UNIVERSITY OF LEEDS

This is a repository copy of *Classical and modern Arabic corpora: Genre and language change*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/131245/

Version: Accepted Version

Book Section:

Atwell, E orcid.org/0000-0001-9395-3764 (2018) Classical and modern Arabic corpora: Genre and language change. In: Whitt, RJ, (ed.) Diachronic Corpora, Genre, and Language Change. Studies in Corpus Linguistics, 85 . John Benjamins , pp. 65-91. ISBN 9789027201485

© John Benjamins 2018.This is an author produced version of a book chapter published in Diachronic Corpora, Genre, and Language Change. Uploaded in accordance with the publisher's self-archiving policy. the publisher should be contacted for permission to re-use or reprint the material in any form. https://doi.org/10.1075/scl.85.04atw.

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

White Rose
university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

"Author Accepted Version" of:
Atwell, E. (2018). Classical and Modern Arabic Corpora: genre and language change.. In R. J. Whitt (Ed.), Diachronic corpora, genre, and language change.. John Benjamins.

Classical and Modern Arabic Corpora: genre and language change

Abstract

Our Artificial Intelligence (AI) research group in the School of Computing at the University of Leeds has collected, analysed and annotated a variety of Arabic corpus resources, and these have been widely used by other researchers. Classical Arabic texts, in particular the Quran and Hadith, are a specialised genre. The Classical Arabic Quran has been analysed, translated, interpreted and annotated by scholars for over a thousand years, resulting in many knowledge sources for rich corpus linguistic annotation. Modern Standard Arabic is the common written standard used throughout the Arab world; but our research with Arabic corpora has covered wider genre and language variation. AI researchers at Leeds University have collaborated with Arabic linguists to develop a number of Classical Arabic corpus resources: the Quranic Arabic Corpus with several layers of linguistic annotation; the QurAna Quran pronoun anaphoric co-reference corpus; the QurSim Quran verse similarity corpus; the Qurany Quran corpus annotated with English translations and verse topics; the Boundary-Annotated Quran Corpus; the Quran Question and Answer Corpus; the Multilingual Hadith Corpus; the King Saud University Corpus of Classical Arabic; and the Corpus for teaching about Islam. We have also developed Modern Arabic corpus resources spanning a range of genres and language types: Arabic By Computer; the Corpus of Contemporary Arabic; the Arabic Internet Corpus; the World Wide Arabic Corpus; the Arabic Discourse Treebank; the Arabic Learner Corpus; the Arabic Children's Corpus; and the Arabic Dialect Text Corpus. Modern Arabic corpus researchers harvest online news, web-pages, and internet social media; these might see to differ markedly from religious texts in language and genre. However, Quran verses are short text snippets, analogous to Twitter tweets or Amazon customer reviews. Quran verses annotated with analyses derived from traditional exegesis or scholars' commentaries can provide rich training data for supervised Machine Learning of language models, in Artificial Intelligence research. So, the language of the Quran may still inform Modern Arabic corpus linguistics and artificial intelligence research, and development of Modern Arabic text analytics tools.

Classical Arabic corpora for religious education and understanding

The University of Leeds is unique in having both an Artificial Intelligence (AI) research group in its School of Computing with research interests in Arabic text analytics and corpus linguistics, as well as a department of Arabic Islamic and Middle Eastern Studies (AIMES) with expert Arabic linguists who advise and collaborate with Artificial Intelligence researchers. Corpus linguistics researchers in the Artificial Intelligence research group of the School of Computing at the University of Leeds have collected, analysed and annotated a wide range of Arabic corpus resources, which illustrate genre and language variation in Arabic. This chapter reviews the range of Classical and Modern Arabic Corpus resources we have developed, and the range of applications they have been used for.

The Classical Arabic Quran is required reading for all faithful Muslims, who believe it comprises the teachings of God passed on by an angel to the prophet Mohammad, to be memorised and recited verbatim by all Muslims in unchanged, un-translated original wording. This differentiates the Quran from other religious texts such as the Bible or the Book of Mormon, which are generally read in translated form. Hence the Classical Arabic Quran is probably the single most widely read text ever. Another Classical Arabic text source of major significance to Muslims is the Hadith, the sayings and deeds of the prophet Mohammad, reported by his followers. The Quran and Hadith are the primary corpora of Classical Arabic, although other Classical Arabic texts corpora have also been collated.

To illustrate the range of language and linguistic annotation in Classical Arabic corpora, here are some of the Classical Arabic corpus resources developed by the Artificial Intelligence research group in the School of Computing at the University of Leeds:

Quranic Arabic Corpus

The Quran and other Classical Arabic texts have until recently not attracted much interest among corpus linguists, whose focus is on modern languages and modern language teaching. However, a growing area of Arabic corpus research by computer science and artificial intelligence researchers is in the development of computing tools and resources to aid access to and understanding of key Islamic texts, in particular the Classical Arabic texts of the Quran and Hadith. Muslims believe the Quran is the message from God passed on by the angel Gabriel to Mohammad, to teach others to recite. The Quran is divided into chapters and verses, with key themes or tropes running through and linking the text: the roles and attributes of Allah or God; the Day of Judgment when the world ends; stories of previous prophets or religious messengers; and rules and laws for faithful Muslims to obey. Mohammad has a special status in Islam, as the recipient and first reciter of the Quran; and his statements and actions are a secondary source of Islamic knowledge, the Hadith. Hadith are the collection of statements and actions of Mohammad reported by his followers.

One advantage of the Quran as a language data-set is that it has been analyzed, translated, interpreted, annotated and documented by scholars for over a thousand years. Such exegesis or critical explanation and interpretation of the text provides expert knowledge sources for rich corpus linguistic annotation. For example, the Tafsir or Quran exegesis of Ibn Kathir is widely respected and read by Muslims; it provides comments and analysis of each verse, including narrative description of morphological features of words, syntactic dependencies between words, meanings or semantics of key concepts, anaphoric references of pronouns to words and concepts in previous verses, cross-references to other verses with similar meanings, and other linguistic and conceptual insights. Artificial Intelligence researchers have codified this information from the narrative text into formal representations of morphology, grammar, dependency structure, anaphoric co-reference, meanings and ontologies, for use in Natural Language Processing tools for analysis of Arabic text. These formal computational models of Classical Quranic Arabic can then be adapted for analysis of Modern Arabic text.

The Quranic Arabic Corpus (Dukes and Atwell 2012, Dukes et al 2013), is an online, freely-accessible resource for learning, understanding and research in linguistics, artificial intelligence, and religious studies. There are several other website offering access to the text of the Quran, in original Arabic and in translations; the Quranic Arabic Corpus is unique in also offering several layers of annotation, including part-of-speech tagging and morphological segmentation (Dukes and Habash, 2010) syntactic analysis using dependency grammar (Dukes and Buckwalter, 2010, Dukes et al 2010), word-by-word English gloss, several parallel verse-by-verse English translations, audio recordings of recitations, and ontology or index of key Quranic entities and concepts (Dukes and Atwell 2012). There are several Modern Standard Arabic treebanks offering syntax tree annotations with each sentence, based on modern linguistic theories of grammar; the Quranic Arabic Corpus treebank is different in that it is built on a deep linguistic model based on the historical traditional grammar known as i'rāb. This traditional description of Quranic Arabic grammatical structure of sentences can engage Quranic Arabic readers as they are more likely to have been exposed to traditional i'rāb than to modern linguistic theories. Furthermore, use of i'rāb allowed us to engage our readers in linguistic annotation of the corpus. The corpus was first annotated with a rule-based tagger program; then online volunteers were invited to collaborate in proofreading the tagging. The Quranic Arabic Corpus morphological tagging resulted from “crowdsourcing”: about one hundred volunteer annotators proofread sections of the Quran text tags and corrected errors. A small group of expert supervisors reviewed proposed changes made by the crowdsource collaborators; each suggested correction to the computational analysis had to be justified with reference to exegesis. We built a linguistic software platform aimed at collaborative crowdsourcing: LAMP, the Linguistic Analysis Multimodal Platform (Dukes and Atwell 2012).

The Quranic Arabic Corpus has been used as a gold standard resource for a range of research on Arabic linguistics, and Arabic Natural Language Processing; for example: Arabic grammatical analysis (Mohammed and Omar 2011, Rabiee 2011), Arabic morphological analysis (Khaliq and Carroll 2013), (Alosaimy and Atwell 2017a,b); Arabic stylometrics (Alqurneh et al 2014), coherence analysis in Arabic translation studies (Tabrizi and Mahmud 2013), Arabic word stemming (Yusof et al 2010), Arabic text summarization (El-Haj et al 2015), Arabic oral-formulaic analysis (Bannister 2014). The Quranic Arabic Corpus has achieved greater social impact than typical Corpus Linguistics research projects: the website has attracted millions of visits, including non-Arabic-speakers who want to gain direct insights into the meanings and teachings of the original Classical Arabic text of the Quran through the linguistic annotations.

QurAna: Quran pronoun anaphoric co-reference corpus

QurAna (Sharaf and Atwell 2012a) (Muhammad 2012) is a specialized annotation data-set to accompany the Classical Arabic Quran corpus, showing pronoun co-reference. Each personal pronoun is tagged with its antecedent: the word or phrase it refers to, in the preceding (or occasionally following) text. In addition, ach personal pronoun is also tagged with its "meaning": a link to the the person, entity or concept that pronoun represents, in a separate Quran ontology or knowledge-base of people, entities and concepts. In Classical Arabic, most verbs are morphologically marked with personal pronoun(s) for subject and/or object(s); hence, there is a higher density of personal pronouns in the Classical Arabic text than in English or other translations. The Classical Arabic Quran has nearly 25,000 personal pronouns, and in QurAna each personal pronoun is tagged with antecedent information. QurAna also provides an ontology or term-index of over one thousand persons, entities and concepts, all linked to specific nouns or phrases in the Arabic text which are referred to by the personal pronouns. Deciding on the implicit reference of a personal pronoun in a text is not always straightforward; but for the Quran, we could follow the co-reference analysis in the Tafsir of Ibn Kathir. We could have at least as much confidence in this as in the alternative method widely used in linguistic corpus annotation projects, of relying on inter-annotator agreement between two casual-worker annotators. The QurAna pronoun anaphoric reference corpus is the first freely downloadable resource of its kind for any type or genre of Arabic. QurAna has been used to guide the analysis and annotation of other Arabic corpora (Zeroual and Lakhouaja

2016) (Hammo et al 2016) (Seddik et al 2015), and in the development of Quran ontologies or knowledge-bases of people, entities and concepts (Alrehaili and Atwell 2014) (Hakkoum and Raghay 2015a,b) (Alromima et al 2015) (Alqahtani and Atwell 2016) (Bentrcia et al 2017).

QurSim: Quran verse similarity corpus

QurSim (Sharaf and Atwell 2012b) (Muhammad 2012) is another type of corpus research resource based on the Classical Arabic Quran. QurSim shows pairs of Quran verses that are related or similar in meaning, according to the Tafsir or Quranic commentary work of Ibn Kathir. This exegesis examined and commented on each verse of the Quran, and noted links to other verses with related meanings and teachings. We text-mined the Tafsir to extract these cross-references, producing a corpus research resource of over 7,600 pairs of related verses. Users can choose a Quran verse and see verses related to this, which link on to a network of further related verses. Interestingly, we found that about one third of pairs of related verses shared one or more key words, indicating that "relatedness" of religious text can be partly predicted by lexical matching. However, two thirds of related verse-pairs had no words in common, so predicting semantic relatedness for cases like these is computationally more challenging, requiring artificial intelligence modelling of context and domain knowledge. The QurSim Classical Arabic corpus resource can be used for research on meaning relatedness, similarity and paraphrasing in short texts. Ibn Kathir's commentary was an analysis of the Classical Arabic source text of the Quran, but the verse-relations can also apply to translations: two Quran verses which Ibn Kathir noted as linked in meaning should still be "related" even after translation to another language. Hence, QurSim is a corpus resource for research on textual similarity and relatedness in another language that has a Quran translation, eg (Basharat et al 2015). QurSim can also be used for extraction and visualization of topics in the Quran (Panju 2014), and as a component in building further Quran resources such as Quranic Arabic Wordnet linking similar lexical items (AlMaayah et al 2104) and Quran ontologies or concept-indexes (Alrehaili and Atwell 2014) (Hakkoum and Raghay 2015a,b) (Alqahtani and Atwell 2016).

Qurany: Classical Arabic Quran with English translations and verse topics.

Qurany (Abbas 2009, Abbas and Atwell 2013) is another corpus research resource based on the Classical Arabic Quran. Qurany was developed to help Quran readers to search for and find verses related to a given concept or concepts. To do this, Qurany encodes the source Arabic text of each verse along with several representations of the meanings or concepts in that verse. Each verse in the Quran is annotated with semantic conceptual category tags, extracted from a respected Quran commentary which includes an index of nearly 1100 concepts or topics with links to the Quran verses, the Mushaf Al Tajweed index. This index shows the main concepts or topics in the Quran, along with the verses each concept appears in. The index was encoded in a Python ontology or knowledge representation formalism. Qurany can be accessed via a web-browser, so that users can navigate the ontology as a tree of main concept-tags, sub-concept-tags, sub-sub-concept-tags etc. Having chosen a specific fine-grained concept-tag, the user can then follow the link to a list of verses tagged with this concept-tag. The concept-tags are available in original Arabic and also in English translation. Each Arabic verse is also annotated with 8 alternative English translations from popular published sources. A verse can be found via Arabic or English keyword-search if any of the original Arabic or English translations contain the keyword(s). Also, the user can opt to see synonyms of keywords, derived from the WordNet synonym-set knowledge-base, to broaden the search-terms. This leads to improved recall: Qurany can show the user more of the verses which are relevant or semantically related to their query. The Qurany dataset is also searchable via standard Google search (or Yahoo, or Bing, or other web-search systems), as it is also online in a single website consisting of a large set of of separate web-pages, one per Quran verse. Each verse-webpage displays Arabic source text, 8 English translation texts, and lists of concept-labels or semantic tags relating to the verse, written in both Arabic and English. This website is compatible with standard web search engines. For example, a Google search for “wine” with the site: parameter set to this version of the Qurany website will match all Quran verses containing the word “wine” in at least one of the English translations or concept-labels; and this in turn allows you to see a range of alternative English translations for these verses, along with the Arabic source text, so you can see different translations or interpretation for the word or concept of “wine”. In this way, Qurany is useful for showing the range of possible translations and sometimes metaphors and euphemisms for a concept such as “wine” (Gehrels 2016) or “fornication” (Wood 2016). The Qurany resource has been used in research in digital religious studies (Clivaz 2013), Quranic Arabic word meanings or lexical semantics (Al-khalifa et al 2010), terminology extraction (Mukhtar et al 2012), formalized knowledge extraction from the Quran (Saad et al 2011, 2013) (Muhammad 2012) (Abed 2015) (Ouda 2015) (Almaayah et al 2016) (Alrehaili and Atwell 2016) (Bentrcia et al 2017), Quran recitation methods (Mahmoud and Hassan 2013), combining and merging formal Quran knowledge representations and ontologies (Atwell et al 2011) (Alqassem 2013) (Dukes et al 2013) (Ahmad 2017), and knowledge-based systems for question answering about the Quran (Baqai et al 2009) (Chelli 2012) (Abdelhamid et al 2013) (Jilani 2013) (Shmeisania et al 2014) (Mohamed at al 2015) (Hakkoum and Raghay 2015a,b) (Bakari et al 2015) (Alqahtani and Atwell

2016) (Hassan and Atwell 2016a) (Kadir and Yauri 2017) (Alqahtani and Atwell 2017).

Boundary-Annotated Quran Corpus

Our Boundary-Annotated Quran Corpus is another type of corpus research resource based on the Classical Arabic Quran. The number of words and sentences in Arabic text depends on precisely how word-boundaries and sentence-boundaries are defined and counted. For the Boundary-Annotated Quran Corpus, we developed a precise computational definition and implementation of sentence and word boundaries, to arrive at a dataset of 77430 words and 8230 sentences of the Classical Arabic text of the Quran. In addition, each word is tagged with phonetic, prosodic and syntactic annotations (Brierley et al 2006 ,2012a,b). The Boundary-Annotated Quran Corpus has been used for research in Arabic prosody modeling and visualization (Brierley et al 2012c, 2014) (Sawalha et al 2012a), Arabic phrase break prediction (Sawalha et al 2013 b,c,d), Arabic speech-to-text transcription (Brierley et al 2016, Sawalha et al 2014a,b, 2017).

Quran Question and Answer Corpus

Our Quran Question and Answer Corpus is another type of corpus research resource based on the Classical Arabic Quran, but this time extending the Quran text to include questions about the Quran, with answers that include one or more verses from the Quran. In effect, each verse is “annotated” with one or more questions about the verse, and explanatory text linking the question(s) to the verse. Question-Answering systems and chatbots have been developed for a variety of domains (Abu Shawar and Atwell 2007, 2015, 2016). Answering questions about Quran teachings is somewhat different from QA in most other domains, in that the answer is usually expected to be a verse from the Quran, or at least to be based on interpretation of a Quran verse. So, we could model answering a question about the Quran as finding the “best-match” verse for the given input (Abu Shawar and Atwell 2004, 2009). A more general approach is to collect a corpus of attested, reputable answers to questions about the Quran as a knowledge-base of question-answer pairs, and then the response to a given new question involves finding the “best match” question in the corpus and presenting the corresponding answer to the user. This has led us to collect Quran question-answer corpus collections from Quran scholars (Hamdelsayed and Atwell 2016a,b, 2017) and Islamic websites with Quran Frequently Asked Questions (FAQs) with answers devised by Islamic experts (Hamoud and Atwell 2016a,b,c).

Multilingual Hadith Corpus

The Quran is believed by Islamic scholars to be the text transcript of messages sent from Allah, and the primary exemplar of Classical Arabic; so the Quran is the focus of both religious and linguistic research. The Hadith are not direct “words of God” but statements about the deeds and saying of Mohammed, and the second most widely used source of Classical Arabic text. There are many sources of Hadith with varying credentials, depending on credibility of the claimed chain of narrators who passed on the Hadith verbally, at least initially. Because Hadith are not claimed to be the literal words of God, unlike the Quran, it is not so imperative that they are read and understood in original Classical Arabic. We collated a Multilingual Hadith Corpus, including parallel texts in Classical Arabic, English, French and Russian (Altoum and Atwell 2016) (Hassan and Atwell 2016a,b,c). In information retrieval experiments, we found that search with Arabic keywords in the Arabic original sub-corpus gave slightly higher accuracy results that equivalent searches in the English, French and Russian equivalents.

KSUCCA King Saud University Corpus of Classical Arabic

The Quran is an exemplar of Classical Arabic, and can be used to extract and study examples of Classical Arabic lexis and grammar, using an Arabic-friendly concordance program such as aConCorde (Roberts et al 2005) or SketchEngine (Kilgarriff et al 2014). However, linguists and lexicographers generally see the need for much larger corpora to study less frequent linguistic phenomena such as lexical co-occurrence patterns, collocations and concordance patterns. For research on multi-word units, collocation patterns, rare words, etc., we need a sufficiently large set of examples of each pattern to be studied; but many words and phrases in the Quran occur only once or a handful of times. The Quran contains about seventy thousand words, depending on how word-boundaries are defined and counted; whereas the British National Corpus, the first very large corpus developed for British English dictionary research, is much larger, about 100 million words..

So, we collaborated with researchers at King Saud University to collect a larger sample of Classical Arabic for lexical pattern research. The 50-million word corpus contains the Quran and some Haddith, and also a range of Classical Arabic texts from around the same period as the Quran was first recited and shortly after. These are predominantly religious texts related to Islam, such as commentaries on the Quran, and biographies of early Islamic scholars. The King Saud University Corpus of Classical Arabic (Alrabiah et al 2013, 2014a,b) allows us to select a word from the Quran and then find many more examples of its use in context in the broader sample of Classical Arabic. The KSUCCA corpus is downloadable from the KSUCCA website, and also searchable online via SketchEngine (Kilgarriff et al 2014). KSUCCA has been used for corpus-based study of Arabic historical linguistics (Alrabiah et al 2014a), and study of distributional lexical patterns around words from the Quran (Alrabiah et al 2014b).

Corpus for teaching about Islam

We also used the Web to collect a specialized ‘corpus’ of texts for university-level teaching about Islam (Atwell et al 2011), for a project on a Web-as-Corpus approach to populating Wikiversity for teaching about Islam and Muslims in language, linguistics and area studies. The language of this corpus is in fact mainly English, but it contains extracts and references from the core Classical Arabic texts.

Modern Arabic corpora for language teaching, lexicography, and text analytics

Modern Arabic text corpora have been developed for a wide range of applications. The earliest Arabic corpora were developed for modern Arabic language teaching and translation studies, to provide representative example texts for teaching and translating modern Arabic. For example, they included magazine articles and similar sources suitable for classroom teaching examples and exercises. A related use of modern Arabic corpora is for dictionary development, for learners and translators; lexicographers exploited modern Arabic corpora to extract concordance examples of lexical items in context, to inform the writing of dictionary word and sense definitions and translations. Interesting dictionary items can occur infrequently in a corpus, so lexicographers required much larger corpora.

Artificial Intelligence research, applying Machine Learning to corpus data to build Natural Language Processing models and tools, is a very different use of corpora, but also requires large Arabic corpora. For lexicography and Artificial Intelligence Machine Learning research, size matters more than genre balance, so researchers tended to harvest the most readily available large-scale sources of Arabic text: online news, web-pages, and more recently, internet social media such as Twitter and FaceBook.

To illustrate the range of genres and language in Arabic corpora, here are some of the modern Arabic corpus resources we have developed by Artificial Intelligence researchers in the School of Computing at the University of Leeds:

ABC: Arabic By Computer

The first Arabic corpus resource project at Leeds University was ABC Arabic By Computer; we built an Arabic text database and glossary system for Arabic language students (Brockett et al 1989). In the 1980s, UK universities could not afford to provide computers for language teaching, as computers were expensive resources which attracted funding only for science and engineering research. We saw a future demand for free open access to Arabic corpus linguistics resources for teaching and research. A major practical hurdle was that computer interfaces at the time only allowed input and output of a very restricted set of characters: Roman alphabet letters A to Z, digits 0 to 9 and a few mathematical symbols. Capture, editing and display of Arabic text required specialist Apple Macintosh hardware and software, including rudimentary Arabic word processing software. Our focus on these technical challenges left less time for linguistic and pedagogical issues such as planning and analysis of text types and genres to be included to match Arabic language syllabuses. The ABC corpus contained a small selection of Arabic magazine articles typed into the Macintosh to provide computer-readable and searchable example texts for Arabic teaching and learning. ABC was a very small example corpus, a taster of things to come.

CCA: Corpus of Contemporary Arabic

Corpus Linguistics research initially concentrated on corpora and tools for English linguistics and language teaching; we wanted to extend corpus linguistics methods and resources to Arabic linguistics and Arabic language teaching. This required an Arabic corpus, so we developed the first freely downloadable million-word Corpus of Contemporary Arabic (Al-Sulaiti and Atwell 2005, 2006). The million-word LOB and Brown corpora of Modern British English and American English published texts (Leech et al 1983a) were widely used in English corpus linguistics and English teaching; so we wanted to develop a comparable collection of Arabic published texts. However, the range and balance of text genres in LOB and Brown were decided by Brown University linguistics researchers in the 1970s, and this might not suit the needs of contemporary Arabic researcher and teachers. So, before collecting texts we examined the range of genres covered in other corpora, and undertook a survey of prospective Arabic corpus users in Arabic natural language processing and Arabic language teaching, to identify user needs and preferences for text-types and genres to be included in our Corpus of Contemporary Arabic. This informed our selection of contemporary online sources of text samples: a range of online magazines, websites, newspapers, radio broadcast transcripts, and emails.

The CCA has been widely used by researchers in Arabic language education, Arabic translation, Arabic natural language processing and Arabic corpus linguists; for example in Arabic lexical profiling (Attia et al

2011), the translation of culturally bound metaphors (Merakchi and Rogers 2013) (Affeich 2011), learning Arabic spelling and vocabulary (Erradi et al 2012), lexical differences in world affairs and sports sections in Arabic newspapers (Abdul Razak 2011), corpus-based sociolinguistics (Friginal and Hardy 2014), sentiment analysis of Arabic text (Itani et al 2017), automated Arabic document classification and clustering (Saad and Ashour 2010) (Froud et al 2010) (Aly and Kelleny 2014), automated Arabic text summarisation (Froud et al 2013) (Al-Saleh and Menai 2016), improving security and capacity for Arabic text steganography (Al-Haidari et al 2009), developing and evaluating concordancers for Arabic text (Atwell et al 2004) (Roberts et al 2006) (Alfaifi and Atwell 2016), Arabic Part-of-Speech tagging (El-Haj et al 2009) (Sawalha and Atwell 2010a), comparative evaluation of Arabic language morphological analysers and stemmers (Sawalha and Atwell 2008, 2013c), Arabic word sense disambiguation (Zouaghi et al 2011), development of Arabic speech recognition systems (Abushariah et al 2012). The methodology for design and collection of the Corpus of Contemporary Arabic was used as a model for corpus development research for other languages and dialects, including Persian (Bijankhan et al 2011), Kazakh (Makhambetov et al 2013), Palestinian (Jarrar et al 2017), Malay (Romli et al 2016), and Igbo (Onyenwe 2017).

Arabic Internet Corpus

To collect the Corpus of Contemporary Arabic text samples, we visited websites and selected sample texts to download “by hand”, a time-consuming process; but at least we did not have to type in the text, as was the case for earlier Brown and LOB corpus projects. Collecting a million-word corpus “manually” is labour-intensive and expensive.

The British National Corpus (BNC), one hundred million words of British English, was a large-scale research effort by a consortium of several British universities and companies, and became an established gold standard for English corpus linguistics research in the 1990s. In practice, Corpus Linguistics and Artificial Intelligence researchers working on other languages could not get funding to manually collate a large general corpus of the size of BNC, 100 million words. The Web-as-Corpus approach (Baroni and Bernardini 2004) offered a more practicable alternative. The Web-as-Corpus approach automates the process of corpus collection from websites: a web-bot visits web-pages and “scrapes” the text. Researchers at Leeds University have used this Web-as-Corpus method to collect Internet corpora for English, and many other languages including French, German, Italian, Spanish, Polish, Russian, and Arabic (Sharoff 2006). We can harness these Internet corpora for linguistic research via a web-page with a concordance and collocation search interface. This collection of web-corpora includes the 176-Million-word Arabic Internet Corpus, which we subsequently lemmatized using the SALMA morphological analysis toolkit

(Sawalha and Atwell 2013a). Such a large corpus can be used to examine examples of uses of rare lexical items and collocations, which occur too infrequently in the smaller Corpus of Contemporary Arabic to give sufficient examples. The web-bot collects all texts that match the search terms: a list of about 100 “seed-terms” or common Arabic words, used by the search-engine to find documents containing the seed-terms. Because there is no direct human control over the types of text to be included, a Web-as-Corpus can contain the wide variety of genres found on the World Wide Web. Identification and classification of genre of web-pages is a challenge for Artificial Intelligence research.

World Wide Arabic Corpus

The Web-as-Corpus web-bot can automatically download web-pages that match the “seed-term” list of search terms, and keep collecting matching web-pages until we have reached a target corpus size. We can specify additional constraints on the search, such as limiting matches to a given web address or domain; for example, limiting matches to URLs ending in .SD limits the corpus to web-pages from Sudan. This allows us to run the web-bot repeatedly, changing the national URL constraint each time, to collect a balanced corpus with equal-sized samples from different Arab countries, representing different national dialects of Arabic. We collected a World Wide Arabic Corpus, analogous to the World Wide English Corpus (Atwell et al 2007), comprising 200,000-word national sub-corpora, to capture country-by-country linguistic and dialect variation in written Arabic. This has been used to study dialect variation in written Arabic text, for example Arabic dialect variation in connectives (Hassan et al 2010, 2013) and variation in Arabic and Arab English in the Arab world (Atwell et al 2009).

Arabic Discourse Treebank

As outlined above, we added a range of linguistic and semantic annotations to the Quran, to make it a richer corpus resource for research in corpus linguistics and artificial intelligence. Modern Arabic corpus texts can also be annotated with linguistic tagging for research. The Leeds Arabic Discourse Treebank (Alsaif and Markert 2010) is a corpus of 537 news texts where 5651 discourse connectives are tagged with discourse relation type, using a custom discourse annotation program. The argument phrases or clauses they connect are also tagged.

Arabic Learner Corpus

The Arabic Learner Corpus (ALC) (Alfaifi and Atwell 2013, Alfaifi et al 2014) is a collection of texts by learners of Arabic in Saudi Arabia. The Arabic Learner Corpus includes 282,732 words, in 1585 short student essays or reports, with an average text sample length of 178 words. The texts are student essays or reports on one of two topics: “A vacation trip” (narrative) and “My study interest” (discussion). 942 students from 67 nationalities and 66 different first languages produced the texts. The Arabic Learner Corpus also includes rich metadata information, in both English and Arabic, which enables researchers to identify key characteristics of a text and its producer. Each text is stored in a separate file, and key characteristics of the text are also encoded in text sample filenames; e.g. S038_T2_M_Pre_NNAS_W_C shows student identifier S038, text number T2, author gender Male, level of study Pre, nativeness NNAS, text mode Written, and place of text production C. The original hand-written sheets are also downloadable as scanned PDF files. A small portion of the learner texts were spoken by the learner and then transcribed; the original 3.5 hours of MP3 audio recordings are also available to download. The ALC is downloadable and/or searchable from several websites, including SketchEngine (Kilgarriff et al 2014). The corpus has been used for research in Arabic native language identification (Malmasi and Dras 2014, 2015), critical discourse analysis (Haider 2016), and automatic text correction (Mohit et al 2014) (Zaghouani et al 2015).

Arabic Children’s Corpus

The Arabic Childrens’ Corpus is a collection of texts written for children (Al-Sulaiti et al 2016); note this is NOT texts written BY children. The Arabic Children's Corpus contains 2950 documents and nearly 2 million words, collected by manually searching the web for suitable texts over a 3-month period. It enabled us to measure variation in Arabic language, and in particular vocabulary, in writing aimed at children compared to an Adult readership, represented by most other Arabic corpora.

Arabic Dialect Text Corpus

Most Modern Arabic texts are in Modern Standard Arabic, an international standard written form taught in schools and used in formal writing across the Arab world. Spoken Arabic can vary significantly in language and vocabulary from Modern Standard Arabic; and there is no one standard spoken form, but a wide variety of dialects. Arabic dialect studies have tended to focus on differences in phonetics and phonology, in the variant pronunciations. Speakers of Arabic dialects (including all Arabic native speakers) generally write in Modern Standard Arabic, in effect “translating” from dialect to MSA. There are few Arabic dialect written texts, or standard writing conventions to capture dialect. However, for text analytics research and applications, it is important to be able to document and handle variation in Arabic dialect vocabulary, morphology and grammar. One way to capture written dialect texts is to record dialect speakers and use speech recognition software for an automatic speech-to-text transcription; this was done for the VarDial’2016 contest, to train and test Machine Learning models for Arabic dialect classification (Alshutayri et al 2016). Another approach is to harvest online sources of spontaneous informal Arabic: online forums and social media where users are likely to write as they speak and not strictly follow Modern Standard Arabic conventions. We harvested Twitter tweets, using location to identify dialect, in an Arabic Twitter Dialect Text Corpus (Alshutayri 2017a); and then extended this by also harvesting online newspaper reader comments (Alshutayri et al 2017b).

Machine Learning from the Quran for Modern Arabic text analytics

On the face of it, there is a mismatch in genre and language between the Quran, a collection of religious chapters and verses, and Modern Arabic corpus sources such as online news, web-pages, and Twitter. So how can we equitably compare Classical Arabic to Modern Arabic? And how can we make use of Classical Arabic corpora to aid Natural Language Processing research targeted at Modern Arabic users?

Much current applied NLP research on Modern Arabic is focussed on short text snippets, such as analysis of Twitter tweets, Facebook comments, or Amazon customer reviews. To train supervised Machine Learning models of language processing, we need language data annotated with appropriate target linguistic analyses. Despite the some vocabulary differences, the Quran verses are also short text snippets; but with the added bonus of linguistic annotations added to the text snippets derived from centuries of expert study. Linguistically annotated Quran verses provide a rich training set for supervised Machine Learning of language models. For example, the Quranic Arabic Corpus morphological annotations were used to train Machine Learning models for Modern Arabic morphological analysis (Khaliq and Carroll 2013); the QurAna anaphoric coreference annotations were used to train Machine Learning models for Modern Arabic anaphora resolution (Seddik et al 2015); the Boundary-Annotated Quran Corpus phonetic and prosodic annotations have been used to train Machine Learning models for Modern Arabic speech-to-text transcription (Brierley et al 2016, Sawalha et al 2017). Despite the differences in genre and language variety, the Classical Arabic text of the Quran can inform Artificial Intelligence and Corpus Linguistics research on Modern Arabic.

References

Abbas, N and E Atwell. 2013. ‘Annotating the Arabic Quran with a classical semantic ontology.’ Proceedings of WACL’2 Second Workshop on Arabic Corpus Linguistics.

Abbas, N, L Aldhubayi, H Al-Khalifa H, Z Alqassem, E Atwell, K Dukes, M Sawalha, and M Sharaf. 2013. ‘Unifying linguistic annotations and ontologies for the Arabic Quran.’ Proceedings of WACL2 Second Workshop on Arabic Corpus Linguistics.

Abbas, N. 2009. ‘Quran Search for a Concept Tool and Website’. MRes Thesis, School of Computing, University of Leeds.

Abdelhamid, Y, M Mahmoud and T El-Sakka. 2013. Using ontology for associating Web multimedia resources with the Holy Quran. In Proceedings of Advances in Information Technology for the Holy Quran and Its Sciences, pp. 246-251, Taibah University, Medina, Saudi Arabia.

Abdul Razak, Z. 2011. ‘Modern media Arabic: a study of word frequency in world affairs and sports sections in Arabic newspapers.’ PhD Thesis, University of Birmingham.

Abed, Q. 2015. Ontology-based approach for retrieving knowledge in Al-Quran. PhD thesis, Universiti Utara Malaysia.

Abushariah, M, R Ainon, R Zainuddin, M Elshafei and O Khalifa. 2012. Arabic speaker-independent continuous automatic speech recognition based on a phonetically rich and balanced speech corpus. International Arab Journal of Information Technology (IAJIT), 9(1), pp.84-93.

Abu Shawar, B and E Atwell. 2004. ‘An Arabic chatbot giving answers from the Quran.’ Proc TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles.

Abu Shawar, B and E Atwell. 2005a. ‘Using corpora in machine-learning chatbot systems.’ International Journal of Corpus Linguistics, vol. 10, pp. 489-516.

Abu Shawar, B and E Atwell. 2005b. ‘A chatbot system as a tool to animate a corpus.’ ICAME Journal: International Computer Archive of Modern and Medieval English Journal, vol. 29, pp.5-24.

Abu Shawar, B and E Atwell. 2009. ‘Arabic Question-Answering via Instance Based Learning from an FAQ Corpus.’ Proceedings of CL2009 Corpus Linguistics.

Abu Shawar, B and E Atwell. 2007. Chatbots: Sind Sie wirklich nu"tzlich? (Chatbots: are they really useful?). LDV-Forum Journal for Computational Linguistics and Language Technology, 22 , pp. 31-50.

Abu Shawar, B and E Atwell. 2015. ALICE chatbot: Trials and outputs. Computacion y Sistemas, 19 (4), pp. 625-632.

Abu Shawar, B and E Atwell. 2016. Usefulness, localizability, humanness, and language-benefit: additional evaluation criteria for natural language dialogue systems. International Journal of Speech Technology, 19 (2), pp. 373-383.

Affeich, A. 2011. La métaphore dans le discours technique d’Internet et son passage de l’anglais vers l’arabe. In Proceedings of JéTou’2011 Journées d'études Toulousaines, Toulouse, France.

Ahmad, N, B Bennett and E Atwell. 2017. Retrieval Performance for Malay Quran. International Journal on Islamic Applications in Computer Science and Technology (IJASAT), 5(2), pp.13-25.

Al-Haidari, F, A Gutub, K Al-Kahsah and J Hamodi. 2009. Improving security and capacity for Arabic text steganography using Kashida extensions. In Proceedings of CSA'2009 Computer Systems and Applications.

Al-Khalifa, H, M Al-Yahya, A Bahanshal, I Al-Odah and N Al-Helwah. 2010. An approach to compare two ontological models for representing Quranic words. In Proceedings of the 12th International Conference on Information Integration and Web-based Applications and Services, pp. 674-678. Paris, France.

Almaayah, M, M Sawalha and M Abushariah. 2016. Towards an automatic extraction of synonyms for Quranic Arabic WordNet. International Journal of Speech Technology, 19(2), pp.177-189.

Alromima, W, R Elgohary, I Moawad and M Aref. 2015. Applying ontological engineering approach for Arabic Quran corpus: a comprehensive survey. In Proceedings of ICICIS International Conference on Intelligent Computing and Information Systems, pp. 620-627, IEEE.

Al-Saif, A, and K Markert. 2010. 'The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic.' Proceedings of LREC'2010: Language Resources and Evaluation Conference.

Al-Sulaiti, L and E Atwell. 2006. 'The design of a corpus of contemporary Arabic.' International Journal of Corpus Linguistics, vol. 11, pp. 135-171.

Al-Sulaiti, L, A Roberts, and E Atwell. 2005. 'The use of corpora and concordance in the teaching of contemporary Arabic.' Proceedings of EuroCALL'2005.

Al-Sulaiti, L, A Roberts, B Abu Shawar, and E Atwell. 2007. 'The Use of Corpus, Concordancer and Chatbot in the Teaching of Contemporary Arabic.' Proceedings of CL'2007 Corpus Linguistics

Al-Sulaiti, L, and E Atwell. 2005. 'Extending the Corpus of Contemporary Arabic.' Proceedings of CL'2005 Corpus Linguistics.

Alfaifi, A and E Atwell. 2012. 'Arabic Learner Corpora (ALC): A Taxonomy of Coding Errors'. Proceedings of ICCA'2012 International Computing Conference in Arabic.

Alfaifi, A and E Atwell. 2013a. 'Arabic Learner Corpus v1: A New Resource for Arabic Language Research.' Proceedings of WACL'2 Second Workshop on Arabic Corpus Linguistics.

Alfaifi, A and E Atwell. 2013b. 'Arabic Learner Corpus: Texts Transcription and Files Format.' Proceedings of CORPORA'2013 International Conference on Corpus Linguistics.

Alfaifi, A and E Atwell. 2014a. 'Tools for Searching and Analysing Arabic Corpora: an Evaluation Study.' Proceedings BAAL-CUP'2014 British

Association for Applied Linguistics and Cambridge University Press Applied Linguistics Workshop.

Alfaifi, A and E Atwell. 2014b. ‘An evaluation of the Arabic error tagset v2.’ Proceedings of AACL’2014 American Association for Corpus Linguistics.

Alfaifi, A and Atwell, Eric. 2015. Computer-Aided Error Annotation A New Tool for Annotating Arabic Error. The 8th Saudi Students Conference, 31 January – 1 February 2015, London, UK.

Alfaifi, A and E Atwell. 2016. Comparative evaluation of tools for Arabic corpora search and analysis. International Journal of Speech Technology, 19(2), pp.347-357.

Alfaifi, A, E Atwell, and G Abuhakema. 2013. ‘Error Annotation of the Arabic Learner Corpus: A New Error Tagset. Language Processing and Knowledge in the Web, vol. 8105, pp.14-22. Springer.

Alfaifi, A, E Atwell, and I Hedaya. 2014. ‘Arabic Learner Corpus (ALC) v2: A New Written and Spoken Corpus of Arabic Learners.’ Proceedings of LCSAW’2014 Learner Corpus Studies in Asia and the World.

Ali,I. 2012. ‘Application of a Mining Algorithm to Finding Frequent Patterns in a Text Corpus: A Case Study of Arabic.’ International Journal of Software Engineering and Its Applications, vol.6(3), pp.127-134.

AlMaayah, M, M Sawalha and M Abushariah. 2014. A proposed model for Quranic Arabic WordNet. In Proceedings of LRE-REL2 2nd Workshop on Language Resources and Evaluation for Religious Texts, Reykjavik, Iceland.

Alosaimy, A., and Atwell, E. 2017a. Joint Alignment of Segmentation and Labelling for Arabic Morphosyntactic Taggers. International Journal of Computational Linguistics.

Alosaimy, A. and Atwell, E. 2017b. Tagging Classical Arabic Text using Available Morphological Analysers and Part of Speech Taggers. Journal for Language Technology and Computational Linguistics.

Alqahtani, M and E Atwell. 2016. Arabic Quranic Search Tool Based on Ontology. In Proceedings of NLDB’2016 International Conference on Applications of Natural Language to Information Systems, pp. 478-485.

Alqahtani, M and E Atwell. 2017. Evaluation Criteria for Computational Quran Search. International Journal on Islamic Applications in Computer Science And Technology, 5(1), pp.12-22.

Alqassem, Z. 2013. Unifying Quranic Analyses into a Single Database. BSc Research Project Report, School of Computing, University of Leeds.

Alqurneh, A, A Mustapha, M Murad, and N Sharef. 2014. ‘Stylometric model for detecting oath expressions: A case study for Quranic texts.’ Literary and Linguistic Computing journal.

Alrabiah, M, A Al-Salman, and E Atwell. 2013. ‘The design and construction of the 50 million words KSUCCA King Saud University Corpus of Classical Arabic.’ Proceedings of WACL’2 Second Workshop on Arabic Corpus Linguistics.

Alrabiah, M, A Al-Salman, E Atwell, and N Alhelewh. 2014. ‘KSUCCA: A Key To Exploring Arabic Historical Linguistics.’ International Journal of Computational Linguistics, vol. 5, pp.27-36.

Alrabiah, M, N Alhelewh, A Al-Salman, and E Atwell. 2014. ‘An Empirical Study On The Holy Quran Based On A Large Classical Arabic Corpus.’ International Journal of Computational Linguistics, vol. 5, pp.1-13.

Alrehaili, S and E Atwell. 2013. ‘Linguistics features to confirm the chronological order of the Quran.’ Proceedings of WACL’2 Second Workshop on Arabic Corpus Linguistics.

Alrehaili, S and E Atwell. 2014. ‘Computational ontologies for semantic tagging of the Quran.’ Proceedings of LRE-Rel 2: 2nd Workshop on Language Resource and Evaluation for Religious Texts, Reykjavik, Iceland.

Alrehaili, S and E Atwell. 2016. A Hybrid-based Term Extraction method on the Arabic text of the Quran. In Proceedings of IMAN'2016 Islamic Applications in Computer Science and Technologies, Khartoum, Sudan.

Alruily, M. 2012. ‘Using Text Mining to Identify Crime Patterns from Arabic Crime News Report Corpus.’ PhD Thesis, De Montford University.

Al-Saleh, A and M Menai. 2016. Automatic Arabic text summarization: a survey. Artificial Intelligence Review, 45(2), pp.203-234.

Altoum, S and E Atwell. 2016. Compilation of an Islamic Hadith Corpus: الشريف النبوي الحديث مدونة تجمع. In Proceedings of ICCA'2016 International Conference on Computing in Arabic, Khartoum, Sudan.

Aly, W and H Kelleny. 2014. Adaptation of Cuckoo Search for Documents Clustering. International Journal of Computer Applications, 86(1).

Attia, M, P Pecina, L Tounsi, A Toral, and J Van Genabith. 2011. ‘Lexical Profiling for Arabic.’ Proceedings of eLex’2011 Electronic Lexicography in the 21st Century.

Atwell, E, L Al-Sulaiti, S Al-Osaimi, S and B Abu Shawar. 2004. A Review of Arabic Corpus Analysis Tools. In Proceedings of JEP-TALN‘2004 workshop on Arabic Language Processing.

Atwell, E, C Brierley, K Dukes, M Sawalha, and A Sharaf. 2011. ‘An Artificial Intelligence Approach to Arabic and Islamic Content on the Internet.’ Proceedings of NITS’2011 3rd National Information Technology Symposium, King Saud University, Riyadh, Saudi Arabia.

Atwell, E, J Arshad, C Lai, L Nim, N Rezapour Asheghi, J Wang, and J Washtell. 2007. ‘Which English dominates the World Wide Web, British or American?’ Proceedings of CL'2007 Corpus Linguistics.

Atwell, E, K Dukes, A Sharaf, N Habash, B Louw, B Abu Shawar, A McEnery, W Zaghouani, and M El-Haj. 2010. ‘Understanding the Quran: a new Grand Challenge for Computer Science and Artificial Intelligence.’ Proceedings of GCCR'2010 Grand Challenges in Computing Research, Edinburgh, Scotland UK.

Atwell, E, L Al-Sulaiti, and S Sharoff. 2009. ‘Arabic and Arab English in the Arab World.’ Proceedings of CL2009 Corpus Linguistics.

Atwell, E, L Al-Sulaiti, S Al-Osaimi, and B Abu Shawar. 2004. ‘A review of Arabic corpus analysis tools’, Proceedings of TALN’2004: Traitement Automatique des Langues Naturelles.

Atwell, E, N Abbas, B Abu Shawar, L Al-Sulaiti, A Roberts, and M Sawalha. 2008. ‘Mapping Middle Eastern and North African Diasporas.’ Proceedings of BRISMES’2008 British Society for Middle Eastern Studies, Leeds, UK.

Atwell, E. (ed.) 1993. ‘Knowledge at Work in Universities - Proceedings of the second annual conference of the Higher Education Funding Council's Knowledge Based Systems Initiative.’ 146pp. Leeds University Press.

Atwell, E. 1982. LOB Corpus Tagging Project: Manual Postedit Handbook. Department of Linguistics and Modern English Language, University of Lancaster.

Atwell, E. 1993. ‘The HEFC's Knowledge Based Systems Initiative.’ AISBQ: Artificial Intelligence and Simulation of Behaviour Quarterly, vol. 83, pp.29-34.

Atwell, E. 2008. ‘Development of tag sets for part-of-speech tagging.’ Ludeling A; Kyto M (ed.) Corpus Linguistics: An International Handbook, Volume 1, pp.501-526. Mouton de Gruyter.

Atwell, E. 2011. ‘Exploiting New Technology and Innovation For Detecting Terrorist Activities.’ Counter Terror Expo, London.

Bakari, W, P Bellot and M Neji. 2015. Literature Review of Arabic Question-Answering: Modeling, Generation, Experimentation and Performance Analysis. In Proceedings of Flexible Query Answering Systems, pp. 321-334.

Bannister, A. 2014. ‘An Oral-Formulaic Study of the Quran.’ Lexington.

Baqai, S, A Basharat, H Khalid, A Hassan and S Zafar. 2009. Leveraging semantic web technologies for standardized knowledge modeling and retrieval from the Holy Qur'an and religious texts. In Proceedings of the 7th International Conference on Frontiers of Information Technology, Abbottabad, Pakistan.

Baroni, M and S Bernardini. 2004. ‘BootCaT: Bootstrapping corpora and terms from the web.’ Proceedings of LREC’2004 Language Resources and Evaluation Conference.

Basharat, A, D Yasdansepas and K Rasheed, 2015. Comparative study of verse similarity for multi-lingual representations of the Quran. In Proceedings of ICAI’2015 International Conference on Artificial Intelligence.

Bentrcia, R, S Zidat and F Marir. 2017. Extracting Semantic Relations from the Quranic Arabic Based on Arabic Conjunctive Patterns. Journal of King Saud University Computer and Information Sciences.

Bijankhan, M, J Sheykhzadegan, M Bahrani and M Ghayoomi. 2011. Lessons from building a Persian written corpus: Peykare. Language Resources and Evaluation Journal, 45(2), pp.143-164.

Brierley, C, E Atwell, C Rowland, and J Anderson. 2013. ‘Semantic Pathways: a Novel Visualization of Varieties of English.’ ICAME Journal of the International Computer Archive of Modern English, vol. 37, pp.5-36.

Brierley, C, M Sawalha, and E Atwell. 2012a. ‘Open-source boundary-annotated corpus for Arabic speech and language processing.’ Proceedings of LREC’2012 Language Resources and Evaluation Conference.

Brierley, C, M Sawalha, and E Atwell. 2012b. ‘Boundary Annotated Qur'an Corpus for Arabic Phrase Break Prediction.’ Proceedings of IVACS’2012 Inter-Varietal Applied Corpus Studies.

Brierley, C, M Sawalha, and E Atwell. 2012c. ‘Visualisation of Prosody in English and Arabic Speech Corpora.’ Proceedinds of AVML’2012 Advances in Visual Methods for Linguistics.

Brierley, C, M Sawalha, and E Atwell. 2014. Tools for Arabic Natural Language Processing: a case study in qalqalah prosody. Proceedings of LREC’2014 Language Resources and Evaluation Conference, pp. 283-287.

Brierley, C, M Sawalha, B Heselwood, and E Atwell. 2016. A Verified Arabic-IPA Mapping for Arabic Transcription Technology, Informed by Quranic Recitation, Traditional Arabic Linguistics, and Modern Phonetics. Journal of Semitic Studies, 61 (1), pp. 157-186.

Brockett A, E Atwell, O Taylor, and M Page. 1989. ‘An Arabic text database and glossary system for students.’ Proceedings of the Seminar on Bilingual Computing in Arabic and English.

Chelli, A. 2012. Advanced Search/Indexing in Holy Quran. Magister Thesis, National Higher School of Computer Science, Algeria.

Chen, H. 2012. ‘Dark Web: Exploring and Data Mining the Dark Side of the Web.’ Springer.

Clivaz, C. 2013. Digital religion out of the book : the loss of the illusion of the 'original text' and the notion of a 'religion of a book'. Scripta Journal vol.25.

Danso, S, E Atwell, O Johnson, A ten Asbroek, S Soromekun, K Edmond, C Hurt, L Hurt, C Zandoh, C Tawiah, J Fenty, S Etego, S Agyei, and B Kirkwood. 2013. ‘A semantically annotated verbal autopsy corpus for automatic analysis of cause of death.' ICAME Journal of the International Computer Archive of Modern and Medieval English, vol. 37, pp.37-69.

Dukes, K and E Atwell. 2012. 'LAMP: a multimodal web platform for collaborative linguistic analysis.' Proceedings of LREC'2012 Language Resources and Evaluation Conference.

Dukes, K and N Habash. 2010. 'Morphological Annotation of Quranic Arabic.' Proceedings of LREC'2010 Language Resources and Evaluation Conference.

Dukes, K and T Buckwalter. 2010. 'A Dependency Treebank of the Quran using Traditional Arabic Grammar.' Proceedings of INFOS'2010 7th Informatics and Systems.

Dukes, K, E Atwell, and A Sharaf. 2010. 'Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank.' Proceedings of LREC'2010 Language Resources and Evaluation Conference.

Dukes, K, E Atwell, and N Habash. 2013. 'Supervised collaboration for syntactic annotation of Quranic Arabic.' Language Resources and Evaluation Journal, vol. 47, pp.33-62.

El-Beltagy, S, and A Ali. 2013. 'Open issues in the sentiment analysis of Arabic social media: A case study.' Proceedings of IIT'2013 Innovations in Information Technology.

El-Haj, M, U Kruschwitz, C Fox. 2015. Creating language resources for under-resourced languages: methodologies, and experiments with Arabic. Language Resources and Evaluation Journal.

El Hadj, Y, I Al-Sughayeir and A Al-Ansari. 2009. Arabic part-of-speech tagging using the sentence structure. In Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt.

Erradi, A, S Nahia, H Almerekhi, and L Al-kailani. 2012. ArabicTutor: a Multimedia m-Learning Platform for Learning Arabic Spelling and Vocabulary. Proceedings of ICMCS'2012 International Conference on Multimedia Computing and Systems.

Friginal, E and J Hardy. 2014. 'Corpus-based Sociolinguistics: A Guide for Students.' Routledge.

Froud, H, R Benslimane, A Lachkar and S Ouatik. 2010. Stemming and similarity measures for Arabic documents clustering. In Proceedings of ISVC'2010 5th International Symposium on I/V Communications pp. 1-4. IEEE.

Froud, H, A Lachkar and S Ouatik. 2013. Arabic text summarization based on latent semantic analysis to enhance Arabic documents clustering. International Journal of Data Mining and Knowledge Management Process (IJDKP) 3(1) pp79-95.

Garside, R and N Smith. 1997. 'A hybrid grammatical tagger: CLAWS4.' in Garside, R, G Leech and A McEnery (eds.) 'Corpus Annotation:

Linguistic Information from Computer Text Corpora.’ Longman, London, pp. 102-121.

Gehrels, S., 2016. Liquid hospitality: wine as the metaphor. The Routledge Handbook of Hospitality Studies, p.247-259.

Google. 2014. Definition of ‘Artificial Intelligence’.

Greene, B and G Rubin. 1971. ‘Automatic grammatical tagging of English.’ Technical report, Department of Linguistics, Brown University.

Haider, A. 2016. A corpus-assisted critical discourse analysis of the Arab uprisings: evidence from the Libyan case. PhD thesis, University of Canterbury, New Zealand.

Hakkoum, A and S Raghay. 2015a. Ontological approach for semantic modeling and querying the Quran. International Journal on Islamic Applications in Computer Science And Technology, p.37-45.

Hakkoum, A and S Raghay. 2015b. Advanced search in the Qur'an using semantic modeling. In Proceedings of AICCSA’2015 Arab International Conference on Computer Systems and Applications, pp.1-4, Marrakech, Morocco.

Hamdelsayed, A. and E Atwell. 2016a. Islamic Applications of Automatic Question-Answering. Journal of Engineering and Computer Science, 17 (2), pp. 51-57.

Hamdelsayed, M, and E Atwell. 2016. Using Arabic Numbers (Singular, Dual, and Plurals) Patterns To Enhance Question Answering System Results. In Proceedings of IMAN’2016 Islamic Applications in Computer Science and Technologies, Khartoum, Sudan.

Hamdelsayed, M, and E Atwell. 2017. Quran Question Answering System Using Arabic Number Patterns (Singular, Dual, Plural). International Journal on Islamic Applications in Computer Science and Technology (IJASAT), 5 (2), pp. 1-12.

Hammo, B, S Yagi, O Ismail and M AbuShariah. 2016. Exploring and exploiting a historical corpus for Arabic. Language Resources and Evaluation Journal, 50(4), pp.839-861.

Hamoud, B. and E Atwell. 2016a. Using an Islamic Question and Answer Knowledge Base to answer questions about the Holy Quran. International Journal on Islamic Applications in Computer Science And Technology (IJASAT), 4 (4), pp. 20-29.

Hamoud B and E Atwell. 2016b. Quran question and answer corpus for data mining with WEKA. In Proceedings of IEEE Conference of Basic Sciences and Engineering Studies, pp. 211-216, Khartoum, Sudan.

Hamoud B and E Atwell. 2016c. Compiling a Quran Question and Answer Corpus :تجميع مدونة اسئلة واجوبة للقرآن الكر. In Proceedings of ICCA'2016 International Conference on Computing in Arabic, Khartoum, Sudan.

Hassan, H, N Daud, and E Atwell. 2010. ‘Connectives in the World Wide Arabic corpus.’ Proceedings of IVACS'2010 Inter-Varietal Applied Corpus Studies, Leeds, UK.

Hassan, H, N Daud, and E Atwell. 2013. ‘Connectives in the World Wide Web Arabic corpus.’ World Applied Sciences Journal (Special Issue of Studies in Language Teaching and Learning), vol. 21, pp.67-72.

Hassan, S and E Atwell. 2016a. Concept Search Tool for Multilingual Hadith Corpus. International Journal of Science and Research (IJSR), 5(4), pp.1326-1328.

Hassan, S and E Atwell. 2016b. Design Requirements for Multilingual Hadith Corpus. International Journal of Science and Research (IJSR), 5 (4), pp. 494-498.

Hassan, S and E Atwell. 2016. Design and Implementing Of Multilingual Hadith Corpus. International Journal of Recent Research in Social Sciences and Humanities, 3 (2), pp. 100-104.

Itani, M, C Roast and S Al-Khayatt. 2017. Corpora for sentiment analysis of Arabic text in social media. In Proceedings of ICICS’2017 International Conference on Information and Communication Systems, pp. 64-69. IEEE.

Jarrar, M, N Habash, F Alrimawi, D Akra and N Zalmout. 2017. Curras: an annotated corpus for the Palestinian Arabic dialect. Language Resources and Evaluation Journal, 51(3), pp.745-775.

Jilani, A. 2013. Parallel corpus multi stream question answering with applications to the Quran. PhD Thesis, University of Huddersfield, UK.

Kadir, R and A Yauri. 2017. ‘Automated semantic query formulation using machine learning approach’. Journal of Theoretical & Applied Information Technology, 95(12) pp.2761-2775.

Karlsson, F, A Voutilainen, J Heikkila, and A Anttila (eds.). 1995. ‘Constraint Grammar: A Language-Independent System for Parsing Running Text.’ Mouton de Gruyter, Berlin and New York.

Khaliq, B. and Carroll, J. 2013. Induction of root and pattern lexicon for unsupervised morphological analysis of Arabic. Proceedings of IJCNLP’2013 International Joint Conference on Natural Language Processing, Nagoya, Japan

Kilgarriff, A. 2007. ‘Re: [Corpora-List] history of corpus linguistics.’ Corpora-List Archive, 6 January 2007.

Kilgarriff, A, V Baisa, J Bušta, M Jakubíček, V Kovář, J Michelfeit, P Rychlý, and V Suchomel. 2014. ‘The Sketch Engine: ten years on.’ Lexicography journal vol.1(1), pp.1-30.

Kilgarriff, A, F Charalabopoulou, M Gavrilidou, J Jonannessen, S Khalil, S Johansson, R Lew, S Sharoff, R Vadlapudi, and E Volodina. 2013 ‘Corpus-based vocabulary lists for language learners for nine languages.’ Proceedings of LREC’2013 Language Resources and Evaluation Conference.

LDOCE Online. 2014. Definition of ‘model.’ Longman Dictionary Of Contemporary English, Online.

Leech, G, R Garside, and E Atwell. 1983a. ‘Recent developments in the use of computer corpora in English language research.’ Transactions of the Philological Society, 1983, pp.23-40.

Leech, G, R Garside, and E Atwell. 1983b. ‘The Automatic Grammatical Tagging of the LOB Corpus.’ ICAME Journal: International Computer Archive of Modern and Medieval English Journal, vol. 7, pp.13-33.

Mahmoud, M. and L Hassan. 2013. Artificial intelligence techniques for extracting individuals recitation of the Holy Quran from its combinations. In Proceedings of Advances in Information Technology for the Holy Quran and Its Sciences, pp. 292-297, Taibah University, Medina, Saudi Arabia.

Makhambetov, O, A Makazhanov, Z Yessenbayev, B Matkarimov, I Sabyrgaliyev and A Sharafudinov. 2013. Assembling the Kazakh Language Corpus. In Proceedings of EMNLP’2013 Empirical Methods in Natural Language Processing, pp. 1022-1031.

Malmasi, S, and M Dras. 2014. Arabic native language identification. In Proceedings of EMNLP 2014 Workshop on Arabic Natural Language, Doha, Qatar.

Malmasi, S and M Dras. 2015. Multilingual native language identification. Natural Language Engineering Journal, pp.1-53.

Merakchi, K and M Rogers. 2013 ‘The translation of culturally bound metaphors in the genre of popular science articles: A corpus-based case study from Scientific American translated into Arabic.’ Intercultural Pragmatics journal, vol.10(2), pp.341-372.

Mohammed, M and N Omar. 2011. ‘Rule based shallow parser for Arabic language.’ Journal of Computer Science, vol.7(10), pp.1505-1514.

Mohamed, R, M Ragab, H Abdelnasser, N El-Makky and M Torki. 2015. Al-Bayan: A Knowledge-based System for Arabic Answer Selection. In Proceedings of SemEval’2015 Semantic Evaluation, pp. 226-230.

Mohit, B, A Rozovskaya, N Habash, W Zaghouani and O Obeid. 2014. The First QALB Shared Task on Automatic Text Correction for Arabic. In the proceedings of the EMNLP 2014 Workshop on Arabic Natural Language, 25 October 2014, Doha, Qatar.

Muhammad, A. 2012. Annotation of conceptual co-reference and text mining the Quran. PhD Thesis, School of Computing, University of Leeds UK.

Mukhtar, T, H Afzal and A Majeed. 2012. Vocabulary of Quranic Concepts: A semi-automatically created terminology of Holy Quran. In Proceedings of INMIC’2012 International Multitopic Conference, pp. 43-46, Islamabad, Pakistan.

Onyenwe, I. 2017. Developing Methods and Resources for Automated Processing of the African Language Igbo. PhD thesis, Department of Computer Science, University of Sheffield.

Ouda, K. 2015. QuranAnalysis: A Semantic Search and Intelligence System for the Quran. MSc Thesis, School of Computing, University of Leeds.

Panju, M. 2014. 'Statistical Extraction and Visualization of Topics in the Quran Corpus.' MMath Thesis, University of Waterloo.

Rabiee, H. 2011. Adapting Standard Open-Source Resources To Tagging A Morphologically Rich Language: A Case Study With Arabic. Proceedings of RANLP'2011 Recent Advances in Natural Language Processing.

Roberts, A, L Al-Sulaiti, and E Atwell. 2005. 'aConCorde: towards a proper concordance of Arabic.' Proceedings of CL'2005 Corpus Linguistics.

Romli, T, A Hassan and H Mohamad. 2016. Equivalent Malay-Arabic Data Corpus Collection. International Editorial and Advisory Board Journal, 4(1), p.66.

Roberts, A, L Al-Sulaiti, and E Atwell. 2006 'aConCorde: Towards an open-source, extendable concordancer for Arabic.' Corpora journal, vol. 1, pp. 39-57.

Saad, M, and W Ashour. 2010. Arabic Text Classification Using Decision Trees. Proceedings of CSIT'2010 12th international workshop on Computer Science and Information Technologies, Moscow and Saint-Petersburg, Russia.

Saad, S, N Salim and S Zainuddin. 2011. An early stage of knowledge acquisition based on Quranic text. In Proceedings of STAIR'2011 Semantic Technology and Information Retrieval, pp. 130-136, Putrajaya, Malaysia.

Saad, S, N Salim and H Zainal. 2013. Rules and Natural Language Pattern in Extracting Quranic Knowledge. In Proceedings of Advances in Information Technology for the Holy Quran and Its Sciences, pp. 407-412, Taibah University, Medina, Saudi Arabia.

Sawalha, M and E Atwell. 2008. 'Comparative evaluation of Arabic language morphological analysers and stemmers.' Proceedings of COLING'2008 Computational Linguistics.

Sawalha, M and E Atwell. 2009. 'Linguistically informed and corpus informed morphological analysis of Arabic.' Proceedings of CL'2009 Corpus Linguistics.

Sawalha, M and E Atwell. 2010a. 'Fine-Grain Morphological Analyzer and Part-of-Speech Tagger for Arabic Text.' Proceedings of LREC'2010 Language Resources and Evaluation Conference.

Sawalha, M and E Atwell. 2010b. 'Constructing and Using Broad-Coverage Lexical Resource for Enhancing Morphological Analysis of Arabic.' Proceedings of LREC'2010 Language Resources and Evaluation Conference.

Sawalha, M and E Atwell. 2011. 'Morphological analysis of classical and modern standard Arabic.' Proceedings OF ICCA'2011 International Computing Conference in Arabic.

Sawalha, M, C Brierley, and E Atwell. 2012a. ’Prosody prediction for Arabic via the open-source boundary-annotated Qur'an corpus.’ Journal of Speech Sciences, vol. 2, pp.175-191.

Sawalha, M, C Brierley, and E Atwell. 2012b. Automatic Analysis of Phrase-Break Prediction for Arabic نصوص في والابتداء للوقف الآلي التحليل والكلاسيكية الحديثة العربية اللغة Proceedings of the International Computing Conference in Arabic (ICCA).

Sawalha, M, E Atwell and C Brierley,. 2012c. Open-Source Boundary-Annotated Qur’an Corpus for Arabic and Phrase Breaks Prediction in Classical and Modern Standard Arabic Text. Journal of Speech Sciences, 2 , pp. 175-191.

Sawalha, M, C Brierley, and E Atwell. 2012d. ‘Predicting phrase breaks in classical and modern standard Arabic text.’ Proceedings of LREC’2012 Language Resources and Evaluation Conference.

Sawalha, M and E Atwell. 2012e. ‘Visualization of Arabic Morphology.’ Proceedings of AVML’2012 Advances in Visual Methods for Linguistics.

Sawalha, M and E Atwell. 2013a. ‘Accelerating the processing of large corpora: using grid computing for lemmatizing the 176 million words Arabic Internet Corpus.’ Proceedings of WACL’2 2nd Workshop of Arabic Corpus Linguistics.

Sawalha, M and E Atwell. 2013b. ‘A standard tag set expounding traditional morphological features for Arabic language part-of-speech tagging.’ Word Structure journal, vol. 6, pp.43-99.

Sawalha, M and E Atwell. 2013c. ‘ Comparing morphological tag-sets for Arabic and English.’ Proceedings of CL’2013 Corpus Linguistics.

Sawalha, M, E Atwell, and M Abushariah. 2013d. SALMA: Standard Arabic Language Morphological Analysis. In Proceedings of ICCSPA’2013 International Conference on Communications, Signal Processing, and their Applications, pp.1-6.

Sawalha, M, C Brierley, and E Atwell. 2014a. Automatically generated, phonemic Arabic-IPA pronunciation tiers for the boundary annotated Qur’an dataset for machine learning. In Proceedings of LRE-Rel’2 2nd Workshop on Language Resource and Evaluation for Religious Text, pp. 42-47.

Sawalha, M, C Brierley, and E Atwell and J Dickins. 2014b. Text Analytics and Transcription Technology. In Proceedings of IMAN’2014 Islamic Applications in Computer Science And Technology, Amman, Jordan.

Sawalha, M, C Brierley, and E Atwell and J Dickins. 2017. Text Analytics and Transcription Technology for Quranic Arabic. International Journal on Islamic Applications in Computer Science and Technology (IJASAT), 5 (2), pp. 45-51.

Seddik, K, A Farghaly and A Fahmy. 2015. Arabic Anaphora Resolution: Corpus of the Holy Quran Annotated with Anaphoric Information. International Journal of Computer Applications, 124(15).

Sharaf, A and E Atwell. 2009. ‘A Corpus-based Computational Model for Knowledge Representation of the Quran’, Proceedings of CL’2009 Corpus Linguistics.

Sharaf, A and E Atwell. 2012a. ‘QurAna: Corpus of the Quran annotated with Pronominal Anaphora.’ Proceedings of LREC’2012 Language Resources and Evaluation Conference.

Sharaf, A and E Atwell. 2012b. ‘QurSim: A corpus for evaluation of relatedness in short texts.’ Proceedings of LREC’2012 Language Resources and Evaluation Conference.

Sharoff, S. 2006. ‘Open-source corpora: using the net to fish for linguistic data.’ International Journal of Corpus Linguistics 11 (4), pp. 435–62.

Shmeisania, H, S Tartirb, A Al-Nassaanc and M Najid. 2014. Semantically Answering Questions from the Holy Quran. In Proceedings of IMAN’2014 Islamic Applications in Computer Science and Technology, Amman, Jordan.

Tabrizi, A, and R Mahmud. 2013. ‘Issues of coherence analysis on English translations of Quran.’ Proceedings of ICCSPA’2013 International Conference on Communications, Signal Processing, and their Applications.

Wiechmann, D and S Fuhs. 2006. ‘Concordance Software.’ Corpus Linguistics and Linguistics Theory journal, vol.2, pp109-130

Wikipedia. 2014a. Definition of ‘Machine Learning’

Wikipedia. 2014b. Definition of ‘Amazon Mechanical Turk’

Wood, P. 2016. The Pen and The Sword: Reporting ISIS. Discussion paper, Shorenstein Center on Media Politics and Public Policy.

Yusof, R, R Zainuddin, M Baba, and Z Yusoff. 2010. ‘Quranic words stemming.’ Arabian Journal for Science and Engineering, vol.35(2), pp.37-49.

Zaghouani, W, T Zerrouki and A Balla. 2015. SAHSOH@ QALB Shared Task: A Rule-Based Correction Method of Common Arabic Native and Non-Native Speakers’ Errors. In Proceedings of ANLP’2015 Arabic Natural Language Processing Workshop, p. 155.

Zeroual, I and A Lakhouaja. 2016. A new Quranic Corpus rich in morphosyntactical information. International Journal of Speech Technology, 19(2), pp.339-346.

Zouaghi, A, L Merhbene and M Zrigui. 2011. Word Sense disambiguation for Arabic language using the variants of the Lesk algorithm. In Proceedings of WORLDCOMP’2011, pp.561-567.