguous. For example, the tag JJ|VBG means a word could be an adjective or a present participle verb, and nothing in the context tells the annotator which is correct. However, we found no demand or need for such a mechanism: in cases of apparent disagreement, discussion and analysis always seems to end in a consensus. This could be because collaborative discussion allows for much more reflection than a single annotator trying to follow guidelines.

5 Evaluation

5.1 Accuracy of Annotations

As explained in section 3.2, although we do have gold standard reference texts, the analyses these contain are not in a format which can be readily cross-matched to corpus annotations, so that we cannot easily apply standard metrics of precision and

[1] Annotation guidelines for gender tagging: http://corpus.quran.com/documentation/gender.jsp

recall to evaluate the corpus against the gold standard. Indirect evidence for having confidence in the accuracy of the annotations can be found by comparing website usage (figure 9) to message board activity (figure 10). The inverse trends indicate that although more people continue to make use of the online annotated resource over time, the number suggested corrections has decreased, since errors are becoming harder to find as accuracy improves.

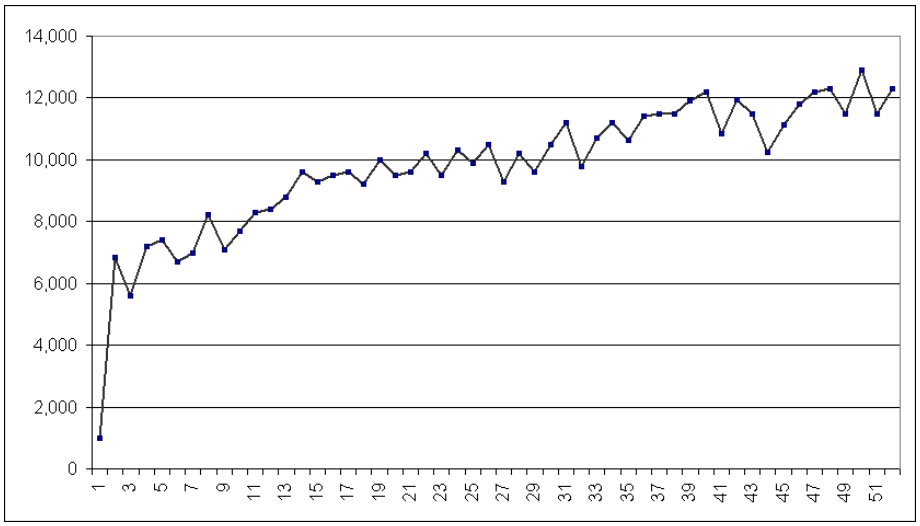

Figure 9 Website visitors per week, over a 1 year period.

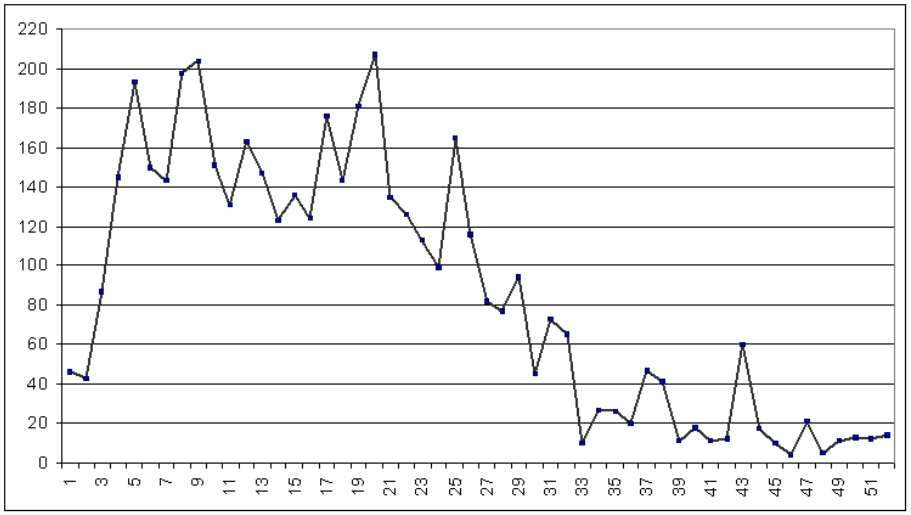

Figure 10 Message board posts per week.

A more precise measurement of accuracy can be obtained through random sampling. In this section we focus on evaluating the accuracy of morphological annotation in the corpus, which is considered to be a stable part of the tagging effort

and near completion. The Quranic Arabic Dependency Treebank currently provides syntactic dependency graphs for 30% of the Quranic text but is still in progress, while morphological annotation covers 100% of the Quran, and has been proofread online over a year's time. We also do not consider the accuracy of ancillary annotation in the corpus, which includes automatically generated phonetic transcriptions, verse audio recitations and translation.

To measure the accuracy of morphological annotation by random sampling, from the 77,430 words in the Quran we take three random non-overlapping samples of 1,000 words each. The words in each sample need not be in sequence or be from the same verses or chapters. The annotations in the corpus for each of these samples are compared to the analyses in the classical references of Quranic grammar. Typically, accuracy does not vary significantly across each of these samples, so that they can be averaged to give an estimated accuracy measure for the corpus as a whole. Figure 11 shows the number of suggestions during the first two 3-monthly periods of online annotation, for the three samples; for the whole Corpus, there were 1801 suggestions by 3 months, and a further 1728 suggestions by 6 months. As can be seen, the number of suggestions during these time periods is overall evenly distributed among the samples, which demonstrates that they are representative of the annotation effort.

	Suggestions per Sample		
Online Project Time	Sample A	Sample B	Sample C
3 months	21	26	23
6 months	19	24	19

Figure 11 Suggestions per random sample.

The accuracy for the morphological annotation of a single word is measured according to strict criteria. A typical word in the Quran will receive multiple tags and features for different items of linguistic information such as segmentation, part-of-speech, gender, person, number, and grammatical case. We consider a word to be accurately annotated only if all of the features have the correct expected values. Figure 12 summarizes the accuracy of morphological analysis, measured by using the same random samples at 5 different stages of annotation. Each stage of annotation builds on the previous stage by reviewing the existing annotations and making further corrections. The initial data used to seed the online collaborative platform is generated through automatic tagging followed by a first pass of offline correction. Supervisors were introduced after three months of online proofreading by website collaborators. Accuracy is measured at each of these stages, as well as at 6 months and at 12 months into the annotation effort.

Online Project Time	**Stage**	**Accuracy**
-	Automatic annotation	77.2%
-	Initial offline correction	89.6%
3 months	Online proofreading without supervisors	92.5%
6 months	Online proofreading with supervisors	96.9%
12 months	-	98.7%

Figure 12 Accuracy of morphological annotation.

The effect of introducing a supervisory role 3 months into the project can be seen from the accuracy measurements in figure 12. During the first three months of annotation (without supervisors) accuracy improved by 2.9%. For the next 3 months with supervisors, accuracy improved by a further 4.4%. It is also relevant to consider the quality of message board suggestions. For the first three months of online annotation (without a supervisory role), 1331 out of 1801 suggestions resulted in valid corrections to annotations (74%). For the following three months of annotation (with a supervisory role) out of a total of 1728 suggestions, 401 of these were vetoed by supervisors, and out of the remaining 1327 suggestions, 1271 resulted in corrections to the corpus annotations (96%) by editors. Introducing a supervisory role later in the project boosted the quality of suggestions considered by editors by 22%, due to supervisors filtering out inaccurate suggestions made by less experienced contributors. This increase in the quality of suggestions allows editors to focus on considering genuine corrections and comparing only these to the gold standard references.

5.2 Comparative Evaluation

In order the compare the methodology of supervised collaboration to crowdsourcing, a simple experiment was conducted using Amazon Mechanical Turk (AMT), an online job marketplace where workers are matched with requesters offering tasks. These AMT tasks are known as HITS (Human Intelligence Tasks), and are often presented in a multiple choice format, or make use of restricted text entry. Although recent work has shown high accuracy in using AMT for simple annotation tasks (Su et al., 2007; Snow et al., 2008), it is not clear how well the AMT approach would perform for deep linguistic annotation for a genre-specific language such as Quranic Arabic. In the AMT experiment, a 500-word part-of-speech tagged section of the Quranic text was put online for correction by Mechanical Turk workers, and was reviewed independently by 6 contributors. To simplify the experiment, only part-of-speech tags were considered instead of the full set of morphological features. This allowed the AMT experiment to run as a simple multiple-choice task. Unlike with the Quranic corpus, AMT workers are paid a small fee for each completed task. These workers are not necessarily Arabic specialists or volunteers interested in the Quran, but can be anyone with the required skills wanting to earn money for participation.

To ensure a baseline level of competency, the experiment required successful completion of a screening test, which asked 5 challenging multiple-choice questions about Arabic grammar. Only those AMT workers passing the screening test participated in the annotation experiment. The initial data given to AMT was a reduced form of the part-of-speech tagset used to seed the online Quranic Arabic Corpus (stage 2 in figure 12, at 89.6% accuracy). This allows for a more accurate comparison between online supervised collaboration and AMT crowdsourcing. The AMT workers were invited to review this tagging and provide corrections. After this review, the final accuracy of the 500-word sample averaged at 91.2% (an increase of 1.6%). This compares with the 92.5% accuracy in figure 12 at stage 3, for initial online collaboration in the Quranic corpus without supervisors. This would suggest

that involving expert supervisors in the collaborative process, as well as encouraging discussion and communication leads to higher accuracy for a deeply annotated resource such as the Quranic corpus. The current estimated accuracy of morphological annotation in the corpus is measured at 98.7%, using the approach of supervised collaboration.

6 Conclusion and Future Directions

In this article, we presented a description of a collaborative effort to morphological and syntactic annotation of the Quran: The Quranic Arabic Corpus. Given the uniqueness of this text and its importance as a sacred religious book, we devised an online supervised collaboration using a multi-stage approach. The different stages include automatic rule-based tagging, initial manual verification, and online supervised collaborative proofreading. The website 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 are promoted to a supervisory role, allowing them to review or veto suggestions made by other collaborators. We show that this approach produces superior and needed quality compared to more common crowdsourcing methods that lack supervision. Given the special characteristics of our task, we decided not to use an existing Wiki platform to host the discussion forum, but decided to integrate the search and feedback mechanisms into a tailor-made architecture. This has proven to be a useful and popular contribution to Quranic Arabic research that provides new ways to study the Quran. The website receives 1,500 interested visitors each day (see figure 13):

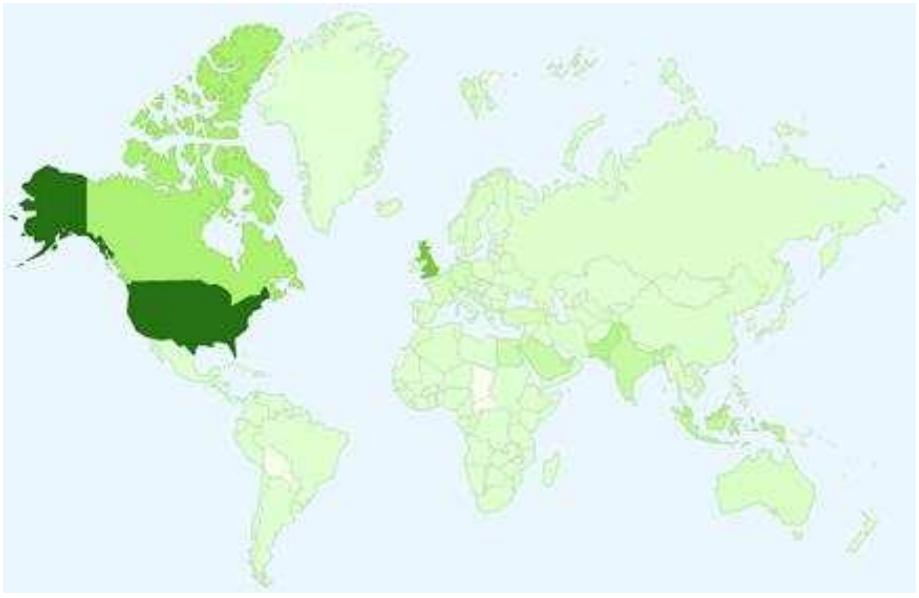

Figure 13 The annotated Quranic Arabic Corpus attracts worldwide interest. The map above shows the distribution of the website's 50,000 users per month by geographic region. Countries with more visitors are shaded darker. (Source: 245,781 visitors between 20 Dec 2009 and 20 May 2010).

There are currently several ongoing academic research projects that use annotated linguistic data from the Quranic Arabic Corpus, which is freely available for download under an open source license. The data has been manually verified by multiple annotators, and the linguistic database is machine readable. This data is being used for training and testing, for example in (Yusof et. al., 2010), where a rule-based stemmer is developed and tested against data from the Quranic Arabic Corpus. They note that although there has been a recent focus on measuring the performance of systems using annotated MSA corpora, little work has been done in evaluating systems using Quranic Arabic. We expect more work to be done in this area, following the publication of the Quranic corpus datasets. Another application is formal semantic modelling of Quranic verses. In (Zaidi et. al., 2010), data from the Quranic Arabic Corpus is used to develop an ontology through extraction rules written using GATE. Our choice of syntactic representation inspired by traditional Arabic grammar may also be applicable to other related languages. Gasser (2010) develops a dependency grammar for the related Semitic language of Amharic, and uses a similar syntactic scheme that also includes hidden nodes in dependency graphs.

Future collaborative work will include further morphological annotation: verb and noun patterns, different types of gender (semantic versus functional gender), and refined segmentation rules adapted from traditional Arabic grammar. Additionally, we plan to finish the syntactic annotation which is in progress. To further improve accuracy, we plan to introduce a concept of 'quality labels' for certain sections, so that passages which are in need of in-depth review and discussion can be more easily identified by volunteer annotators. As well as making suggestions, it may also be useful for contributors to mark sentences that they have checked and found correct. We are also working on developing a first draft of full-coverage anaphoric resolution for pronouns in the Quran, which will be displayed online alongside the existing corpus annotations for collaborative review. Finally, we plan to integrate the treebank into other standard tools for computational linguistics and language processing, such as the open-source NLTK and GATE toolkits, to widen take-up of the Quranic Arabic Corpus as a training and testing corpus for general NLP research. We have previously argued that understanding the Quran, and other texts widely considered to be notable, should be a grand challenge for Computer Science and Artificial Intelligence (Atwell et. al., 2010). We are also interested in possibilities for collaborating with other annotation projects, via reuse of our collaborative architecture for online linguistic analysis and research.

Acknowledgments

We would like to thank Lydia Lau and Katja Markert at the School of Computing, University of Leeds for providing invaluable feedback and numerous suggestions to improve the quality of this article. We thank Wajdi Zaghouani at the Linguistic Data Consortium, University of Pennsylvania for assistance in devising the Amazon Mechanical Turk experiment for tagging the Quran via crowdsourcing. We also acknowledge the hard work of the supervisors and other volunteer collaborators involved in online annotation of the Quranic Arabic Corpus.

References

N. Abbas (2009). Qurany: A Tool to Search for Concepts in the Quran. MSc Research Thesis, School of Computing, University of Leeds.

J. Alderete, P. Tupper and S. A. Frisch (2009). Phonotactic Learning without A Priori Constraints: A Connectionist Analysis of Arabic Cooccurrence Restrictions.

S. Al-Mubarakpuri (2003). Tafsir Ibn Kathir. Darussalam Publishers.

A. Al-Saif and K. Markert (2010). The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic. Language Resources and Evaluation Conference (LREC). Valletta, Malta.

N. Amara and F. Bouslama (2005). Classification of Arabic Script Using Multiple Sources of Information: State of the Art and Perspectives. International Journal on Document Analysis and Recognition. 5(4)(2005),195-212.

H. Ansari (2000). Learning the Language of the Quran. MMI Publishers.

E. Atwell (2008). Development of Tagsets for Part-of-Speech Tagging. Corpus Linguistics: An International Handbook. Mouton de Gruyter.

E. Atwell, K. Dukes, A. Sharaf, N. Habash, B. Louw, B. Abu Shawar, T. McEnery, W. Zaghouani, M. El-Haj (2010). Understanding the Quran: A New Grand Challenge for Computer Science and Artificial Intelligence. ACM/BCS Visions of Computer Science. Edinburgh, Scotland.

D. Bamman, M. Francesco and G. Crane (2009). An Ownership Model of Annotation: The Ancient Greek Dependency Treebank. Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories, Milan.

A. Bies, M. Ferguson, K. Katz, and R. MacIntyre (1995). Bracketing Guidelines for Treebank II Style, Penn Treebank Project. University of Pennsylvania, Philadelphia.

A. Bies and M. Maamouri (2003). Penn Arabic Treebank Guidelines. http://www.ircs.upenn.edu/arabic

C. Bow, B. Hughes and S. Bird (2003). Towards a General Model of Interlinear Text. Proceedings of EMELD Workshop.

K. Böhm and E. Daub (2008). Geographical Analysis of Hierarchical Business Structures by Interactive Drill Down. Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information. Irvine, California.

J. Carletta (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. Computational Linguistics, 22(2):249–254.

J. Chamberlain, U. Kruschwitz and M. Poesio (2009). Constructing an Anaphorically Annotated Corpus with Non-experts: Assessing the Quality of Collaborative Annotations. Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources.

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

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

K. Dukes, E. Atwell and A. M. Sharaf (2010). Syntactic Annotation Guidelines for the Quranic Arabic Treebank. Language Resources and Evaluation Conference (LREC). Valletta, Malta.

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

M. Gasser (2010). A Dependency Grammar for Amharic. Workshop on Language Resources and Human Language Technologies for Semitic Languages, Language Resources and Evaluation Conference (LREC). Valletta, Malta.

N. Habash (2007). Arabic Morphological Representations for Machine Translation. Arabic Computational Morphology: Knowledge-based and Empirical Methods. Springer.

N. Habash (2010). Introduction to Arabic Natural Language Processing. Synthesis Lectures on Human Language Technologies, Graeme Hirst, Editor. Morgan & Claypool Publishers.

N. Habash, R. Faraj and R. Roth (2009). Syntactic Annotation in the Columbia Arabic Treebank. Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt.

N. Habash, R. Gabbard, O.Rambow, S. Kulick and M. Marcus (2007). Determining Case in Arabic: Learning Complex Linguistic Behavior Requires Complex Linguistic Features, Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Prague, Czech Republic.

N. Habash and O. Rambow (2007). Arabic Diacritization through Full Morphological Tagging, Proceedings of the North American chapter of the Association for Computational Linguistics (NAACL), Rochester, New York.

N. Habash, O. Rambow and R. Roth (2009). MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt.

N. Habash and R. Roth (2009). CATiB: The Columbia Arabic Treebank. *Proceedings of (ACL'09)*, Suntec, Singapore.

J. Hajič, O. Smrž, P. Zemanek, J. Snaidauf, and E. Beska (2004). Prague Arabic Dependency Treebank: Development in Data and Tools. Proceedings of the NEMLAR International Conference on Arabic Language Resources and Tools.

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

A. Kittur and R. Kraut (2010). Beyond Wikipedia: Coordination and Conflict in Online Production Groups. Proceedings of the 2010 ACM Conference on Computer Supported Cooperative work. Savannah, Georgia, USA.

G. Kruijff (2006). Dependency Grammar. The Encyclopedia of Language and Linguistics, 2nd Edition. Elsevier Publishers.

E. Lane (1992). Arabic-English Lexicon. Islamic Texts Society.

M. Maamouri, A. Bies and T. Buckwalter (2004). The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. NEMLAR Conference on Arabic Language Resources and Tools, Cairo, Egypt.

J. Mace (2007). Arabic Verbs. Bennett & Bloom.

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

S. Nowak and S. Rüger (2010). How Reliable are Annotations via Crowdsourcing: A Study about Inter-annotator Agreement for Multi-label Image Annotation. Proceedings of the International Conference on Multimedia Information Retrieval, Philadelphia, Pennsylvania.

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

A. Pietersma (2002). A New Paradigm for Addressing Old Questions: The Relevance of the Interlinear Model for the Study of the Septuagint. Bible and Computer: The Stellenbosch AIBI-6 Conference.

B. Salih (2007). al-*i'rāb al*-mufassal li-*kitāb allāh al*-murattal ("A Detailed Grammatical Analysis of the Recited Quran using *i'rāb*"). Dar Al-Fikr, Beirut

M. Sawalha and E. Atwell (2010). Fine-Grain Morphological Analyzer and Part-of-Speech Tagger for Arabic Text. Language Resources and Evaluation Conference (LREC). Valletta, Malta.

A. Sharaf (2009). Knowledge Representation in the Quran. Interim Report, University of Leeds.

A. Soudi, A. Bosch and G. Neumann (Editors, 2007). Introductory Chapter, Arabic Computational Morphology: Knowledge-based and Empirical Methods. Springer.

O. Smrž (2007). Functional Arabic Morphology. Formal System and Implementation PhD Thesis, Charles University, Prague, Czech Republic.

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

R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng (2008). Cheap and Fast - But is it Good? Evaluating Non-expert Annotations for Natural Language Tasks. Proceedings of EMNLP.

Q. Su, D. Pavlov, J. Chow, and W. Baker (2007). Internet-Scale Collection of Human-Reviewed Data. Proceedings of WWW 2007.

R. Yusof, R. Zainuddin, M. Baba and Z. Yusof (2010). Qur'anic Words Stemming. Arabian Journal for Science and Engineering (AJSE).

A. Wilson (2000). Conceptual Glossary and Index to the Vulgate Translation of the Gospel according to John. Hildesheim, Olms-Weidmann.

W. Wright (2007). A Grammar of the Arabic Language. Simon Wallenberg Press.

S. Zaidi, M. Laskri and A. Abdelali (2010). Arabic collocations extraction using Gate. *IEEE ICMWI'10*. Algiers, Algeria.

I. Zitouni, J. S. Sorensen, and R. Sarikaya. (2006) Maximum Entropy Based Restoration of Arabic Diacritics. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia.