An Artificial Intelligence Approach to Arabic and Islamic Content on the Internet

Conference Paper · January 2011

DOI: 10.13140/2.1.2425.9528

5 authors, including:

Eric Atwell
University of Leeds
370 PUBLICATIONS **2,295** CITATIONS
SEE PROFILE

Majdi Sawalha
University of Jordan
46 PUBLICATIONS **321** CITATIONS
SEE PROFILE

An Artificial Intelligence approach to Arabic and Islamic content on the internet

Eric Atwell, Claire Brierley, Kais Dukes, Majdi Sawalha, Abdul-Baquee Sharaf

I-AIBS Institute for Artificial intelligence and Biological Systems,

School of Computing, University of Leeds, Leeds LS2 9JT, UK

Keywords *Artificial Intelligence, Corpus Linguistics, Arabic, Quran, Knowledge, Resources*

Abstract: *We review a range of Artificial Intelligence and Corpus Linguistics research at Leeds University on Arabic and the Quran, which has produced a range of software and corpus datasets for research on Modern Standard Arabic and more recently Quranic Arabic .Our work on Quranic Arabic corpus linguistics has attracted widespread interest, not only from Arabic linguists but also from Quranic students, and the general public. We see a great potential impact of Artificial Intelligence modelling of the Quran. This leads us to present a proposal for further research: the Quranic Knowledge Map.*

Introduction

The Natural Language Processing research group, part of the Institute for Artificial Intelligence and Biological Systems (I-AIBS) at Leeds University, has been involved in research on Arabic natural language processing and corpus linguistics for over a decade. Our early work focused on tools and corpus resources for analysis and modeling of Modern Standard Arabic. More recently, we have worked with Quranic Arabic. We view the Quran as a rich data-set for Artificial Intelligence and Machine Learning research.

The Quran is held by Muslims to be a single-authored text, the direct words of God (Allah), conveyed by the angel Gabriel to Mohammed 1355-1378 years ago, and later transcribed verbatim to be used as the sole authoritative source of knowledge, wisdom and law. A challenge for Artificial Intelligence researchers is to represent this knowledge, wisdom and law in computer systems: to build intelligent systems which can answer any question with knowledge from the Quran, and can help society, both Muslim and non-Muslim, to understand and appreciate the Quran.

Artificial Intelligence research on language and text is generally based on a Corpus, a machine-readable data-set of the text being researched, enriched with metadata and tags or annotations showing morphological analyses, Part-of-Speech tags, etc. Examples include the LOB Lancaster-Oslo/Bergen Corpus of one million words of British English texts from published written sources such as newspapers and books; the British National Corpus of 100 million words from both written and spoken sources, for example London teenager conversations; and the Leeds Arabic Internet Corpus of c170 million words of Arabic texts harvested from WWW Arabic websites. The Quran in comparison is a concise data-set, a text of less than 80,000 words, sequenced in chapters and verses. Muslims hold that the original data format was spoken Classical Arabic, captured faithfully in a sophisticated transcription system: it was vital to transcribe accurately the exact value and relative location of every consonant, vowel, and pause, every morpheme, affix and clitic, every word, verse, and chapter. Access to the Quran has traditionally been through the text: many Muslims learn to memorise and recite the verbatim data-set. Access to the underlying knowledge, wisdom and law

requires interpretation and inference; much knowledge is encoded via subtle use of words, grammar, allusions, links and cross-references. For over a thousand years, scholars have sought to extract knowledge and laws from the text, and have built up a much larger Tafsir or corpus of analyses, interpretations and inference chains. Computer Science and Artificial Intelligence presents the opportunity to re-analyse the text data, enrich the Quranic Arabic text with annotations capturing the linguistic structure, and from this extract and capture the underlying knowledge in a Knowledge Representation and Reasoning formalism, to enable automated, objective inference and querying.

Other religions also have defining books, for example the Christian Bible, and the Jewish Tanakh, which could also be amenable to Artificial Intelligence analysis and modelling. The Quran makes a good first case study as it is more concise and more homogeneous. The Bible and Tanakh are larger collections of texts by a range of authors over a longer period with a variety of literary genres including allegories, historical narrative, poetry, genealogy, and explicit exposition of various types of law; whereas the Quran is a single text, widely accepted as the work of a single author, in a consistent genre and style. Consequently, the Tafsir or canon of Quranic interpretations is narrowly focussed on this smaller core text, whereas the body of commentaries on the Bible and Tanakh is broader and less homogenous. Of course, lessons learnt in computing research on the Quran as a first case study could later be extended to the Bible, Tanakh, and other religious texts.

Artificial Intelligence research at Leeds University on Arabic and the Quran

An early survey of Arabic corpus linguistics tools (Atwell et al 04) found few publicly-available Arabic language computing resources; but we found that Machine Learning from a suitable Corpus could be used to adapt generic NLP techniques to Arabic (Abu Shawar and Atwell 04, 05). This required an Arabic text training set, so we developed the first freely-available open-source Corpus of Contemporary Arabic (Al-Sulaiti and Atwell 06), and Arabic concordance visualisation toolkit (Roberts et al 06). The Corpus of Contemporary Arabic has been widely re-used in Arabic NLP research, for training and evaluation of systems. We also developed tools for Modern Arabic text analytics: morphological analysis, stemming, and tagging (Sawaha and Atwell 08, 09, 10a), broad-coverage Arabic lexical resource (Sawalha and Atwell 10b) and a Discourse Treebank for Modern Standard Arabic akin to the Penn English Discourse Treebank (Al-Saif and Markert 10).

We are pioneering NLP research on the Quran, extending our text analytics techniques to Classical Arabic, including chatbot development (Abu Shawar and Atwell 04); conceptual search tool for the Qur'an based on a Quranic scholar's index of the Quran (Abbas 09); formal knowledge representation (Sharaf and Atwell 09); morphological analysis and Part-of-Speech tagging of Quran verses (Dukes et al 11, Sawalha and Atwell 11), syntactic annotation showing grammatical dependency structure of verses (Dukes and Habash 10, Dukes et al 10); text-mining and machine-learning to classify Quran chapters and verses (Sharaf and Atwell 11). Our Quranic Arabic Corpus website http://corpus.quran.com/ has become a widely-used resource, not just by Arabic and Quranic researchers, but by general public wanting online tools to explore and understand the Quran. This has led us to propose "Understanding the Quran" as a new

Grand Challenge for Computer Science and Artificial Intelligence for 2010 and beyond, to the British Computer Society and Association for Computing Machinery conference on the future of Computing research (Atwell et al 10). The Quranic Arabic Corpus is an online annotated linguistic resource which shows the Arabic morphological features for each word of the Qur'an (Dukes and Habash 10). This corpus is being extended by incorporating traditional Arabic grammar analysis of the Quran (I'rab) as machine-readable dependency Treebank (Dukes and Buckwalter 10, Dukes et al 10). Also, work is undergoing towards extracting concepts and named-entities in the Qur'an, establishing ontological links and relationships among these concepts and resolving pronominal anaphoric references to these concepts. Further future planned research include extending these works to include a WordNet type resource for the Qur'anic nouns, adjectives and verbs, a FrameNet type of semantic frames, a discourse corpus for the Qur'an and an enhanced powerful search tool for the Qur'an enabling search over various linguistic, stylistic and conceptual terms. Our intention is to extend our NLP analysis to other classical Arabic texts - like Hadith and early scholar's works, for example to collate a corpus of Arabic texts fundamental to Islamic Finance, and codify these in a linguistic Knowledge Representation formalism to enable formalised querying and retrieval. We have also been asked to consider extending our corpus-based analytic approach to defining books of other religions, for example the Christian Bible, and the Jewish Tanakh, which could also be amenable to Artificial Intelligence analysis and modelling. Lessons learnt in computing research on the Quran as a first case study could later be extended to the Bible, Tanakh, and other religious texts, and will help us to formalise similarities and differences between these knowledge sources.
Our resources are open-source rather than commercial; this is why they have been widely re-used, compared to resources kept "in-house" by other Arabic NLP research groups. Our Quranic Arabic Corpus website http://corpus.quran.com/ shows the advantages of making resources open-source : publications, press articles, Message Board for feedback, Google Analytics visualisation of global distribution of visitors to the website.

Potential impact of Artificial Intelligence modelling of the Quran

Information Retrieval and Information Extraction systems already exist, enabling Quranic scholars to access the Quran texts on a number of existing websites; so why is Understanding the Quran a grand challenge for Artificial Intelligence?

1) Understanding Islam is a major societal issue:

- In Western media and WWW, one of the commonest collocations of "Islamic" is "terrorist", fuelled by and fuelling conceptions that Islam is a threat. Western schools, universities and the general public need an objective, impartial online Quran Expert to learn about Islam and understand its implications for society.

- Some non-Arabic-speaking Muslims may also be ignorant of the deeper meanings in the Quran, despite memorising the sounds of the verses. An impartial online Quran Expert could help them explore and understand the deeper teachings of the Quran for themselves.

2) Current systems can, in principle, answer "factoid" questions from the source text, like "Are all angels male?"; but many potential questions are more difficult and

contentious to answer via text-match, requiring a new Knowledge Representation and Reasoning formalism capable of capturing complex, subtle knowledge encoded in the Classical Arabic, and inferencing in new ways which mirror the thousand-year–old traditions of scholarly analysis and interpretation.

3) In principle, we could use any book as training data for Knowledge Extraction research. The Quran stands out as the source of a large collection of analysis and interpretation texts, known as Tafsir, which could provide a Gold Standard "ground truth" for AI knowledge extraction and knowledge representation experiments. An additional sub-challenge is to encode the Tafsir interpretations in our Knowledge Representation formalism, so we can cross-check for compatibility and consistency with knowledge extraction results from the Quran corpus. In principle, we should aim to be able to reproduce computationally every sound inference and interpretation in the Tafsir. We may find some computational results are incompatible with specific Tafsir inferences and/or conclusions, which will shed new light on traditional interpretations. We may also find some new computational results which are not in any Tafsir interpretation, thus adding to the canon of Islamic wisdom.

4) The system will be used and relied on by billions of Muslims, and also billions of non-Muslims who want to understand the influence of Islam around the world. This is not just an issue of systems scalability and robustness, since the likes of Google and Yahoo etc can already handle huge volumes of queries. It is also vital that answers are always "logically consistent and correct" in that they are consistent with and supported by evidence from the source text, and have a demonstrable chain of inference. We accept that Google sometimes gets wrong matches between text keywords and the concepts we are looking for, and we can live with "acceptable error rates"; but for Quran users, any false inference may be unacceptable.

5) Existing Quran websites offering limited search and analysis are already popular with scholars and the general public; for example, the Quranic Arabic Corpus research project has been snowed under by volunteer contributors. This makes Understanding the Quran an ideal vehicle for research in computer-supported collaborative working. The potential demand / market for an online Quran Expert is huge, a flagship achievement for Computer Science which will capture the public imagination.

In summary, Understanding the Quran is a grand challenge for society, for western public education, for Muslim-world education, for knowledge representation and reasoning, for knowledge extraction from text, for systems robustness and correctness, for online collaboration. Understanding the Quran is a major new Grand Challenge for Computer Science and Artificial Intelligence.

A proposal for further research: the Quranic Knowledge Map

We want to extend research on Arabic and Islamic content on the internet. One way forward is a proposal for a challenging computational project. We propose the Quranic Knowledge Map - a structured large-scale online resource for understanding the Quran, the religious text of Islam. This would be both a machine-readable structured database of linguistic and semantic information to enable further research, as well as a highly useful educational website. Such a resource would be of interest to the many people

wanting to learn Arabic, and to members of the general public wanting to know more about Islam and the Quran. In addition, The Quranic Knowledge Map would also be of interest to professional Arabic linguistics, computational linguists, and students of the Quran and Quranic researchers. In addition, by advancing the existing methodology for construction of similar systems, a set of reusable tools is planned for computational linguistics, knowledge engineering and information retrieval researchers.

As with any other major religious text, the Quran contains knowledge and information, both spiritual and philosophical in nature, as well practical guidelines. The challenge for a large-scale computing project is to represent this knowledge and information as an online structured computational resource, enabling further advanced applications. This would include not only a detailed structured linguistic and semantic database, but also an intelligent system capable of using these datasets to answer questions related to knowledge contained in the Quran. Based on experiencing constructing a smaller-scale linguistic resource, the Quranic Arabic Corpus, there is a huge online demand for access to high-quality Quranic knowledge.

Accessing Quranic Knowledge

Information retrieval and information extraction systems already exist, so why is understanding the Quran a large-scale computational challenge? Understanding Islam is a major societal issue. In Western media and WWW there are many negative collocations of Islam, fuelled by and fuelling conceptions that Islam as a religion is negative. Western schools, universities and the general public need an objective, impartial online Quranic Knowledge Map to learn about Islam and understand its implications for society.

Some non-Arabic-speaking Muslims may also be ignorant of the deeper meanings in the Quran, despite memorizing the sounds of the verses. An impartial online Quranic Knowledge Map could help them question and understand the teachings of the Quran for themselves. Current systems can, in principle, answer "factoid" questions from the source text, like "Are all angels male?"; but many potential questions are more difficult and contentious to answer via text-match, requiring a new knowledge representation and reasoning formalism capable of capturing complex, subtle knowledge encoded in the Classical Arabic text, and inferencing in new ways which mirror the thousand-year–old traditions of scholarly analysis and interpretation.

In principle, we could use any book as training data for knowledge extraction research. The Quran stands out as the source of a large collection of analysis and interpretation texts, known as *tafsir*, which could provide a gold standard "ground truth" for AI knowledge extraction and knowledge representation experiments. An additional sub-challenge is to encode the Tafsir interpretations in our Knowledge Representation formalism, so we can cross-check for compatibility and consistency with knowledge extraction results from the Quran corpus. In principle, we should aim to be able to reproduce computationally every sound inference and interpretation in the *tafsir*. We may find some computational results are incompatible with specific *tafsir* inferences and/or conclusions, which will shed new light on traditional interpretations. We may

also find some new computational results which are not in any *tafsir* interpretation, thus adding to the canon of Islamic wisdom.

Existing Quranic websites are hugely popular online, with the top sites having millions of regular daily users, mostly from those of the 1.5 billion Muslims worldwide who are also internet-enabled. The Quranic Knowledge Map would be accessible by both Muslims and non-Muslims who want to understand the influence of Islam around the world. This is not just an issue of systems scalability and robustness, since the likes of Google and Yahoo etc can already handle huge volumes of queries. It is also vital that answers are always "correct" in that they are consistent with and supported by evidence from the source text, have a demonstrable chain of inference. We expect Google to sometimes get it wrong, and we can live with "acceptable error rates"; for Quran users, any false inference may be unacceptable.

Understanding the Quran is a grand challenge for society, for western public education, for Muslim-world education, for knowledge representation and reasoning, for knowledge extraction from text, for systems robustness and correctness, and for online collaboration. We propose the construction of the Quranic Knowledge Map to address the fact that understanding the Quran is a major new grand challenge for computer science and artificial intelligence.

Development of the Quranic Arabic Corpus

The Quranic Arabic Corpus is a sequence of planned linguistic datasets organized at the University of Leeds, but developed through online collaborative annotation worldwide. To date, morphological and syntactic datasets have been released, and work is in progress to develop a semantic ontology of the Quran. Although much deeper annotation of the Quran is planned, these existing datasets have been proven to be hugely popular. The Quranic Arabic Corpus website attracts 50,000 users per month, each making use of the detailed linguistically annotated Quran – the first of its kind.

These datasets were created not only by developing reusable natural language processing tools for Classical Arabic, but by leveraging the large and interested set of dedicated volunteer annotators who wish to contribute to the project. Hundreds of volunteer experts have collaborated online to develop a highly accurate linguistically annotated Quranic text that has proven to be highly useful as an educational resource, and also for further understanding the detailed intended meanings of the Quran itself.

Positive feedback has been received from thousands of members of the general public, as well as dozens of leading academics across the fields of computational linguistics and religious studies. These institutions include Harvard, Stanford, Pennsylvania, University of Texas at Arlington, and dozens of others. For a brief summary of widespread interest in project from the general public as well as many academics, see: http://corpus.quran.com/feedback.jsp. As an illustrative example of positive feedback, the Center for Middle Eastern Studies at Harvard is interested in applying the same novel techniques used to develop the Quranic Arabic Corpus to the Hebrew Torah:

> I'm extremely impressed with the Qur'an project. Your implementation of Qur'anic grammar alongside the text and the resulting interlinear format will be a major tool for researchers and instructors everywhere.

Existing Quran websites offering limited search and analysis are already popular with scholars and the general public; for example, the Quranic Arabic Corpus has been snowed under by volunteer contributors. This makes Understanding the Quran an ideal vehicle for research in computer-supported collaborative working. The project has received much positive feedback from academics with regards to the novel uses of collaborative annotation:

> The work reported here is exceptional and many of the approaches and solutions of the project are really worth noting for the research community.

The potential demand and market for the Quranic Knowledge Map is huge, a flagship achievement for Computer Science which will capture the public's imagination. It is expected that the lessons learnt in developing the existing Quranic Arabic Corpus will provide useful and relevant experience for the proposed much larger Quranic Knowledge Map.

Modular Design

The Quranic Knowledge Map aims to include and extend the existing Quranic Arabic Corpus, although is much larger in scope and aim. The existing Quranic Arabic Corpus consists of datasets for syntax and morphology, as well as initial research tools for natural language processing of classical Arabic, a genre with includes the Quran as well as the related Islamic knowledge contained in the Hadith (canonical sayings of the Prophet Muhammad). The Quranic Arabic Corpus will be extended to include a variety of additional datasets covering different levels of linguistic and semantic representation. In addition to data, the project includes software infrastructure, and a planned set of highly useful and relevant end-user applications.

A key methodology in the development of the Quranic Knowledge Map is the approach of building a structured sequence of related modules, which come together to form the final project. The three vertical columns in Figure 1 organize the modules around the three main contributions of the project:

- Infrastructure. A set of tools used to develop the Quranic Knowledge Map, but tools that should be easily reusable and customizable for related work. These include Natural Language Processing tools, tools for online collaborative annotation, and tools for knowledge engineering and automated reasoning.
- Datasets. A sequence of independent yet related datasets is planned for the Quranic Knowledge Map. These structured databases will encode in the knowledge contained in the Quran in a machine-readable way, with each dataset focusing on a different level of annotation. Each of these datasets is expected to be highly useful for further research and worthy in publication and distribution in itself.
- End-user applications. These form the main contribution of the Quranic Knowledge Map to society, i.e. to interested members of the general public who will use and access the system.

The three horizontal layers in Figure 1 reflect the fact that the Quranic Knowledge Map is inter-disciplinary in nature. Modules are organized into these three layers to reflect

the different nature of the modules at each level. Although there is expected to be overlap between the expertise used to develop each module, the three broad fields of concern to the Quranic Knowledge Map are:

- Core Text Annotation: Analyzing, processing, parsing and tagging Classical Arabic.
- Further Linguistics: Deep understanding of the orthography, syntax, morphology, grammar and lexicon of the Quran.
- Quranic Knowledge: Knowledge Representation, Automated Reasoning and Question Answering

Infrastructure

Developing the infrastructure modules is a software engineering challenge. A robust set of reusable components is required to develop and make use of annotated datasets, as well as to support the set of proposed online user applications. These infrastructure modules are planned to not to be used for the Quranic Knowledge Map, but also be designed and implemented in such a way so as to be easily reusable in further related projects.

Classical Arabic Natural Language Processing

An Arabic Natural Language Processing Toolkit is proposed to provide a unified framework for a NLP tools for Arabic, specifically designed initially for Classical Arabic (the genre of the Hadith and the Quran) as opposed to Modern Standard Arabic, although much overlap is expected. These tools will include a robust morphological analyzer and syntactic parser. The existing research prototypes developed for the Quranic Arabic Corpus will provide useful context and background experience for these modules. There is currently a lot of demand and interested in linguistic processing tools for Arabic, but tools for the genre of Quranic Arabic remained relatively unexplored.

Collaborative Annotation

A major focus of the Quranic Knowledge Map will be collaborative annotation. Natural language processing tools which include sophisticated machine learning and statistical analysis will be used to develop the initial versions of datasets. Given the sensitive nature of the Quran as a religious text it is essential that all data is rigorously checked and verified against acceptable sources. Infrastructure for online collaborative annotation is required to allow interested volunteer users to view and amend data. It is expected that unlike a completely open system such as Wikipedia, the Quranic Knowledge Map will be open for review, but final corrections will only be incorporated into the datasets following expert review by trusted members of the online collaborative community. This model has proven to be very efficient and effective for developing the existing datasets in the Quranic Arabic Corpus. The online collaborative annotation framework is expected to be directly reusable in future related projects.

	INFRASTRUCTURE	DATASETS	APPLICATIONS
	Reusable Computational Tools	**The Quranic Arabic Corpus**	**Online User Access**
CORE TEXT ANNOTATION	**Software for Classical Arabic NLP** Arabic Morphological Analyzer Arabic Syntactic Parser Arabic Natural Language Toolkit Multi-lingual Word Alignment	**Quranic NLP Datasets** Quranic Arabic Text Morphological Tagging Syntactic Treebank Translations & Audio	**Baseline Quranic Resource** Online Tagged Quran Morphological Search Quranic Grammar Annotations Interlinear Translations
FURTHER LINGUSTICS	**Tools for Collaborative Annotation** Lingustic Database Manual Annotation Tools Online Collaborative Annotation	**Quranic Lingustic Datasets** Pronoun Resolution Named Entity Resolution Quranic WordNet	**Quranic & Arabic Linguistics** Electronic Lexicon & Dictionary Word-sense Disambiguation Arabic Educational Resources
QURANIC KNOWLEDGE	**Knowledge Representation** Semantic Annotation Framework Semantic Database QA & Information Retrieval Automated Reasoning and Inference	**Quranic Knowledge Datasets** Ontology of Quranic Concepts Quranic PropBank / FrameNet Related Texts: The Hadith Knowledge Representation	**Quranic Knowledge Online** Concept Topic Map & Search Verse similarly concordance Quran to Hadith Linkage Question Answering

Figure 1. Modules in the Quranic Knowledge Map.

.

Automated Reasoning using Artificial Intelligence

Intelligent components will be required to support automated reasoning, inference and question answering. Such components are not necessarily only applicable to the Quran, and should work with any related knowledge encoded using the same semantic formalism.

Datasets

Although comprehensive, developing the following datasets is realistic since the Quran is a closed domain, consisting of fewer than 80,000 words of Classical Arabic.

- The existing Quranic Arabic Corpus Datasets will be enriched and extended to include deep morphological and syntactic tagging of the Quran.
- Quranic WordNet is a well-defined project that will produce a detailed linked and categorized lexicon of the Quran, including word senses and dictionary definitions in both Arabic and English.
- An ontology of Quranic concept is planned which will provide a formal categorization of all the concrete and abstract concepts listed in the Quranic text. Each conceptual element in the ontology will be assigned a unique identifier. This will allow the tasks of pronoun resolution and named entity tagging to easily map to the relevant concepts in the ontology.
- Further annotation of the Quran will include semantic representation. Quranic PropBank or Quranic FrameNet would leverage the established annotation methodologies of related projects, and apply these to the Quran.
- A deep semantic dataset is planned to cover subtle knowledge representation in Quranic Verses.
- The datasets need not only be restricted to the Quran. The related corpus of Hadith provides a valuable context around the Quranic verses. Extended the Quranic Arabic Corpus to include Hadith is an often request feature by users of the existing Quranic Arabic Corpus.

Each of these datasets is expected to be encoded using modern XML standards, and will include high-quality documentation. These will provide much needed resources for further research, not only for the Quran, but also for systems wanting to use the Quran as a "gold standard" for building further systems that process Arabic.

Applications

Each of the above datasets is expected to be accessible through an online interface, including advanced search features. This will lead to a detailed online annotated Quran, allowing users unprecedented access to a wealth of structure information including not only the Quranic text itself, but related grammatical and linguistic information.

Proposed online applications include:

- Online Tagged Quran
- Morphological Search
- Quranic Grammar Annotations
- Interlinear Translations
- Quranic & Arabic Linguistics
- Electronic Lexicon & Dictionary
- Word-sense Disambiguation

- Concept Topic Map & Search
- Verse similarly concordance
- Quran to Hadith Linkage
- Automated Question Answering

These are not only educational aids useful for study and teaching, but also provide a means for the general public to quickly and easily find relevant information directly in the Quran related to Islam, as well as contextual information surrounding the online information.

Collaborators to develop the Quranic Knowledge Map

Given its inter-disciplinary nature, developing the Quranic Knowledge Map will require contributions from different backgrounds. This would include people with the following expertise:

- Management and coordination: leading researchers acting in a supervisory capacity, ideally with all or most of the relevant inter-disciplinary skills, or least with access to relevant experts when needed
- Religious studies experts, with relevant experience in the Quran and exegesis (*tafsir*); and also experts in other religious texts
- Full time annotators, familiar with Arabic and the Quran
- NLP people, with good proven Arabic language computation skills
- Software engineers for general infrastructure and web development, not necessarily with NLP skills
- E-learning experts, ideally with a background in developing Arabic online language-learning or religious educational resources

In addition, it is expected that the project will leverage a large body of existing expert volunteers worldwide through collaborative annotation - a proven methodology used to develop the Quranic Arabic Corpus, if the right framework and infrastructure has been put in place.

These collaborators from a range of disciplines are required for success. An outline of the research project in terms of research work packages is the following:

WP1 Project Management

WP2 Design:

2.1 User requirements analysis

2.2 Design and specification

WP3: Implementation

3.1 Online collaboration framework

3.2 Morphological and syntactic taggers

3.3 Tagset design: morphosyntactic and dependency tags

3.4 Interaction and visualization

WP4: Annotation: tagging and proofreading

WP5: Validation and User Evaluation: Case Studies

WP6: exploring applications in Artificial Intelligence research

6.1 Machine Learning of annotations, to tag other related texts

6.2 Learning similarity, links and bridging

WP7: e-learning customization

References

Abbas, Noorhan. 2009. Qurany 'Search for a Concept' Tool and Website. Unpublished thesis, School of Computing, University of Leeds. Website: http://quranytopics.appspot.com/

Abu Shawar, Bayan; Atwell, Eric. 2004. *An Arabic chatbot giving answers from the Qur'an* in: Bel, B & Marlien, I (editors) **Proc TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles**, Fez, Morocco.

Abu Shawar, Bayan; Atwell, Eric. 2005. *Using corpora in machine-learning chatbot systems*. **International Journal of Corpus Linguistics**, vol. 10, pp. 489-516.

Al-Saif, Amal; Markert, Katja. 2010. *The Leeds Arabic Discourse Treebank: Annotating Discourse*
Connectives for Arabic. in **Proc LREC'2010: Language Resources and Evaluation Conference**, Valetta, Malta.

Al-Sulaiti, Latifa; Atwell, Eric. 2006. *The design of a corpus of contemporary Arabic*. **International Journal of Corpus Linguistics**, vol. 11, pp. 135-171.

Atwell, Eric; Al-Sulaiti, Latifa; Al-Osaimi, Saleh; Abu Shawar, Bayan. 2004. *A review of Arabic corpus analysis tools* in: Bel, B & Marlien, I (editors) **Proc TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles**, Fez, Morocco.

Atwell, Eric; Al-Sulaiti, Latifa; Sharoff, Serge. 2009. *Arabic and Arab English in the Arab World* in: **Proc CL2009 International Conference on Corpus Linguistics**, Liverpool, England.

Atwell, Eric et al. 2010. *Understanding the Quran: a new Grand Challenge for Computer Science and Artificial Intelligence* in **Proc GCCR'10 Grand Challenges in Computing Research for 2010 and beyond**, Edinburgh, Scotland.

Dukes, Kais; Atwell, Eric; Sharaf, Abdul-Baquee. 2010. *Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank* in **Proc LREC'2010: Language Resources and Evaluation Conference**, Valetta, Malta.

Dukes, Kais; Habash, Nizar.. 2010. *Morphological Annotation of Quranic Arabic*. in **Proc LREC'2010: Language Resources and Evaluation Conference**, Valetta, Malta.

Dukes, Kais; Buckwalter, Tim. *A Dependency Treebank of the Quran using Traditional Arabic Grammar*. In **Proc 7th International Conference on Informatics and Systems.** Cairo, Egypt.

Roberts, Andrew; Al-Sulaiti, Latifa; Atwell, Eric. 2006 *aConCorde: Towards an open-source, extendable concordancer for Arabic*. **Corpora journal**, vol. 1, pp. 39-57

Sawalha, Majdi; Atwell, Eric. 2008. *Comparative evalreligionsuation of Arabic language morphological analysers and stemmers* in: **Proc COLING'2008 22nd International Conference on Computational Linguistics**, Manchester, England.

Sawalha, Majdi; Atwell, Eric. 2009. *Linguistically Informed and Corpus Informed Morphological Analysis of Arabic* in: **Proc CL2009 International Conference on Corpus Linguistics**, Liverpool, England.

Sawalha, Majdi; Atwell, Eric. 2010a. *Fine-Grain Morphological Analyzer and Part-of-Speech Tagger for Arabic Text* in **Proc LREC'2010: Language Resources and Evaluation Conference**, Valetta, Malta.

Sawalha, Majdi; Atwell, Eric. 2010b. *Constructing and Using Broad-Coverage Lexical Resource for Enhancing Morphological Analysis of Arabic* in **Proc LREC'2010: Language Resources and Evaluation Conference**, Valetta, Malta.

Sharaf, Abdul-Baquee; Atwell, Eric. 2009. *A Corpus-based Computational Model for Knowledge Representation of the Quran* in: **Proc CL2009 International Conference on Corpus Linguistics**, Liverpool, England.