LAMP: A Multimodal Web Platform for Collaborative Linguistic Analysis

Kais Dukes and Eric Atwell

I-AIBS Institute for Artificial Intelligence and Biological Systems,
School of Computing, University of Leeds, United Kingdom.
E-mail: sckd@leeds.ac.uk, e.s.atwell@leeds.ac.uk

Abstract

This paper describes the underlying software platform used to develop and publish annotations for the Quranic Arabic Corpus (QAC). The QAC (Dukes, Atwell and Habash, 2011) is a multimodal language resource that integrates deep tagging, interlinear translation, multiple speech recordings, visualization and collaborative analysis for the Classical Arabic language of the Quran. Available online at http://corpus.quran.com, the website is a popular study guide for Quranic Arabic, used by over 1.2 million visitors over the past year. We provide a description of the underlying software system that has been used to develop the corpus annotations. The multimodal data is made available online through an accessible cross-referenced web interface. Although our Linguistic Analysis Multimodal Platform (LAMP), has been applied to the Classical Arabic language of the Quran, we argue that our annotation model and software architecture may be of interest to other related corpus linguistics projects. Work related to LAMP includes recent efforts for annotating other Classical languages, such as Ancient Greek and Latin (Bamman, Mambrini and Crane, 2009), as well as commercial systems (e.g. Logos Bible study) that provide access to syntactic tagging for the Hebrew Bible and Greek New Testament (Brannan, 2011).

Keywords: Arabic Corpus, Treebank, Quran, Collaborative Annotation

1. Introduction

Over the last several decades, the development and use of annotated corpora has grown to become a major focus of research for both linguistics and computational natural language processing. Annotated corpora provide the empirical evidence that is used to advance various theories of language (Sampson and McCarthy, 2005). Annotated data is also used by computational linguists to engineer state-of-the-art natural language systems and resources including electronic lexicons (Kucera and Francis, 1967; Hajic, et al., 2007; Brierley and Atwell, 2008), part-of-speech taggers (Leech, et al., 1983; Brants, 2000; Spoustová et al., 2009; Søgaard, 2011) and syntactic parsers (Atwell, et al., 1984; Collins, 1999; Charniak, 2000; Nivre, et al., 2007).

Two of the main challenges that arise when developing annotated corpora are the often prohibitively high costs required to manually construct the annotated data, as well as the extra effort required to make this data easily available and accessible. This is especially important for popular texts that are studied in detail by members of the general public, such as central religions texts. The fact that texts such as the Quran, Hebrew Bible and Greek New Testament are written in ancient classical languages and are of wide public interest makes them an important dataset for linguistic annotation and natural language processing.

The Quranic Arabic Corpus was developed through a model of collaborative annotation, where volunteers proofread part-of-speech tagging and syntactic analysis, and then suggest corrections through an online message-board forum (Dukes, Atwell and Habash, 2011). Figure 1 shows an example sentence annotated using a hybrid dependency-constituency syntactic tagging scheme. The first line indicates the chapter, verse and word numbers from the Quran (sequenced from right to left, in line with the Arabic script), followed by an interlinear word-by-word translation into English. Part-of-speech tags are assigned to the morphemes that form each word. Finally, syntactic dependencies between words and phrases are annotated using Arabic syntactic dependency labels. The annotated corpus also includes further additional tagging, including named entity references, an ontology of semantic concepts and an automatically generated phonetic transcription. For further details on the full annotation scheme used in the Quranic Arabic Corpus, see (Dukes et al., 2011).

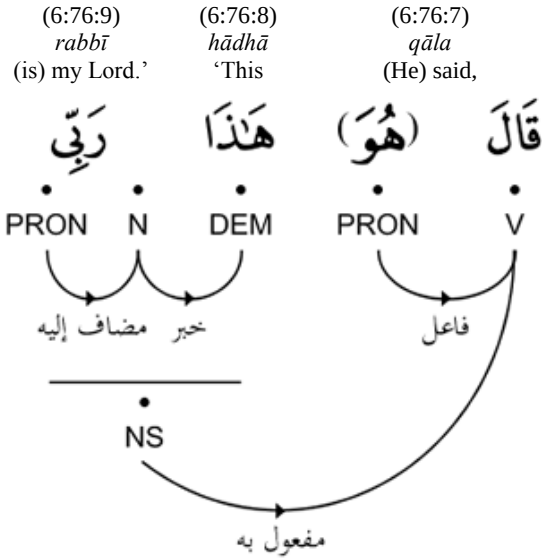

Figure 1: Morphological and syntactic annotation for the Quranic verse: *He said, 'This is my Lord.'*

Related efforts for other classical languages such as Greek and Latin (Bamman, et al., 2009) have also involved collaboration between Classical linguists and computer scientists to develop tagged and parsed corpus resources, to be

used by Classical scholars and Computational Linguistics researchers. However, the Quranic Arabic Corpus is different in attracting much wider interest from Muslims worldwide wanting to use the resource to help them understand and appreciate the Quran. This has enabled us to benefit from several hundred volunteers who have cross-checked the annotation in the QAC against historical works based on the traditional Arabic grammar known as *i'rāb* (إعراب) (Salih, 2007). To encourage volunteer annotation, and to provide a high quality resource, accessibility and usability of the Quranic Corpus were key concerns in its design. The resource is intended not only as a computational dataset, but also as a learning and study guide for the Quran. In order to meet the two challenges of dealing with large scale volunteer annotation and building a highly accurate, useful study and educational tool, we decided to develop a custom software architecture to annotate the QAC, and decided to make the annotations freely available online.

2. Linguistic Analysis Multimodal Platform

Although our LAMP architecture has been used successfully for the Quranic Arabic Corpus website, we argue that the technology concepts we present are a general model of online corpus development, applicable to related datasets. In order to support a reusable design and efficiently handle a large number of users, LAMP is implemented as a set of layered Java modules. Figure 2 shows the three layers used in LAMP: a multimodal database, a set of computational-linguistic components, and an online web interface. These three layers are used for data storage, data processing, and data access respectively.

3. Multimodal Linguistic Data

The multimodal database stores annotated text in XML format, as well as media files for audio and video, and scanned manuscripts that have not yet been converted into text using OCR. The most important data is the Arabic text of the Quran itself. However, given that the Quran is primarily an oral tradition, the database also includes seven audio recitations of the Quran. These different recitations reflect different readings of the text, each with subtle differences in prosodic stress. In addition to speech recordings, the latest version of the corpus (version 0.5) is planned to include video recordings of Quranic recitation as part of the multimodal data

The database also stores seven parallel translations of the Quran into English, as well as an interlinear translation (see http://corpus.quran.com/wordbyword.jsp). Whereas Christians usually read the Bible translated into their native language, by tradition all Muslims are encouraged to study and appreciate the Quran in its original Classical Arabic form. The seven verse-by-verse English translations and the interlinear word-by-word literal translation are popular and useful aids for the large proportion of readers who speak English as a first or second language.

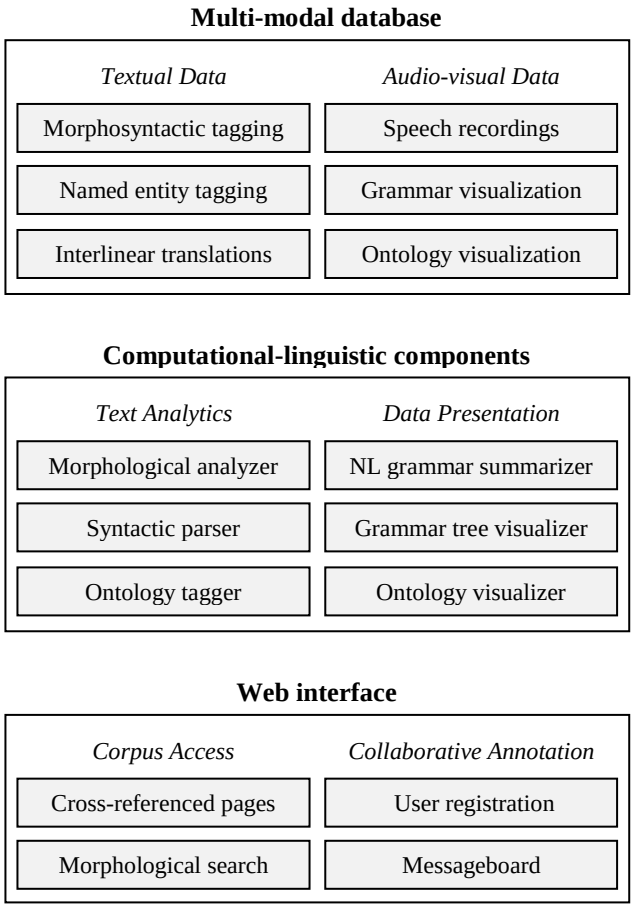

Figure 2: LAMP architecture diagram, illustrating the three layers of the software design.

All of the multimodal data has been aligned at sentence-level. For each 6,236 verses in the Quran, a hyperlinked page on the website gives access to the original Arabic script, audio, video and English translations. The multimodal data is also hyperlinked to annotations for morphological segmentation, part-of-speech tagging and syntactic analysis (Dukes and Buckwalter, 2010), which reference historical works of Quranic grammar. Multimodal data is especially useful for an educational study guide such as the QAC, to help engage the student, and to allow the student to choose from a range of different viewpoints of the core text

To efficiently allow the data to be quickly accessed online, several indexes have been added to the database. For fast web search, the Quranic text has been indexed by keywords for English, and by morphological stems and roots for Arabic. In addition, the Quranic script has been indexed by chapter, verse and word number.

For displaying online, the database stores snapshots of automatically generated visualizations of annotated treebank sentences and ontological data. The multimodal database is not static. As linguistic annotations are improved through new user suggestions, updated tags are stored together with snapshots of updated visualizations for the corrected sentences in the corpus.

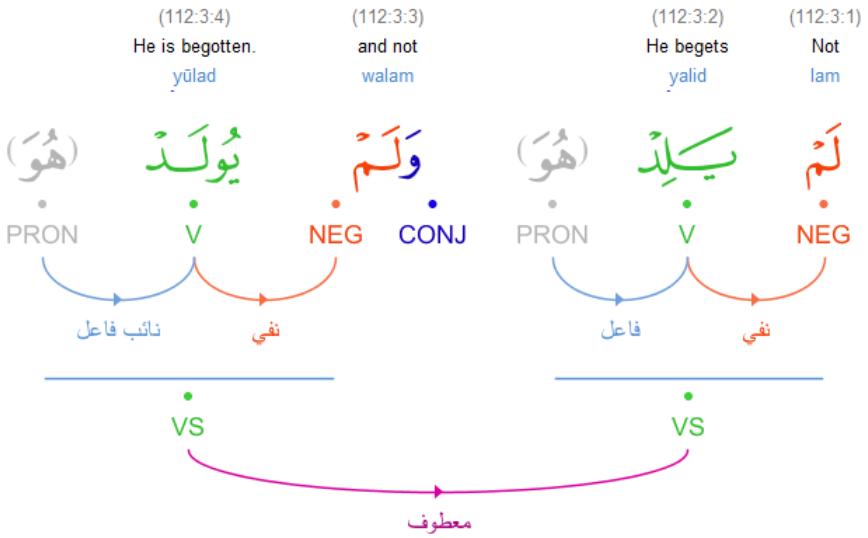

Figure 3. Graphical color-coded representation of treebank syntax in the Quranic Arabic Corpus.

4. Computational Linguistic Components

The second layer in LAMP is a set of computational linguistic components for both analysis of Quranic text, as well as tools that make use of the resulting annotated data. A morphological analyzer (Dukes and Habash 2010) and a syntactic parser (Dukes and Habash, 2011) were designed specifically for the Classical Arabic language of the Quran and have been used for initial development of the corpus annotations, before online proofreading by volunteers.

To produce the graphical representation of syntax in the treebank (Figures 1 and 3), we use custom visualization algorithms to illustrate the sentence structure of Arabic dependency grammar through color-coded links between words. These diagrams are hyperlinked to the rest of the annotated data (Dukes, Atwell and Sharaf, 2010a).

Of possible interest to related projects is our use of natural language generation (NLG). The database stores POS tags and syntactic disambiguation as a sequence of abbreviated machine-readable tags. We apply tools to publish annotations as concise grammatical summaries. For example, the Classical Arabic word *fafataqnāhumā* (ففتقناهما) may be translated as '*then we parted them both'*. A grammatical description is produced using NLG (Figure 4).

> 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* (ف ت ق). The suffix (نا) is an attached subject pronoun. The attached object pronoun is third person dual.

Figure 4: Automatic natural language generation of grammatical summaries using morphosyntactic tagging.

5. Quranic Arabic Corpus Website

The third and final layer of LAMP is the web interface. We publish the data stored in the multimodal database online and make use of the linguistic software tools to enrich the data and make it more easily accessible. Each word in the Quran has several layers of annotation including phonetic transcription, morphological segmentation, part-of-speech tagging and syntactic analysis.

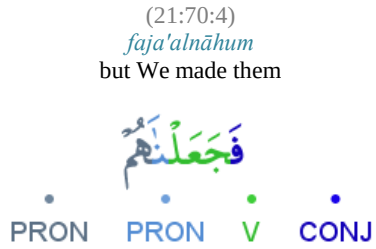

Figure 5. Linguistic 'breakdown' of an Arabic word in the Quran showing multiple levels of annotation.

Each of these layers is shown on a single 'word breakdown' webpage with multiple hyperlinks to further detailed information (see Figure 5). The QAC website also offers a morphological search feature, not available on other related Quranic study sites. This feature is made possible by the richly annotated data stored in the linguistic database.

For example, the word ذهب in Classical Arabic has two readings, as either the noun 'gold' or the verb 'go'. Searching using by POS tag and root, the occurrences of the correct reading can be easily found in the corpus.

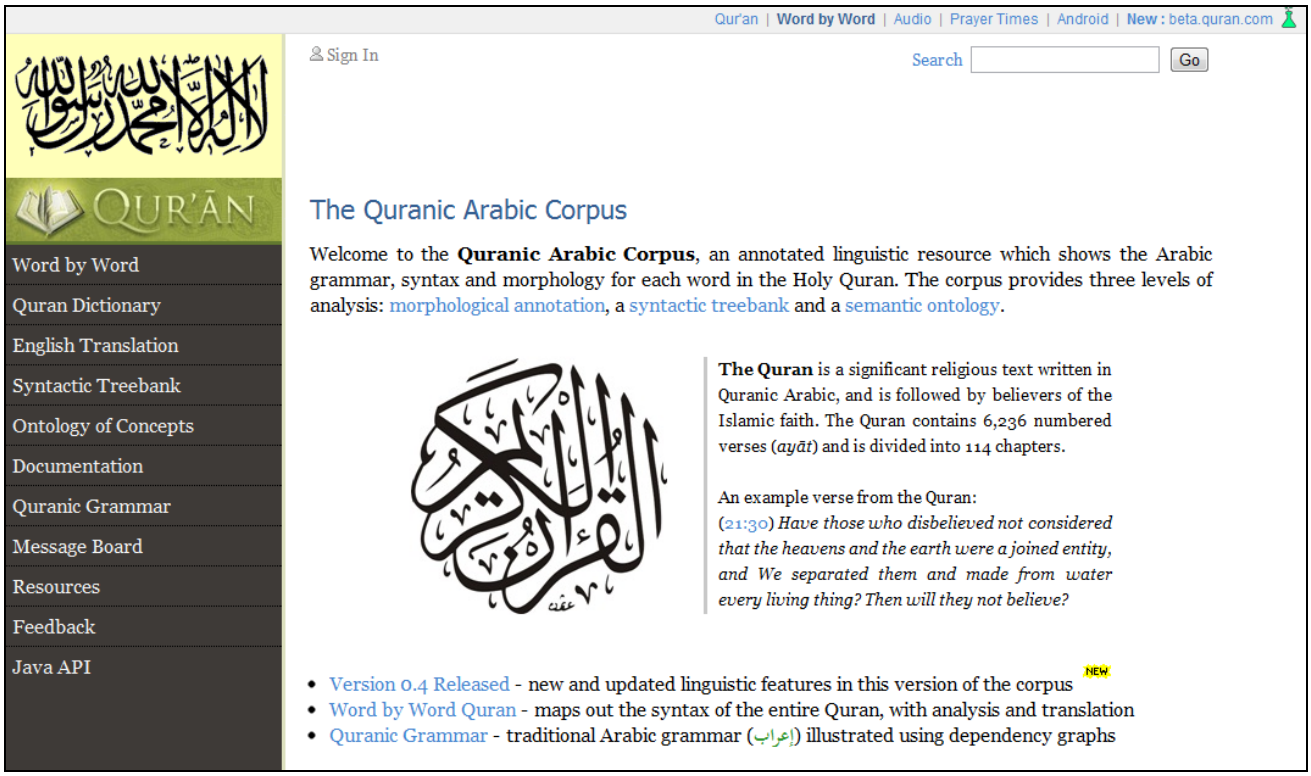

Figure 6. The Quranic Arabic Corpus website (http://corpus.quran.com).

The Quranic Arabic Corpus website also supports collaborative annotation. While browsing the annotated corpus, users are encouraged to submit corrections and suggestions for improvements. Our model for online annotation is supervised collaboration. The website includes a set of annotation guidelines that new visitors are encouraged to read (Dukes, Atwell and Sharaf, 2010b). If the online annotation does not match the guidelines, or appears to be in error, corrections can be suggested using an online message board. A few select supervisors, chosen for consistently making high-quality corrections, review the suggestions on the message board and discuss these with other members of the website. If consensus is reached on a correction or improvement, the database is updated with the new tagging (Dukes, Atwell and Habash, 2011).

Figure 6 shows the homepage of the Quranic Arabic Corpus The left-hand navigation pane lists QAC features in order of popularity of use: Word by Word, Quran Dictionary, English Translations, Syntactic Treebank, Ontology of Concepts, Documentation, Quranic Grammar, Message Board, Resources, Feedback, Java API.

5.1 Word by Word

The Quranic Arabic Corpus started out as a research project in morphological and syntactic analysis and annotation of the source Arabic text, and we added the word-by-word English translation initially to accommodate non-Arabic-speaking AI research collaborators. However, it soon became clear that this is the most popular 'annotation level' for non-research users, as it gives a more literal access to the source Arabic words to members of the general public who speak English as a first or second language. Furthermore, in the feedback section of the website, a frequently-requested extension to the current website is the addition of word-by-word translations into other languages, such as French, German, Urdu, Malay.

5.2 Quran Dictionary

The user can select an Arabic root, and see the inflected and derived forms used in the Quran (see Figure 7).

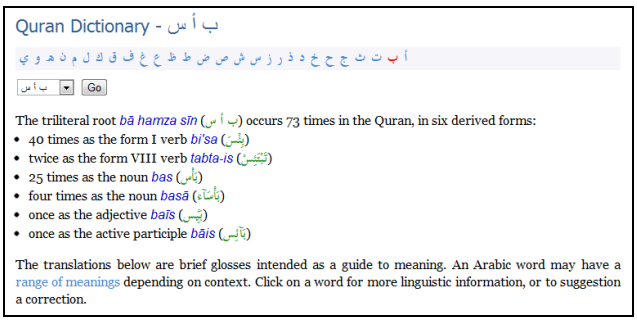

Figure 7. Quran Dictionary: root *bā hamza sīn* (ب أ س) occurs 73 times in the Quran, in six derived forms.

The Quran dictionary also offers:

- Verb Concordance: a list of verbs grouped by root and form, and sorted by frequency.
- Lemma Frequency list: list of lemmas split by part-of-speech and sorted by frequency, where lemma groups word-forms that differ only by inflectional

(as opposed to derivational) morphology, and do not vary in meaning.

- Morphological Search: search for words in the Quran by specifying part-of-speech, and/or morphological features, and/or root or lemma or stem.

5.3 English Translations

For a chosen chapter and verse, this displays the Arabic and seven widely-known English interpretations: Sahih International; Pickthall; Yusuf Ali; Shakir; Muhammad Sarwar; Mohsin Khan; Arberry. Click on the Arabic text to see the word-by-word detailed literal translation and morphological analysis. The English Translation page also links to 11 alternative Recitations, and Dependency graph - syntactic analysis (*i'rāb*) for the verse.

5.4 Syntactic Treebank

For a selected chapter and verse, this shows the graphical color-coded representation of treebank syntax, as illustrated in Figures 1 and 3. Users can also see the Arabic grammar description (إعراب) for the verse, adapted from the grammatical analysis provided by the Quran Printing Complex.

5.5 Ontology of Concepts

The Quranic Ontology uses knowledge representation to define the key concepts in the Quran, and shows the relationships between these concepts using predicate logic. The fundamental concepts in the ontology are based on the knowledge contained in traditional sources of Quranic analysis, including the *Hadīth* of the prophet Muhammad , and the *Tafsīr* (Quranic exegesis) of Ibn Kathīr. Named entities in verses, such as the names of historic people and places mentioned in the Quran, are linked to concepts in the ontology as part of named entity tagging. An overview diagram shows a visual representation of the ontology: the graph is a network of 300 linked concepts with 350 relations.

As well as listing the major concepts in the Quran, the ontology also defines a set of semantic relations between these concepts. The most important relation is the set membership relation '*instance*' in which one concept is defined to be an instance or individual member of another group. For example the relation 'Satan is a jinn' in the ontology would represent the knowledge contained in the Quran that the individual known as *Satan* belongs to the set of sentient creations named the *jinn*. Other concepts in the ontology are grouped into logical categories, according to the properties that they share. For example, the Sun, Earth and Moon are classified under 'Astronomical Body'.

In the Word by Word view, pronouns are hyperlinked to concepts in the ontology in order to resolve anaphoric reference. For example, verse 97:1 literally means 'we revealed it'; but through traditional Quranic exegesis (*Tafsīr Ibn Kathīr*) we know that this verse refers to Allah revealing the Quran. The analysis shows these referents:

- PRON – subject pronoun → *Allah*.
- PRON – 3rd person masculine singular object pronoun → *Quran*.

Named entities found in the Arabic text of the Quran are also linked to concepts in the ontology. Ontology users can also browse a Concept Map, showing each concept, its definition, subcategories, related concepts, location in the visual map, and predicate logic relations with subclasses and instances. There is also a Topic Index: click on a concept in the list to see a summary of the topic, and a list of all occurrences of that concept in the Quran.

5.6 Documentation

The non-expert reader can benefit from detailed tutorial-style explanations of:

- Quranic script: phonetic transcription, pause marks, verse marks.
- Morphology: part-of-speech tags, morphological features.
- Syntax: dependency graphs, syntactic relations, phrase tags.

5.7 Quranic Grammar

The Morphology and Syntax tutorials in the overview Documentation include many cross-reference hyperlinks to more detailed descriptions of Quranic Grammar categories and concepts. The Grammar section of the website provides a set of guidelines for annotators who wish to contribute to the project. This description of Quranic grammar is also useful for further computational analysis, as well as for linguists researching the language of the Quran, and for those with a general interest in the Arabic language. The Quranic Grammar documentation has five main subsections:

5.7.1 The Syntax of Nominals

The nominals are one of the three basic parts-of-speech according to traditional Arabic grammar. These include nouns, pronouns and adjectives. The following sections describe the syntax of nominals:

- Gender: semantic, morphemic and grammatical gender.
- Adjectives: these follow and depend on the noun that they describe.
- Possessives: the possessive construction of *iḍāfa* (إضافة) is used with the genitive case.
- Apposition: two nouns placed side by side, both with the same syntactic function.
- Specification: *tamyīz* (تمييز) specifies the degree of a head word.
- Numbers: the *murakkab* (مركب) dependency is used to annotate digit compounds.

5.7.2 Verbs, Subjects and Objects

The verbs form the second of the three basic parts-of-speech. The corpus documentation describe the syntax of verbs in the Quran, as well as case rules for subjects and objects of verbs:

- Verb forms: the different forms of verbs found in Quranic Arabic.
- Subjects and objects: these will inflect for different cases according to syntactic function.
- The verb *kāna*: a special group of verbs with different case rules.
- Verb moods: the subjunctive and jussive moods of the imperfect.
- Imperative verbs: commands, requests and negative prohibitions using the imperfect jussive.

5.7.3 Phrases and Clauses

Traditional Arabic grammar defines a set of dependencies linking different types of phrases and clauses:

- Preposition phrases: these use the genitive case and can attach to nouns or verbs.
- Coordinating conjunctions: these connect two words, phrases or clauses.
- Subordinating conjunctions: together with relative pronouns these introduce subordinate clauses.
- Conditional sentences: formed of two clauses, the condition and the result.

5.7.3 Adverbial Expressions

The accusative case ending *mansūb* is used in various grammatical constructions, which include adverbial expressions and objects:

- Circumstance: the circumstantial accusative (حال).
- Cognate accusative: the *mafūl muṭlaq*.
- Accusative of purpose: *l-mafūl li-aj'lihi*.
- Comitative objects: *l-mafūl ma'ahu*.

5.7.4 The Syntax of Particles

The particles are the third of the three basic parts-of-speech in traditional Arabic grammar. The following annotation guidelines discuss common syntactic constructions involving particles:

- The particle *alif*: interrogative and equalizational uses of *hamza*.
- The particle *inna*: a special group of particles with their own case rules.
- The particle *fa*: conjunction, resumption and cause particles.
- Vocative particles - these can place a noun into one of two grammatical cases.
- Exceptive particles - may place a noun into the accusative case according to the type of exception.

5.8 Message Board

This is for volunteers to review the linguistic annotations in the corpus: '…If you come across a word and you feel that a better analysis could be provided, you can suggest a correction online by clicking on an Arabic word.' At the time of writing, 1068 messages (and subsequent threads) are still under discussion; and 5229 resolved comments on morphological and syntactic tagging have been archived.

5.9 Resources

A wide range of resources are made available for other researchers, including publications, bibliography, data download, release notes, and mailing list.

5.9.1 Publications

Academic research publications: a list of research articles and papers by Quranic Arabic Corpus researchers; and a list of citations, as well as references to the Quranic Arabic Corpus in other research papers.

- Newspaper reviews: such as an interview with Kais Dukes in The Muslim Post discussing the research; and a review in the University of Leeds Reporter newsletter.
- Blog reviews: such as Examiner.com review of the website and discussions of how related techniques might apply to other texts.

5.9.2 Bibliography

A comprehensive list of textbooks and other scholarly resources consulted in developing the Quranic Arabic corpus, with a summary of the contribution of each book, including:

- Textbooks and references on Arabic grammar for English-speakers.
- English translations of the Quran.
- Dictionaries of the Quran.
- Quran websites and online resources.
- Arabic grammar resources.

5.9.3 Data Download

This allows researchers to download the Quranic Arabic Corpus morphological data, if they agree to the terms and conditions of the GNU General Public License, and to the terms of use: permission is granted to copy and distribute verbatim copies of this file, but changing it is not permitted. Annotation can be used in any website or application, provided its source (the Quranic Arabic Corpus) is clearly indicated, and a link is made to http://corpus.quran.com to enable users to keep track of changes. The copyright notice is required to be included in all copies of the text.

5.9.4 Release Notes

A summary of improvements since the previous release; the current version 0.4 includes: increased coverage for the syntactic Treebank, to 40% of the Quran; revised morphological analysis taking account of over 500 feedback comments; improved Quran dictionary and lemmatization, with concordance lines from Quranic verses as context; readability and navigation improvements; more accurate tagging of proper nouns, with new named entities added to the semantic ontology; more accurate tagging for particles *wāw* and *fa*. Version 0.4 of the morphologically annotated corpus is freely available for download from the website.

5.9.5 Mailing List

We maintain a mailing list for Quranic computing academic researchers, to discuss related research issues; the website has a hyperlink to an archive of past discussions.

5.10 Feedback

The Quranic Arabic Corpus development has been guided by over 300 user comments, including suggestions for additional features, and feedback on a wide range of uses for academic research and study of the Quran. There are also subsections for Frequently Asked Questions (with answers), contacts, and acknowledgements of contributors to the project: academic collaborators, annotation proofreaders and supervisors, and sources of verified Arabic text of the Quran, MP3 audio files, and English translations.

5.11 Java API.

The Quranic Arabic Corpus includes a set of Java APIs for accessing and analyzing the Holy Quran, in its authentic Arabic form. The Java library is released as an open source project, in order to encourage computational analysis of the Quran. We invite others to contribute to this project, with the long term aim of providing a publicly available set of computational tools for linguistic analysis of the Quran in Arabic. The Java API is organized into three parts: The Quranic text itself, a set of *access* APIs, and a set of *analysis* APIs. The distinction between accessing and analyzing the Quran is that access is concerned with representing the Arabic text (e.g. chapters, verses, letters and diacritics), whereas the analysis API is built on top of this, providing more sophisticated tools for computational linguistics.

The Java developer's guide includes an overview, and specific notes on: Uthmani script; orthography model; simple encoding; Unicode serialization; transliteration; analysis table; search API; and examples of code using the Java API. There is also standard Java API documentation generated for each individual Java package and class; and for advanced users, notes on orthography internals, build instructions, and test coverage reports.

6. Conclusion and Future Work

The Quranic Arabic Corpus is a popular annotated linguistic resource, with over 1.2 million visitors over the past year. Developing the corpus annotations, and making them easily accessible online was aided by our development of LAMP, our Linguistic Analysis Multimodal Platform.

Our plans for future work include extending the existing Quranic Arabic Corpus with further annotations, but also making LAMP available for download as a separate framework. We hope that other corpus linguistics projects will find the platform useful, particularly for annotation research on texts which attract wider public interest, such as religious texts, or popular works of literature.

Although most tagging efforts result in machine readable resources, software tools such as visualization and natural language generation can be used to make annotated data more easily accessible. For widely used and studied texts such as the Quran, an accessible and easy-to-use website is essential to engage the general public and encourage the large number of interested visitors to suggest corrections and participate in collaborative annotation.

7. References

Eric Atwell, Geoffrey Leech, Roger Garside. (1984). Analysis of the LOB Corpus: progress and prospects. In Aarts, J, Meijs, W (Eds), *Corpus Linguistics: Proceedings of the ICAME 4th International Conference on the Use of Computer Corpora in English Language Research*. Amsterdam: Rodopi. pp40-52

David Bamman, Francesco Mambrini and Gregory Crane. (2009). An Ownership Model of Annotation: The Ancient Greek Dependency Treebank. In *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories*. Milan, Italy.

Rick Brannan. (2011). Why Another Greek New Testament? In *Bible Technologies Conference (BibleTech)*. Seattle, Washington.

Thorsten Brants. (2000). TnT: A Statistical Part-of-Speech Tagger. In *Proceedings of Applied Natural Language Processing Conference*. Seattle, Washington.

Claire Brierley, Eric Atwell. (2008). ProPOSEL: A prosody and POS English lexicon for language engineering. In: *Proceedings of LREC'08: Language Resources and Evaluation Conference*

Eugene Charniak. (2000). A Maximum-Entropy-Inspired Parser. In *Proceedings of the 1st Annual Meeting of the North American Chapter of the ACL (NAACL)*. Seattle.

Michael Collins. (1999). Head-driven Statistical Models for Natural Language Parsing. *Ph.D. Thesis, University of Pennsylvania*.

Kais Dukes, Eric Atwell and Nizar Habash. (2011). Supervised Collaboration for Syntactic Annotation of Quranic Arabic. To appear in *Language Resources and Evaluation Journal (LREJ): Special Issue on Collaboratively Constructed Language Resources*.

Kais Dukes, Eric Atwell and Abdul-Baquee Sharaf. (2010a). Online Visualization of Traditional Quranic Grammar using Dependency Graphs. *The Foundations of Arabic Linguistics Conference*. Cambridge.

Kais Dukes, Eric Atwell and Abdul-Baquee Sharaf. (2010b). Syntactic annotation guidelines for the Quranic Arabic Dependency Treebank. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*. Valletta, Malta.

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

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

Kais Dukes and Nizar Habash. (2011). One-step Statistical Parsing of Hybrid Dependency-Constituency Syntactic Representations. In *Proceedings of the International Conference on Parsing Technologies (IWPT)*. Dublin, Ireland.

Jan Hajic, Jarmila Panevová, Zdenka Urešová, Alevtina Bémová, Veronika Kolárová and Petr Pajas. (2003). PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories*. Växjö, Sweden.

Henry Kucera and Nelson Francis. (1967). Computational Analysis of Present-Day American English.

Goeffrey Leech, Roger Garside, Eric Atwell. (1983) The Automatic Grammatical Tagging of the LOB Corpus *ICAME Journal of the International Computer Archive of Modern English* Vol.7.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel and Deniz Yuret. (2007). The CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of EMNLP-CoNLL*.

Bahjat 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.

Geoffrey Sampson and Diana McCarthy. (2005). Corpus Linguistics: Readings in a Widening Discipline. *Continuum*.

Anders Søgaard. (2011). Semi-Supervised Condensed Nearest Neighbor for Part-of-Speech Tagging. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*. Portland, Oregon.

Drahomíra Spoustová, Jan Hajič, Jan Raab and Miroslav Spousta. (2009). Semi-supervised Training for the Averaged Perceptron POS Tagger. In *Proceedings of the 12th EACL Conference*. Athens, Greece.