Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank

Kais Dukes, Eric Atwell and Abdul-Baquee M. Sharaf

School of Computing, University of Leeds, LS2 9JT, United Kingdom

E-mail: sckd@leeds.ac.uk, csc6ea@leeds.ac.uk, a.m.sharaf08@leeds.ac.uk

Abstract

The Quranic Arabic Dependency Treebank (QADT) is part of the Quranic Arabic Corpus (http://corpus.quran.com), an online linguistic resource organized by the University of Leeds, and developed through online collaborative annotation. The website has become a popular study resource for Arabic and the Quran, and is now used by over 1,500 researchers and students daily. This paper presents the treebank, explains the choice of syntactic representation (إعراب), and highlights key parts of the annotation guidelines. The text being analyzed is the Quran, the central religious book of Islam, written in classical Quranic Arabic (c. 600 CE). To date, all 77,430 words of the Quran have a manually verified morphological analysis, and syntactic analysis is in progress. 11,000 words of Quranic Arabic have been syntactically annotated as part of a gold standard treebank (إعراب القرآن الكريم). Annotation guidelines are especially important to promote consistency for a corpus which is being developed through online collaboration, since often many people will participate from different backgrounds and with different levels of linguistic expertise. The treebank is available online for collaborative correction to improve accuracy, with suggestions reviewed by expert Arabic linguists, and compared against existing published books of Quranic Syntax.

1. Introduction

Annotating an Arabic corpus presents a set of unique challenges when compared to linguistic annotation for texts in other languages, due to complex orthography and highly inflected morphology (Habash, 2007; Habash, Rambow & Roth, 2008). Annotation guidelines are especially important for a corpus developed through online collaboration. Correct annotation of the Quran requires not only a deep understanding of Arabic linguistics, but also of the source material, the Quran itself. 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 in certain cases. 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 origins of modern dependency grammar (Kruijff, 2006; Owens, 1988).

In the Arab-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). The key insight in developing the Quranic Arabic Dependency Treebank is that instead of using an alternative theory of Arabic syntax, the treebank should attempt to adopt as much of traditional *i'rāb* as possible. This contrasts with the approaches used in other recent Arabic treebanks, but has brought many benefits to the project. For example, the Penn Arabic Treebank (Maamouri, Bies & Buckwalter, 2004) follows constituency phrase structure grammar whereas the Prague Arabic Treebank (Smrz & Hajic, 2006) uses a form of dependency grammar known as Functional Generative Description. Using familiar syntax and terminology for the Quranic Arabic Treebank has attracted volunteer Quranic scholars and expert Arabic linguists to the project. In addition, the many detailed published works on Quranic syntax can be leveraged to verify syntactic annotation for each verse of the Quran.
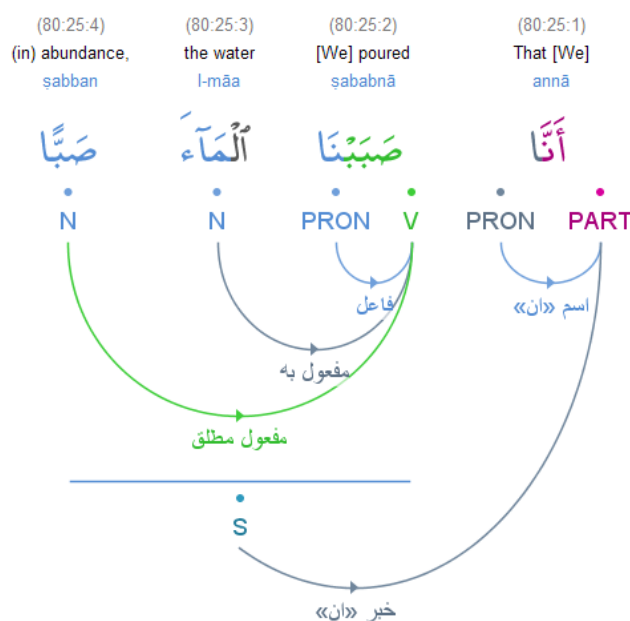

Figure 1: A hybrid dependency graph.

This paper is organized as follows. Section 2 introduces traditional Arabic grammar and describes the annotation process, including a description of the syntactic relations used to label dependency graphs. Section 3 highlights key parts of the full annotation guidelines[1], and Section 4 concludes.

---

[1] The treebank and accompanying documentation are available online at http://corpus.quran.com/treebank.jsp.

2. Syntactic Annotation of Quranic Arabic

2.1 Traditional Arabic Grammar (إعراب)

Arabic is a morphologically rich language, and is highly inflected. One motivation for the historic development of traditional Arabic grammar has been to understand functional inflection. 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 case ending, which results in the noun being pronounced in a slightly different way, and written using different vowelized diacritics. Similarly, imperfect verbs (فعل مضارع) are found in three main moods (the indicative, subjunctive or jussive). 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 a 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 and is written using an alternative vowelized case ending (Mace, 2007; Muhammad, 2007).

To relate inflection to syntactic function for the entire Arabic language requires a sophisticated grammatical framework, capable of handling multiple parts-of-speech, and a wide variety of linguistic constructions and grammatical dependencies. By adopting traditional Arabic grammar, as an educational resource the Quranic Treebank is more accessible to the wider public, and in addition the project attracts a larger number of volunteers including experts who have received formal training in *i'rāb*. Using more familiar terminology also speeds up the syntactic annotation process (Habash, Faraj & Roth, 2009).

However, traditional *i'rāb* is challenging to represent computationally. Unlike in English, where words are typically assigned a single part-of-speech, the fundamental syntactic unit in *i'rāb* is not a word, but morphological word segments. Quranic Arabic is morphologically rich, and often a single word will consist of a stem with multiple fused prefixes and suffixes. Each of these morphological segments is assigned a part-of-speech in traditional Arabic grammar, and can take an independent syntactic role in the sentence that influences inflection (Figure 1). Syntactic dependencies between morphological word-segments is a unique complexity not found in languages such as English. For example, an Arabic noun with a fused preposition prefix will always be inflected for the genitive case (Akesson, 2001). Together these two morphological segments form a syntactic preposition phrase (جار ومجرور), even though this written as a single whitespace-delimited word.

The Quranic Treebank introduces a novel approach to annotating these traditional Arabic grammatical relations. Dependency graphs are used to visualize the syntax of the Quran. This is not only a useful educational resource, but is also a machine-readable representation of Quranic grammar suitable for further research. The syntactic representation adopted in the treebank is a hybrid dependency / constituency phrase structure model. This is motivated by the fact that the Quranic treebank closely follows traditional grammar, and this representation is flexible enough to represent nearly all aspects of traditional syntax. Dependency graphs are used in the treebank to show relations between words, but relations between phrases are also possible by introducing non-terminal nodes.

Figure 1 shows a hybrid dependency graph. Arabic is read from right-to-left and directed edges in the graph point from dependent nodes toward head nodes. The terminal nodes are morphological segments. The graph also makes use of a non-terminal phrase node. This node, marked as *S*, represents a sentence which fills the role of a predicate. The above analysis could be collapsed into a pure dependency graph without non-terminal nodes, by using a transformation in which a relation that ends at a node is applicable to the entire sub-graph headed by that node. However, by using non-terminal nodes, the treebank more accurately follows historical analysis, since traditional Arabic grammar often describes relations between phrases, as well as between words and word segments. This representation has also been found to be more easily understood by annotators who are native Arabic speakers, who use existing published works of Quranic grammar as a reference to verify syntactic annotation in the Treebank.

2.2 The Syntactic Annotation Process

The annotation methodology used in the Quranic Arabic Dependency Treebank follows an iterative approach, involving different stages of annotation. A rule-based dependency parser developed specifically for Quranic Arabic is used to perform initial syntactic analysis, with an F-measure accuracy of 78% (Dukes & Buckwalter, 2010).
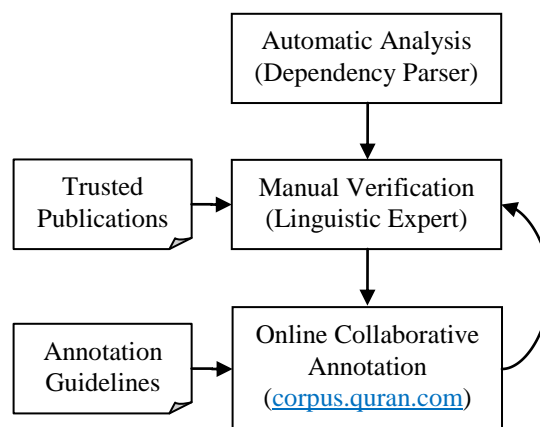

Figure 2: Iteration stages in the annotation process.

The manual stages do not involve annotators performing complete syntactic annotation, but rather correction of automatic annotation performed by the dependency parser. Using a parser not only speeds up annotation but encourages greater internal consistency. The same construct should get the same automatic analysis, leaving proofreaders to focus on correcting exceptional cases.

Cat*	Rel	Arabic	Description
1	adj	صفة	Adjective
	poss	مضاف إليه	Possessive construction
	pred	مبتدأ وخبر	Predicate of a subject
	app	بدل	Apposition
	spec	تمييز	Specification
	cpnd	مركب	Compound (numbers)
2	subj	فاعل	Subject of a verb
	pass	نائب فاعل	Passive subject
	obj	مفعول به	Object of a verb
	subjx	اسم كان	Subject of a special verb
	predx	خبر كان	Predicate of a special verb
	impv	أمر	Imperative
	imrs	جواب أمر	Imperative result
	pro	نهي	Prohibition
3	gen	جار ومجرور	Preposition phrase (PP)
	link	متعلق	PP attachment
	conj	معطوف	Coordinating conjunction
	sub	صلة	Subordinate clause
	cond	شرط	Condition
	rslt	جواب شرط	Result
4	circ	حال	Circumstantial accusative
	cog	مفعول مطلق	Cognate accusative
	prp	المفعول لأجله	Accusative of purpose
	com	المفعول معه	Comitative object
5	emph	توكيد	Emphasis
	intg	استفهام	Interrogation
	neg	نفي	Negation
	fut	استقبال	Future clause
	voc	منادي	Vocative
	exp	مستثني	Exceptive
	res	حصر	Restriction
	avr	ردع	Aversion
	cert	تحقيق	Certainty
	ret	اضراب	Retraction
	prev	كاف	Preventive
	ans	جواب	Answer
	inc	ابتداء	Inceptive
	sup	فجاءة	Surprise
	exh	تحضيض	Exhortation
	exl	تفصيل	Explanation
	eq	تسوية	Equalization
	caus	سببية	Cause
	amd	استدراك	Amendment

*Categories: 1=Nominal dependencies, 2=Verbal dependencies, 3=Phrases and clauses, 4=Adverbial dependencies, 5=Particle Dependencies

Figure 3: Edge labels for syntactic dependency relations.

The second stage of annotation involves manual verification and correction by an Arabic linguistic expert. Using this approach, a single annotator working part-time was able to produce an accurately annotated syntactic dependency treebank of 11,000 words in three months, amounting to 14% of the total 77,430 words in the Quran. The syntactic parses are initially verified by comparing against both existing trusted publications of Quranic grammar, as well as the full annotation guidelines for the project (see Figure 2).

Given the importance of the Quran as a central religious text, a wide variety of interested volunteers regularly participate in the annotation effort online, effectively turning the project into a community effort through online collaborative annotation. While researchers and students make use of the annotated corpus, they are able to add comments to any annotation that they might disagree with, or that they feel requires further clarification. This leads to discussion with other users through an online message board forum (http://corpus.quran.com/messageboard.jsp).

The Quranic grammar message board promotes active discussion, with over 4,000 messages posted over the past 6 months. Some online discussion involves inaccurate suggestions by beginners that are usually resolved through a deeper understanding of Quranic grammar. However, when genuine corrections are presented through online collaborative annotation, these are then referred back to a linguistic expert, who can verify these suggestions against both the annotation guidelines and trusted publications of Quranic syntax, which include books on Quranic grammar, as well as Arabic morphological dictionaries (Nadwi 2006; Omar, 2005; Siddiqui 2008; Wightwick & Gaafar, 2008). General users are also encouraged to use these types of additional information before posting suggested corrections.

2.3 Syntactic Dependency Relations

Traditional Arabic grammar defines several syntactic dependency relations, such as an adjective describing a noun, or a subject relation linking a noun to the verb on which it depends. Figure 3 shows a complete list of the syntactic dependency relations currently annotated in the Quranic Arabic Dependency Treebank. The full list of part-of-speech tags used to label word segments are discussed as part of morphological annotation of the Quranic Arabic Corpus (Dukes and Habash, 2010).

Each of the syntactic relations shown in Figure 3 is used to label edges in dependency graphs in the Quranic Treebank. The list of Arabic dependency tags are taken directly from traditional Arabic grammar, and mapped to equivalent English terms as found in comprehensive publications on Arabic grammatical theory (Haywood & Nahmad, 2005; Ryding 2008). This approach contrasts to other Arabic treebanks (such as the Penn and Prague treebanks) where existing tagging schemes for other languages such as English are adapted to Arabic.

3. Annotation Guidelines

The syntactic annotation guidelines for the Quranic Treebank have been built up over time, and developed during the course of the project. The guidelines are added to whenever a new linguistic construction is discussed during online collaborative annotation that requires further clarification in order to enforce consistency in the corpus. This section highlights key parts of the syntactic annotation guidelines which illustrate a variety of different syntactic constructions in Quranic Arabic, and discusses how these are handled in the traditional Arabic grammar of *i'rāb* (إعراب). The full set of guidelines covering a wider range of linguistic constructions is available online at: http://corpus.quran.com/documentation/grammar.jsp.

3.1 Verbs, Subjects and Objects

Traditional grammar places linguistic constraints on the possible analysis of a sentence. On such constraint is that every verb must have a subject. This will be either an explicit terminal node of the graph (a word or morphological word segment), or otherwise an implicit hidden node used to fill this syntactic role. A verb may optionally accept an object, and ditransitive verbs will take two objects.
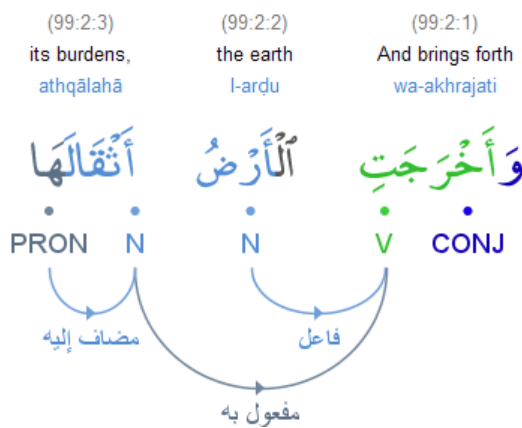

Figure 4: A verb with its dependent subject and object.

Reading Figure 4 from right-to-left, the verb is followed by a subject and then its object. VSO word order is typical in Arabic, although other word orders are also possible and are not ambiguous, since a subject will always be inflected for the nominative case, and objects are always found in the accusative case (Haywood & Nahmad, 2005).

Passive verbs do not have subjects associated with them. Instead, traditional Arabic grammar defines a syntactic role named *nāib fā'il* (نائب فاعل) which may be translated as the "passive subject representative". As with active verbs, a similar constraint exists so that this role must always be filled either explicitly or else implicitly through a hidden node. Figure 5 shows an example of a passive verb followed by its subject representative.
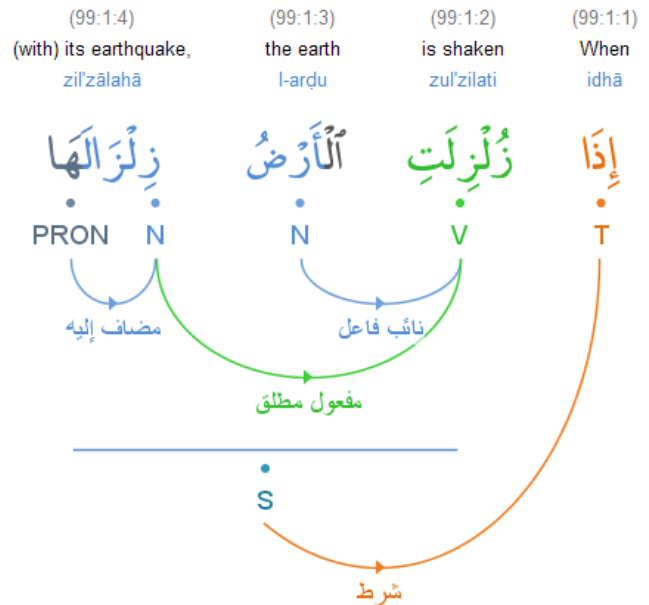

Figure 5: Syntactic annotation of a passive verb.

The above dependency graph also contains a conditional relation between the first word (99:1:1) and the following phrase. In Arabic, the word *idhā* appears as a conditional particle when used in a temporal sense, and is usually translated as "when". The clause following this word will be the protasis of a conditional statement, and will often be a clause or sentence beginning with a verb. The other two dependencies in the graph are the cognate accusative (مفعول مطلق), and the possessive construction (مضاف إليه) also known as the genitive construction.

3.2 Hidden and Empty Nodes

Quranic Arabic is a pro-drop language. Certain verbs imply a pronoun subject through inflection which may be dropped from the sentence (Fischer & Rodgers, 2002). Traditional Arabic grammar restores these dropped words which are known as *damīr mustatir* (ضمير مستتر). Although this adds no new additional information to a sentence, the advantage of this approach is that these nodes satisfy constraints and can be referenced later, for example as part of anaphora resolution. Different inflected hidden pronouns are used depending on the verb's person, gender and number. An additional benefit of showing implicit hidden pronouns in the treebank is that an annotator can quickly determine if the verb has been tagged with correct inflection features.

Figure 6 shows two sentences related through conjunction. Each sentence has a verb with an implicit subject pronoun, shown in gray and in brackets in the dependency graph. In addition to hidden nodes, dependency graphs may also include empty nodes used to fill syntactic roles. These are shown in the treebank using the asterisk notation (*). For a discussion of empty nodes, see (Dukes & Buckwalter, 2010)
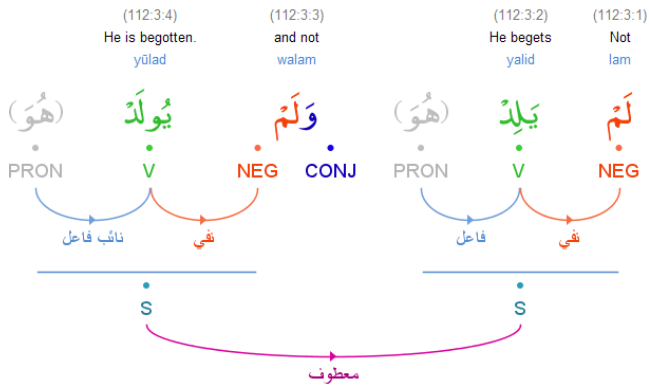
Figure 6: Implicit hidden pronouns.

3.3 Preposition Phrase Attachment

Prepositions are easily identified in Quranic Arabic since they always modify the following noun which will be found in the genitive case. The preposition and its object form a phrase in traditional Arabic grammar known as *jār wa majrūr* (جار ومجرور). A dependency relation named *muta'aliq* is used to annotate preposition phrase (PP) attachment. This relation may be translated as "link" or "attachment". A constraint of the grammar is that a preposition phrase must always be linked to another head node, which is usually either a verb or a noun (Ryding, 2008). Deciding the location of attachment depends on context. Most often a preposition phrase will be attached to its preceding verb, as shown in Figure 7.
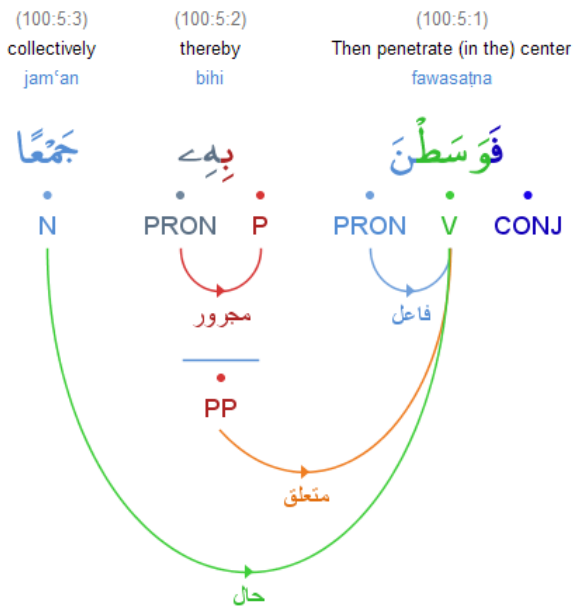

Figure 7: PP-attachment to a verb.

In Arabic, there is no direct equivalent of the English present tense copula verb, and equational sentences (such as "*Mankind are ungrateful to their Lord*") are represented by writing two nouns side-by-side, with both in the nominative case. The first noun will be the subject, and the second noun the predicate. When a preposition phrase is used in an equational sentence, it is typically attached to the predicate.

Certain chapters of the Quran begin with a preposition phrase used as an oath (Rafai, 1998). In this case the preposition will be a particle of oath, usually *wāw*. To satisfy the PP-linking constraint, the preposition phrase will attach to an implicit node such as the hidden verb "I swear by" (see Figure 8). Although a preposition phrase must always be linked to another head node, it not always through attachment. For example, consecutive sequences of preposition phrases may be related through conjunction or through apposition.
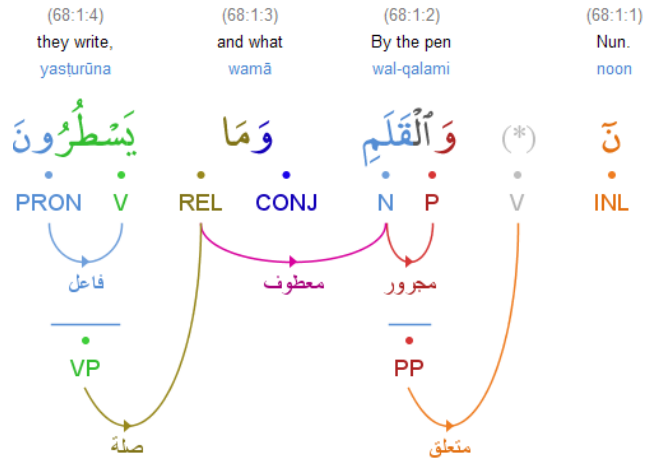

Figure 8: PP-attachment to a hidden node.

4. Conclusion and Future Work

The full annotation guidelines that are presented in this paper are available online at the Quran corpus website (http://corpus.quran.com), to enable online collaborative annotation. The website has attracted a wide variety of visitors including NLP researchers, many non-academics wanting to learn more about the Quran, and interested volunteers who are familiar with the source material and traditional grammar. The aim of traditional Quranic studies is to throw light upon the meanings of the Quranic text. Adopting the grammar framework of *i'rāb* and traditional analytics expertise will lead to an enriched corpus. The markup is not only machine-readable, but can be an aid to human understanding of the Arabic source for non-Arabic speakers.

For example, particles such as *annā* in Figure 1 can be difficult to translate faithfully into other languages. Different English translations of the Quran use "that", "how", "for" or some other construct (Awde & Smith, 2004). The dependency analysis will help readers further in uncovering the detailed intended meanings of each verse and sentence.

As well as morphological and syntactic analysis, a third planned phase of annotation in the corpus will be a semantic layer, following completion of the syntactic treebank. It is hoped the resource will become more directly amenable to computational semantic modeling by annotating the text using semantic role labeling, or by representing semantics using first-order predicate logic.

5. References

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

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

Nicholas Awde and Kevin Smith (2004). Arabic-English/English-Arabic Dictionary. Bennett & Bloom.

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

Kais Dukes and Tim Buckwalter (2010). A Dependency Treebank of the Quran using Traditional Arabic Grammar. 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.

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.

Nizar Habash, Owen Rambow and Ryan Roth (2008). MADA+TOKAN: Quick Manual.

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

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.

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

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.

Ryan Roth, Owen Rambow, Nizar Habash, Mona Diab and Cynthia Rudin (2008). Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking. In Proceedings of the Conference of American Association for Computational Linguistics (ACL08).

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

Abdelhadi Soudi, Antal van den Bosch and Gunter Neumann (2007). Arabic Computational Morphology: Knowledge-based and Empirical Methods. Springer.

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