Unifying linguistic annotations and ontologies for the Arabic Quran

Noorhan Abbas, Luluh Aldhubayi,
Hend Al-Khalifa, Zainab Alqassem,
Eric Atwell, Kais Dukes, Majdi Sawalha,
Abdul-Baquee Muhammad Sharaf
University of Leeds and King Saud University
language@comp.leeds.ac.uk

Current Web/text search allows search by keywords. Semantic Web research aims to allow search for “concepts”, which may not be just character-strings in a document. This involves adding semantic **tags** to each document, which encode the semantic concepts in the document which users are likely to want to find. The set of concept-tags and the relationships between these concepts is called an ontology: the ontology maps the “meaning-space” of a document. For the Quran, a range of different ontologies have been developed:

- each word has been tagged with lexical semantic concept tags: in other words, each word has been morphologically analyzed, to enable search for key features of word-types (for example, search for all imperative verbs, which show explicit instructions). At least three different morphological analysis schemes or tag-sets have been applied to the Quran (Dror et al 2004; Dukes et al 2011; Sawalha and Atwell 2013). So, a challenge at this level is to combine or unify the alternative sets of morphological feature tag-sets.

- each verse has been annotated with concept semantic tags, to enable search for all verses sharing a given concept (for example, search for all verses relating to the Day of Judgement). The set of “concepts” applying to Quran verses is more open to debate than the set of morphological features for Classical Arabic words; at least three different verse-level ontology concept-sets have been applied to the Quran (Al-Khalifa et al 2009, 2010; Abbas 2009; Dukes and Atwell 2012). So another challenge is to combine or unify the sets of verse-level concepts into a single ontology.

- semantic tags can show patterns or links, such as links between pronouns and their antecedents and/or conceptual entity referents (Sharaf and Atwell 2012b), for example, to find all references to Mohammed, even when not named explicitly; and links between related verses (Sharaf and Atwell 2012a), for example, to find all verses similar to a given verse. These links are another aspect to be integrated into a unified Quran ontology.

So, we need to unify rival alternative concept tag-sets at each level; and to unify the concept-sets across levels. A further challenge is that different projects have adopted different representations or formats for the annotated text; such as CSV text-file, or XML, or HTML, or other database format. We need to identify a common format for the different linguistic annotations and ontology mark-ups to map onto. We have begun by mapping some of the annotations into SketchEngine format. SketchEngine is a widely-used tool for corpus linguistics research, so the format has been tested and reused widely. Also, uploading a unified Quran resource to SketchEngine makes it widely accessible for corpus linguistics research.

References

Abbas, N. 2009. Quran 'Search for a Concept' Tool and Website. Unpublished thesis, University of Leeds.

Al-Khalifa, H., Al-Yahya, M. Bahanshal, A. and Al-Odah I. 2009. SemQ: A Proposed Framework for Representing Semantic Opposition in the Holy Quran using Semantic Web Technologies. The 2009 International conference on the Current Trends in Information Technology (CTIT’09), Dubai, UAE.

Al-Khalifa, H, Al-Yahya, M, Bahanshal, A, Al-Oud, I, and Al-Helwa, N. 2010. An Approach to Compare Two Ontological Models for Representing Quranic Words. the 12th International Conference on Information Integration and Web-based Applications & Services (iiWAS2010), Paris, France.

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

Dukes, K; Atwell, E. 2012. LAMP: A Multimodal Web Platform for Collaborative Linguistic Analysis. LREC'2012 Language Resources and Evaluation Conference. Istanbul, Turkey.

Dukes, K; Atwell, E, Habash, N. 2011. Supervised Collaboration for Syntactic Annotation of Quranic Arabic. Language Resources and Evaluation Journal, pp.1-30.

Sawalha M, Atwell E. 2013. A standard tag set expounding traditional morphological features for Arabic language Part-of-Speech tagging. Word Structure Journal, to appear.

Sharaf, A; Atwell, E. 2012a. QurSim: A corpus for evaluation of relatedness in short texts. LREC'2012 Language Resources and Evaluation Conference. Istanbul, Turkey.

Sharaf, A; Atwell, E. 2012b. QurAna: Corpus of the Quran annotated with Pronominal Anaphora. LREC'2012 Language Resources and Evaluation Conference. Istanbul, Turkey.