A Dependency Treebank of the Quran using Traditional Arabic Grammar

Kais Dukes[1] and Tim Buckwalter[2]

[1] *School of Computing, University of Leeds LS2 9JT, United Kingdom*
[2] *University of Maryland, College Park MD 20742,* USA
[1] sckd@leeds.ac.uk
[2] tbuckwal@umd.edu

***Abstract* – The Quran is a significant religious text, followed by the 1.5 billion believers of the Islamic faith worldwide. The text dates to 610-632 CE and is written in Quranic Arabic, the direct ancestor language of modern standard Arabic in use today. This paper presents the Quranic Arabic Dependency Treebank (QADT) and reports on the approaches and solutions used to apply Natural Language Processing to the unique and challenging language of the Quran. This project differs from other Arabic treebanks by providing a deep computational linguistic model based on historical traditional Arabic grammar (إعراب القرآن الكريم). The treebank is part of the Quranic Arabic Corpus (http://corpus.quran.com), a popular free Arabic resource developed at the University of Leeds.**

Motivated by the importance of the Quran as a central religious text, we also report on how online collaborative annotation was used to bring together Quranic scholars and Arabic language experts to ensure a high level of accuracy for grammatical analysis of the entire Quran.

***Keywords* – Treebank, Syntax, Morphology, Quran, Arabic, Corpus, Part-of-Speech Tagging, Dependency Grammar**

I. Introduction

The Quranic Treebank was developed to allow researchers interested in the Quran to get as close as possible to the original Arabic text and understand its intended meanings through grammatical analysis. Because the users of the Quranic Treebank are mostly students of the Quran and Arabic language researchers, a large focus of the project has been on visualization of the grammatical annotation.

The treebank provides two levels of analysis: morphological annotation and syntactic representation. The morphological annotation has been completed and verified, and all of the 77,430 words in the Quran have been divided into constituent morphological segments (figure 1). The treebank also introduces the novel approach of displaying Quranic syntax using dependency graphs, which show how each word in a sentence is related and what role it plays in building up a complete syntactic structure (see figure 2). To date, 2,500 syntactic dependency graphs have been annotated, covering 11,000 words of the Quran.

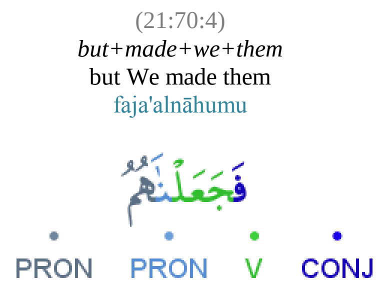

Figure 1: Morphological segmentation of an Arabic word in the Quranic Treebank.

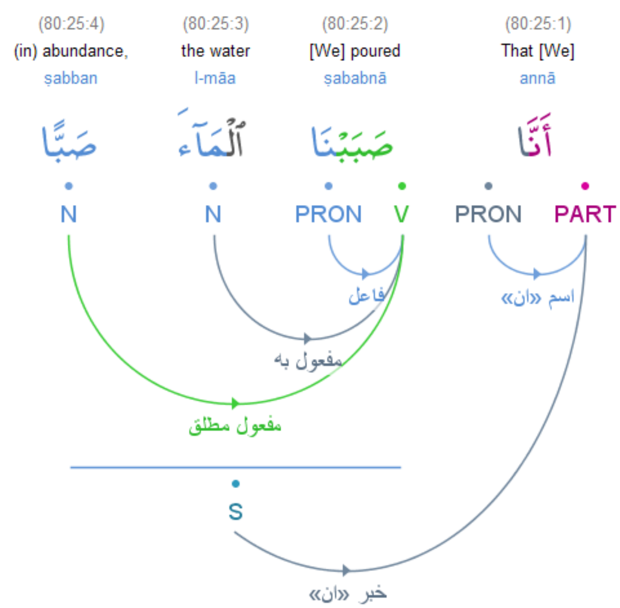

Figure 2: Syntactic dependency graph for verse (80:25) of the Quran.

Previous related work includes the morphological analysis of the Quran performed at the University of Haifa [7], however this automatic processing of the Quranic text was not completed and remains manually unverified with multiple possible analyses for each word. Previous syntactic work includes the three major Arabic treebanks that have been recently developed: the Penn Arabic Treebank [4] [19], the Prague Arabic Dependency Treebank (PADT) [15] [26] and the Columbia Arabic Treebank (CATiB) [12] [14]. Each of these treebanks has a different scope and aim, and each has its own form of representation for modeling Arabic syntax. 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 Arabic treebanks to the Quranic Arabic Dependency Treebank:

Treebank	**Dependency**	**Features**	**Traditional**
Penn	no	yes	no
Prague	yes	yes	no
Columbia	yes	no	yes (subset)
Quran	yes (hybrid)	yes	yes

Figure 3: Comparison of Arabic Treebanks.

The second column in the table above indicates if a treebank uses dependency grammar, which shows the relationships between pairs of words, or if constituent phrase structure is used. The next column indicates if 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 in figure 3 specifies if traditional Arabic syntax is used. Both the Penn and the Prague Arabic Treebanks use complex models of syntactic representation which are not intuitive to native speakers of Arabic, often requiring lengthy training to be able to participate in the annotation effort. The Columbia Treebank uses a limited subset of traditional Arabic grammar which is sufficient for further development of statistical parsing, and also allows for rapid annotation with minimal user training.

The Quranic Treebank extends this approach of using traditional syntax by attempting to represent as much of traditional Arabic grammar as possible. This leads to morphological annotation and dependency graphs which use familiar terminology, and enables anyone who is already experienced with Quranic syntax to immediately participate in the annotation effort. Using traditional grammar also enables the dependency graphs to be verified against the many existing books and publications on Quranic syntax. The grammar of the Quran has been studied and documented in detail for over 1000 years – far longer than for most other languages. In fact, traditional Arabic grammar is widely recognized as one of the origins of modern dependency grammar [18] [22]. This new treebanking project provides for the first time a structured database based on this wealth of existing traditional analysis.

II. A Computational Model of Quranic Arabic

A. Part-of-speech and Dependency Tags

The linguistic model used by the Quranic Arabic Dependency Treebank is divided into three layers: orthography, morphology and syntax. The data is stored in XML format, and a Java object model is provided with the treebank as an API to query the data. For each word in the Quran, the representation in the morphological layer divides the word into its constituent segments, and assigns a part-of-speech tag and an inflection feature matrix to each segment using name-value pairs [11] [27]. Figure 4 below shows how this morphological information is presented online to website users. Natural Language Generation technology (NLG) is used to provide concise English and Arabic summaries of the inflection features stored in the Quranic linguistic database:

(36:61:2)
you worship Me
u'budūnī

V – 2nd person masculine plural imperative verb
PRON – subject pronoun
PRON – 1st person singular object pronoun

فعل أمر والواو ضمير متصل في محل رفع فاعل
والياء ضمير متصل في محل نصب مفعول به

Figure 4: Online display of morphological feature tagging for a word in the Quran.

The Quranic Treebank also includes information not found in other tagged Arabic corpora, which includes the root for each word, a word-by-word interlinear translation into English, and an automatically generated phonetic transcription. All terminology used in the Quranic Treebank and its associated software is taken directly from traditional Arabic grammar (إعراب), and mapped to equivalent English terminology. An attempt has been made to cover as much as possible of the traditional Arabic syntax of the Quran. The online corpus annotation guidelines provide detailed documentation for the tags, inflection features and dependency graph edge labels [8]. In this paper, a summary of key part-of-speech tags and dependency relations is shown in figures 5 and 6 below:

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

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

Figure 5: Part-of-speech tagset for morphological segments (prefixes, stems and suffixes).

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 representative
	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

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

Figure 6: Edge labels for syntactic dependency relations (excluding particle relations).

B. Modeling Traditional Syntax (إعراب) *using Dependency Graphs*

The syntax of traditional Arabic grammar is represented in the Quranic Treebank using hybrid dependency graphs. Graphs are mathematical structures which consist of nodes, and edges which link nodes together. In linguistic terms, a dependency graph is a way to visualize the structure of a sentence by showing how different words relate to each other using directed links called dependencies. The graphs are stored in the XML Quranic linguistic database.

Most relations in the dependency graphs used in the Quranic Treebank will be between terminal nodes. These are nodes which correspond to morphological word segments and will have part-of-speech tags. Since traditional Arabic grammar often describes relations between phrases, a dependency graph may also include non-terminal phrase nodes. It is for this reason that the linguistic framework in the Quranic Treebank is termed a hybrid dependency-phrase structure grammar. Phrase nodes are shown under horizontal blue bars, as can be seen in the graphs throughout this paper. The extent of the bar shows which words or word segments form the phrase. Using phrases is a natural way to relate two groups of words, for example when one clause is connected to another through conjunction. Another good use of phrases is to describe preposition phrase attachment, where the preposition and its genitive noun form a preposition phrase (PP) [3] [17].

Figure 7 below shows a dependency graph that describes the syntax of verse (99:1). The Arabic syntactic constructions annotated in this graph include a conditional clause (شرط), a passive verb subject representative (نائب فاعل), and a cognate accusative (مفعول مطلق). The last dependency on the left is a possessive construction (مضاف إليه) also known as the genitive construction:

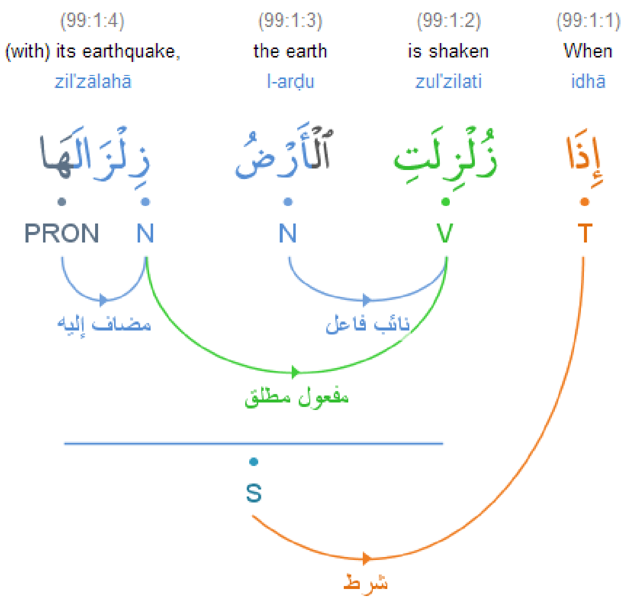

Figure 7: Dependency graph for verse (99:1).

In the Quranic Treebank, the convention used is that all edges in a dependency graph point from a dependent node towards its head node. More than one node can be dependent on the same head node, such as the verb in figure 7. However, each node in a syntactic dependency graph can point directly to at most one other node. This means that a node cannot directly depend on two or more other nodes in the graph, and that each node has at most one unique head. Mathematically, the syntax graphs in the Quranic Treebank are all dependency trees (directed acyclic graphs [6]). The directed graphs also satisfy the property of connectivity, so that each graph will have a single unique root node.

C. Computational Modeling of Elision (حذف) and Reconstruction (تقدير)

Quranic Arabic is a pro-drop language. Certain verbs imply a pronoun subject through inflection and the pronoun can be dropped from the sentence. Traditional Arabic grammar reconstructs these dropped pronouns describing them as implicit *damīr mustatir* (ضمير مستتر). Other parts-of-speech besides pronouns can function as implicit words, depending on the sentence being analyzed. In traditional Arabic grammar, implicit words omitted through elision (حذف) are generally known as *mah'dhūf* (محذوف), and the process of reconstructing a sentence is known as *taqdīr* (تقدير). It is important to note that no new information or meaning is added through the syntactic reconstruction of a sentence. In some sense, reconstruction is a form of "syntactic normalization" which allows implicit syntactic roles to be made explicit.

This reconstruction is performed automatically in the Quranic Treebank through the inflection features tagged for each verb (person, number and gender), and serves a number of purposes. Firstly, it allows for a consistent analysis when comparing against traditional books and existing publications of Quranic syntax. In addition, the further higher levels of linguistic analysis in traditional exegesis require these reconstructed pronouns, the most important being pronoun resolution. This is a future planned feature of the Quranic Arabic Corpus in which pronouns will be mapped to concepts in an ontology, which includes reconstructed dropped pronouns as well as explicit personal and possessive pronouns. Figure 8 below shows an example of this, where the reconstructed implicit pronouns ("He") refer to Allah (God), according to widely accepted traditional commentaries of the Quran, such as *tafsir ibn kathīr*:

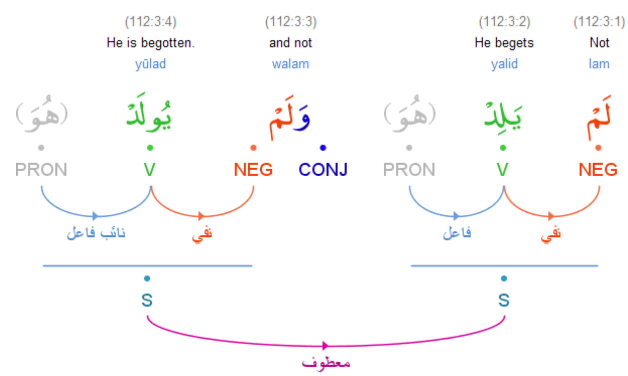

Figure 8: Reconstructed implicit pronouns for verb subjects in verse (112:3).

D.Empty Nodes

In the syntactic analysis for verse (68:1) shown in figure 9 below, an empty node has been introduced as part of the process of traditional syntactic normalization (تقدير). According to a linguistic constraint enforced in the analysis of traditional Arabic grammar, all preposition phrases (PP) must be attached (متعلق) to the head word that they modify. Since this chapter of the Quran begins with a preposition phrase, syntactic normalization is applied so that the PP is attached to an empty node in the graph which represents an implicit verb *mah'dhūf* (محظوف). As such, the implied translation of this verse is the oath "(I swear) by then pen…"

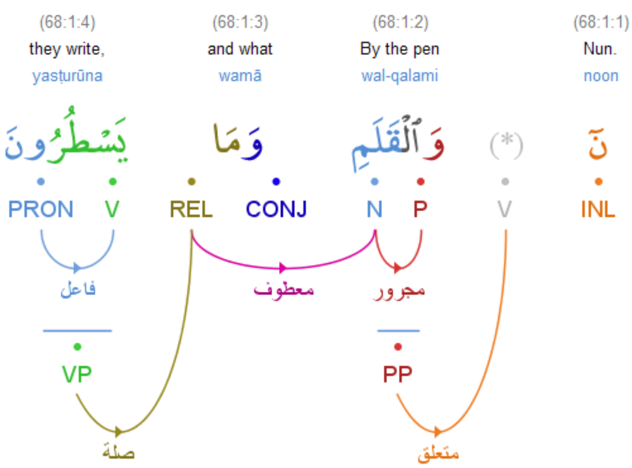

Figure 9: An empty verb node in verse (68:1).

III. Development of the Quranic Arabic Dependency Treebank

A. Java Implementation

The custom linguistic software used in the Quranic Arabic Corpus is implemented using Java, and consists of 75,000 lines of programming code that was developed over an 18-month period. The software is made freely an available as open source under the GNU public license.

Java is suitable for developing such a large code base since it encourages the software to be divided into a distinct set of components (see figure 10 below). The largest component in terms of number of lines of code is the Traditional Grammar Rule Engine. This is a set of approximately 1000 linguistic constraints written as Java rules (20,000 lines of code), which were manually typed up using several traditional Arabic grammar textbooks as a reference [1] [10] [16] [20] [21] [23] [24]. Usually such computational grammar rules are induced from a corpus through statistical analysis of an existing treebank. This was not easily possible for this project given the unique form of Quranic Arabic, and the different choice of syntactic representation compared to existing Arabic treebanks. The rule engine is used for two purposes: validation of existing annotation, and producing new annotation through automatic parsing. Existing morphological and syntactic annotation in the corpus cannot break any of the grammar rules, and so the rule engine provides a useful validation check that ensures the data is consistent, and guards against human annotator error.

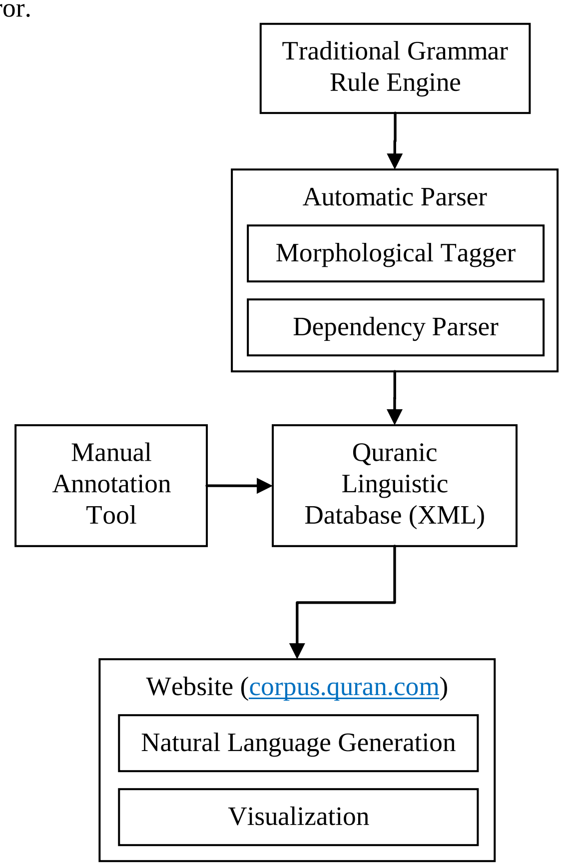

Figure 10: Java components in the Quranic Arabic Corpus software.

The other large Java component is the manual annotation tool, which allows morphological analysis and syntactic dependency graphs to be displayed and manually edited.

B. Automatic Parsing (إعراب) of Quranic Arabic

The Quranic Treebank has only been made possible through the development of recent advances in computational Arabic, and also by the efforts of a small dedicated team of Quranic Arabic experts who diligently reviewed the morphological and syntactic annotation. The grammatical analysis was developed in three stages: automatic parsing, initial offline manual verification, and then online public collaborative annotation.

As shown in figure 10 above, the automatic parser is divided into two subcomponents, a morphological tagger, and a dependency parser. Both of these are based on recent natural language technology. The morphological tagger [9] was derived from the Perl-based Buckwalter Arabic Morphological Analyzer (BAMA) [5] that was used to annotate the Penn Arabic Treebank. All 4 Arabic Treebanks (Penn, Prague, Columbia and The Quranic Treebank) use either the Buckwalter Arabic lexicon directly, or a derivation of it to perform initial automatic morphological analysis [4] [13] [15]. For the Quran, a domain-specific optimization was used. Typically BAMA operates on undiacritized text, and produces a list of possible analyses for each word that requires further manual disambiguation in order to select the correct morphological analysis. However, the Quranic Treebank uses a fully automated morphological analyzer. Since the Quran contains diacritics, edit distance was used as a metric to select the most likely BAMA analysis. In addition, the Traditional Grammar Rule Engine was applied to discard analyses that were not possible on linguistic grounds, by using surrounding words as context. The second subcomponent (the dependency parser) follows Joakim Nivre's parsing algorithm, which came top in the 2007 CoNLL task on dependency parsing. The Quranic parser's algorithm is based on Nivre's. However, syntactic rules are not deduced automatically from a pre-existing treebank, but instead the Traditional Grammar Rule Engine is used to drive parser actions.

Figure 11 shows the F-Measure of both subcomponents. According to this measure of performance, the morphological tagging subcomponent is 77% accurate. The error is mostly due to the fact that the Buckwalter lexicon was intended for modern standard Arabic and does not contain many words found only in the Quran. Another difficulty in adapting the Buckwalter analyzer was differences in spelling, since the Quran is 1,400 years old. This was addressed through the use of pre-processing step before morphological analysis, in order to normalize orthographic and lexical variation (e.g. use of the *alif khanjarīya* diacritic and multiple different uses of the letter *hamza*). The second subcomponent, the rule-based dependency parser covers only 68% of the full grammar of the Quran, but the grammar rules it does know about lead to very accurate automated analysis (91% precision).

Parser Component	**Precision**	**Recall**	**F-Measure**
Morphological Tagger	72%	83%	77%
Dependency Parser	91%	68%	78%

Figure 11: Evaluation of automatic parsing of the Quran (إعراب القرآن الكريم).

C. *Online Collaborative Annotation*

The accuracy figures reported in figure 11 relate to automatic parsing. The accuracy of the current morphological and syntactic annotation presented on the Quranic Arabic Corpus website is much higher, since the work has undergone several stages of manual correction and verification after the initial step of automatic parsing.

Work to improve accuracy is currently ongoing. The Quranic Treebank is verified online via collaborative annotation through volunteer corrections. This is similar to Wikipedia in that anybody can suggest new information. However, final acceptance of grammatical analysis into the corpus annotation requires approval, after validation using the Traditional Grammar Rule Engine, and verifying against published sources of Quranic syntax.

The benefit of adopting a model of collaborative annotation is that many interesting points-of-view can be presented and discussed. For most cases of disagreement, the different opinions are usually resolved through a deeper understanding of the Arabic language and of the Quran. In general, consistent analysis is encouraged by constantly improving the annotation guidelines when exceptional cases such as these are encountered by volunteer annotators. The discussion of such cases takes place publicly on an open message-board forum. Different types of discussion found on the annotator message board include correct part-of-speech, case inflection, gender tagging and differing opinions on accuracy of interlinear translation against traditional sources.

IV. APPLICATIONS

The Quranic Arabic Treebank is useful for further research of the Arabic language and the Quran. The Treebank is based on standard traditional Arabic grammar, and it is freely available for download under an open source license. The linguistic data has been manually verified by multiple annotators, and the linguistic database is machine readable. The following academic research is currently using data from the Quranic Treebank and from the Quranic Arabic Corpus:

A. *Statistical Analysis of the Arabic Language*

Paul Tupper (Simon Fraser University, Canada). This research focuses on inducing Arabic phonotactics [2]. Linguistic data from the Quranic Arabic Treebank is used as part of a project to identify phonetic word patterns in unrestricted Arabic text.

B. *Construction of Arabic Ontologies*

Soraya Zaidi (Algeria and at the University of New Mexico). This research project [28] uses tagged data from the Quranic Corpus to automatically generate an ontology.

C. Knowledge Representation in the Quran

Abdul-Baquee M. Sharaf (University of Leeds). Related work by the Language Research Group at the University of Leeds. This PhD research builds on the Quranic Arabic Treebank as part of a project for further deep annotation and analysis of the Quran, including semantic modeling of the knowledge contained in each verse [25].

D. Information Retrieval for the Quran

Henda Sfaxi (INSAT Research Centre, Tunsia). Linguistically tagged data from the Quranic Treebank is being used to build a search engine for the Quran by adding annotated word forms to an inverted search index.

E. OpenIslam.org

Idris Mokhtarzada (University of Maryland). Open Islam is a research website that uses linguistically annotated data from the Quranic Arabic Corpus. The project provides an online utility for studying Quranic words and their roots, as well as a tool that allows registered users to keep notes on verses of the Quran.

V. Conclusions and Future Work

The Quranic Treebank is part of the larger Quranic Arabic Corpus project, which is 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. Future work on the treebank 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. Complete coverage of the Quran is planned for the syntactic treebank. Currently dependency graphs cover 11,000 out of the 77,430 words in the Quran, although morphological analysis is available for the entire Quran. In addition, we hope to integrate the treebank into other standard tools for computational linguistics and language processing such NLTK as GATE.

References

[1] J. Akesson. *Arabic Morphology and Phonology: Based on the Marah Al-Arwah by Ahmad b. 'Ali Mas'ud.* Brill, 2001.

[2] J. Alderete, P. Tupper, S. A. Frisch. *Phonotactic learning without a priori constraints: A connectionist analysis of Arabic cooccurrence restrictions*, 2009.

[3] H. Ansari. *Learning the Language of the Quran.* Centre of Religious Studies and Guidance, 1997.

[4] A. Bies and M. Maamouri. *Penn Arabic Treebank Guidelines*. http://www.ircs.upenn.edu/arabic, 2003.

[5] T. Buckwalter. *Buckwalter Arabic Morphological Analyzer version 1.0*. Linguistic Data Consortium, University of Pennsylvania, 2002.

[6] M. A. Covington. *A Fundamental Algorithm for Dependency Parsing*. Proceedings of the 39th Annual ACM Southeast Conference, 2001.

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

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

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

[10] W. Fischer and J. Rodgers. *A Grammar of Classical Arabic: Third Revised Edition*. Yale University Press, 2002.

[11] N. Habash. Arabic Morphological Representations for Machine Translation. 2007.

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

[13] N. Habash, O. Rambow and R. Roth. *MADA+TOKAN: Quick Manual*, 2008.

[14] N. Habash and R. Roth. *CATiB: The Columbia Arabic Treebank*. Technical Report, Center of Computational Learning Systems, Columbia University, 2009.

[15] J. Hajic, O. Smrz, P. Zemanek, J. Snaidauf, and E. Beska. *Prague Arabic Dependency Treebank: Development in Data and Tools*. Proceedings of the NEMLAR international conference on Arabic Language Resources and Tools, 2004.

[16] J. A. Haywood and H. M. Nahmad. *A New Arabic Grammar of the Written Language*. Lund Humphries Publishers, 2005.

[17] A. Jones. *Arabic Through the Qur'an.* Islamic Texts Society, 2005.

[18] G. Kruijff. *Dependency grammar*. The Encyclopedia of Language and Linguistics 2nd edition, Elsevier Publishers, 2006.

[19] M. Maamouri, A. Bies and Tim Buckwalter. *The Penn Arabic Treebank: Building a Large-scale Annotated Arabic Corpus.* In NEMLAR Conference on Arabic Language Resources and Tools, Cairo, Egypt, 2004.

[20] E. Muhammad. *From the Treasures of Arabic Morphology.* Zam Zam Publishers, 2007.

[21] A. A. Nadwi. *Vocabulary of the Holy Quran*. Millat Book Centre, 2006.

[22] J. Owens. *The Foundations of Grammar: An Introduction to Medieval Arabic Grammatical Theory*. John Benjamins Publishers, 1988.

[23] J. Rafai. *Basic Quranic Arabic Grammar*. Ta-Ha Publishers Ltd, 1998.

[24] K. C. Ryding. *A Reference Grammar of Modern Standard Arabic*. Cambridge University Press, 2008.

[25] A. M. Sharaf. *Knowledge Representation in the Quran*. Interim Report, University of Leeds, 2009.

[26] O. Smrz and J. Hajic. *The Other Arabic Treebank: Prague Dependencies and Functions*. In Arabic Computational Linguistics: Current Implementations, CSLI Publications, 2006.

[27] A. Soudi, A. Bosch and G. Neumann. *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer, 2007.

[28] S. Zaidi. *Review of Textual Terminology Tools for Ontologies Building*, MIC'09 Management International conference, Sousse, Tunisia, 2009.