Using the Web to Model Modern and Quranic Arabic.

Eric Atwell, School of Computing, Leeds University

ABSTRACT:

This chapter is not about a specific Arabic linguistics research project, but about using the Web to collect and promote computing resources for Arabic corpus linguists. Natural Language Processing research can supply useful corpus-based resources in many domains, from healthcare to counter-terrorism. An initial survey found few freely available Arabic corpus linguistics resources; but we found that Machine Learning could be harnessed to adapt generic NLP techniques to Arabic. This required an Arabic text training set, so we developed the first freely downloadable Corpus of Contemporary Arabic, and Arabic concordance visualisation toolkit. We also developed tools for Modern Arabic text analytics: morphological analysis, stemming, and tagging, and Arabic discourse analysis. We have also extended our analytics techniques to Classical Arabic in the Quran, including question-answering, knowledge representation, and syntactic annotation. The Corpus of Contemporary Arabic and the Quranic Arabic Corpus have been widely re-used in Arabic Corpus Linguistics and Computational Linguistics research, for training and evaluation. Our Quranic Arabic Corpus website has become a widely used resource, not just by Arabic and Quranic researchers, but by the 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, Artificial Intelligence, and Corpus Linguistics.

1. Introduction

This chapter is not about a specific Arabic corpus, nor about the use of a corpus in an Arabic linguistics research project. I work in the School of Computing within the Faculty of Engineering at Leeds University, and engineers build things for others to use; so our contribution to Arabic Corpus Linguistics has been to develop a range of Arabic language computing resources – corpora and software tools – for as wide a range of users as possible, including not just linguists but also computing and artificial intelligence researchers, religious scholars, and the general public.

In the School of Computing at Leeds University, we are not Arabic linguists, but we enjoy working with Arabic linguists. To explain our motivation for contributing to Arabic Corpus Linguistics, I will outline some examples of Artificial Intelligence and Corpus Linguistics research at Leeds University where we have worked with “end-users” in interesting and challenging domains. We may have little or no expertise in a domain, but nevertheless can apply Machine Learning to text data from the domain to produce useful results.

Next I explain what I mean by the phrases ‘Using the Web’, ‘to Model’, ‘Modern (Arabic)’, and ‘Quranic Arabic’. This leads into a summary of Web-based software and corpus datasets developed by Leeds University researchers, covering Modern Standard Arabic, and Classical Arabic of the Quran.

I end with some ideas for further development of Arabic Corpus Linguistics resources. Most Arabic linguistic research focuses on Modern Standard Arabic and modern Arabic dialects. However, modern Arabic linguists, lexicographers and language teachers need to recognize and deal with the religious terms and quotations from Quranic Arabic that can appear in modern Arabic texts. Furthermore, Quranic Arabic Corpus research may be a minority interest in linguistics, but it has huge potential for impact on society and the general public including billions of Muslims worldwide who want to study and understand the Quran.

2. Artificial Intelligence for Corpus Linguistics

I like the concise Google definition of Artificial Intelligence as ‘the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.’ (Google 2014). Much early AI research involved trying to encode human expert knowledge as sets of formal rules in Expert Systems or Knowledge Based Systems, e.g. (Atwell 1993a,b). Most current AI theory and systems incorporate Machine Learning: ‘a scientific discipline that deals with the construction and study of algorithms that can learn from data. Such algorithms operate by building a model based on inputs and using that to make predictions or decisions, rather than following only explicitly programmed instructions.’ (Wikipedia 2014a). A corpus is a collection of texts, representing a language use, topic or task; corpus linguists study some specific feature of a language using a corpus to provide empirical evidence of the language feature. A corpus is also useful as a training and/or test data-set for Machine Learning, particularly if annotated or tagged with linguistic information by linguists. An example of the trend towards Machine Learning relevant to Corpus Linguistics is development of Part-of-Speech taggers. PoS-taggers based on explicit grammar rules written by linguists, such as TAGGIT (Greene and Rubin, 1971) or Constraint Grammar taggers (Karlsson et al 1995), have largely been supplanted by PoS-taggers based on Machine Learning models trained with a PoS-tagged corpus, for example CLAWS, the Constituent Likelihood Automatic Word-tagging System (Leech et al 1983a,b, Garside and Smith 1997).

An example use of Machine Learning to learn from a corpus is for classifying cause of death in Verbal Autopsies (Danso et al 2013). In some developing countries, when someone dies a doctor may not be available to diagnose and certify cause of death. A Verbal Autopsy can be used as a proxy to ascertain likely cause of death via an interview with a close relative of the deceased, for example a mother after her baby died. Bodies such as the WHO World Health Organization use Verbal Autopsies to gather statistics on major causes of death, to inform health policies. Danso et al (2013) worked with a corpus of 11,741 Verbal Autopsies from Ghana, which had been sent to the London School of Hygiene and Tropical Medicine, where doctors diagnosed the likely cause of death. Thus, each text in the Verbal Autopsy corpus was tagged with its class: cause of death. A Machine Learning classifier was used to learn patterns linking features of each Verbal Autopsy to cause of death. This classifier can in principle be used to predict cause of death in future Verbal Autopsy datasets, without need for doctors; the classifier is not good enough to use in front-line health care, to confidently diagnose each individual case, but its overall predictions of prevalence of major causes of death may be sufficiently accurate to guide health funding policy.

Another example of Machine Learning to learn from a corpus is for classifying “suspiciousness” of texts. The project Detecting Terrorist Activities: Making Sense (Atwell 2011, Brierley et al 2013) aimed to develop systems to better manage data collected in connection with alleged terrorist plots. Very few of the texts collected from subjects of interest are actually suspicious; the task is ‘like looking for a needle in a haystack’ (Zolfagharifard, 2009). I prefer the analogy of looking for threads in a haystack: Machine Learning aims to find ‘interesting’ texts from the corpus, and ‘threads’ linking them. To train a Machine Learning classifier, we used a training corpus where ‘interesting’ texts are marked or tagged.

In the School of Computing at Leeds University, we are not Arabic linguists; so, how can we be involved in Arabic corpus linguistics research? Machine Learning requires data tagged and classified by experts, to train an automated classifier algorithm. We do not have to “understand” the data or the tagging to apply Machine Learning; we just work with the experts to provide classifiers, and then leave the experts to evaluate the accuracy and usefulness of the results. Of course, I want to work in application areas I find interesting, challenging and useful. I am not a clinician, but maybe Machine Learning can help classify cause of death in Verbal Autopsies; and assisting WHO to target life-threating diseases is surely worthwhile. I am not a counter-terrorism expert, but maybe Machine Learning can help detect terrorist threads in data, and detecting terrorist activities is a fascinating challenge of national and international importance. Arabic is a major international language, yet has fewer computational and corpus linguistics resources than major European languages such as English, French, German, Spanish; this invites and merits research on Arabic corpus linguistics resource development. An initial survey (Atwell et al 04) found few publicly available Arabic language computing resources; but we found that Machine Learning could be used to adapt generic NLP techniques to Arabic (Abu Shawar and Atwell 04, 05a,b). The Classical Arabic of the Quran is of even wider interest, since billions of Muslims worldwide want to understand the Quran as the key source text of their religion; corpus resources developed for the Quran will be useful not only to linguists but to a much broader user community.

3. Using the web, to model, modern Arabic, and Quranic Arabic.

The Web is a very useful resource for Arabic Corpus Linguistics, as source of Arabic corpus data. For example, the Corpus of Contemporary Arabic (Alsulaiti and Atwell 2006) was collated by selecting the genres to be included, and then scouting for websites with appropriate data. A more recent trend in Corpus Linguistics is to use a web-crawler tool like BootCat (Baroni and Bernadini 2004) to automate harvesting of web-page text. At Leeds, Sharoff (2006) harvested large corpora in many languages from web-pages, including the 176 million words Arabic Internet Corpus, which was then automatically lemmatized as case-study of high-performance computing (Sawalha and Atwell 2013) to enable concordance and collocation analysis by Arabic lemma.

Another use of the Web is to annotate a corpus, via ‘crowd sourcing’. The morphological and syntactic tagging of the Quranic Arabic Corpus was achieved by starting with an automated tagger, butting the results online, and then inviting website users to report mistakes and propose corrections. Volunteers can build a shared resource if they share an interest, and the Quran attracts many more interested volunteers than other online corpus resources. For Arabic corpus linguistics tagging projects which are not as “interesting’ to volunteers, researchers can hire and manage online annotators via websites such as Amazon Mechanical Turk, ‘an online crowdsourcing marketplace’ (Wikipedia, 2014b).

A third use of the Web is to publicize and promote re-use of corpora. Often corpus linguists collect a corpus with their own research questions in mind, and then at the end of their project may (or may not) consider how to make the corpus re-usable by others. There are a range of options, with pros and cons: donating to an established Web-based corpus repository, such as ICAME, ELRA, LDC; depositing with a conference website repository, such as the LREC language resources map; setting up a project website and adding a corpus download page, such as the Corpus of Contemporary Arabic website; setting up a customised corpus website for users, with search and corpus use tools, such as the Quranic Arabic Corpus website; donating the corpus to existing specialist corpus websites already hosting other related corpora, such as the ArabiCorpus website; donating to an academic corpus-user website with resources for a wider range of languages, such as Intellitext (Wilson et al 2010); and donating to a commercial corpus-user website catering for the language technology industry as well as academics, such as SketchEngine (Kilgarriff et al 2014). To achieve widest re-use of your corpus (and subsequent citations), it is best to devote time to what Machine Learning researchers call an ‘ensemble’ approach, combining several methods; and that to maximize impact, you should also publish research papers in a wide range of conferences and journals beyond your initial research focus, to add academic credibility as well as reaching a wider audience. The web can then be harnessed to maximize the reach and subsequent citation of your conference and journal papers (and hence re-use of your corpus) by adding them to your own website, your University open-access repository, and a range of reference-sharing “academic social-networking” websites such as Academia.edu, ResearchGate.net, ResearchFish.com, LinkedIn.com, Mendeley.com, Zotero.org etc.

Arabic corpus and tool builders can use the web in these three ways, to build, annotate and promote new Arabic corpus resources. But probably the most common use of the Web by Arabic corpus linguists is not to create new corpora, but to find and use existing online Arabic corpora and analysis tools. Early surveys (Atwell et al 04, Al-Sulaiti and Atwell 2006) found few freely-available online Arabic corpus linguistics resources; but today a Google search for “Arabic corpus” produces ‘About 87,900 results’ including many links to online free-to-use resources.

In the title of this chapter, I use the verb ‘model’ in the sense ‘to do a computer representation or scientific description of a situation or event’ (LDOCE online, 2014). This covers use of a corpus to guide development of linguistic theories or models; and also covers use of a corpus as ‘training data’ for Machine Learning of a language analysis model. Traditional Arabic grammar descriptions in classical textbooks can be modelled in computer representations. In the Quranic Arabic Corpus, narrative text descriptions of the grammar of each verse has been taken from Tafsirs, Islamic scholarly textbooks of commentary and analysis of the Quran, and this narrative text has been modeled in formal syntax tree diagrams, to produce a Classical Arabic Treebank. Similarly, the SALMA morphological analysis toolkit implements a computational model based on traditional Arabic morphology theory from established Arabic grammar and morphology textbooks.

Most of the chapters in this book on Arabic Corpus Linguistics deal with corpora of various modern forms of Arabic, including Modern Standard Arabic and modern dialects; and Arabic corpus linguistics researchers can find a growing range of modern Arabic corpora, including news, web, dialects, and even quite specialized genres such as the ‘Dark Web’ terrorism corpus (Chen, 2012). This reflects the importance of Arabic as a modern international language. The Quran, the core holy text of Islam, is written in Classical Arabic of around 1400 years ago, a time when Latin was still widely used in Europe, and well before English appeared as a new language. Despite its antiquity, Quranic Arabic is of immense importance even in the modern world, because all Muslims (about one quarter of the world’s population) are required to learn and memorize the Quran in its original language. Words, phrases and quotes from the Quran and related Classical Arabic Islamic texts can readily be found in Modern Arabic corpora, in effect a kind of “code switching”. And Quran vocabulary and quotations even crop up in other language corpora: Muslims with mother tongues other than Arabic are also required to study the Quran in its Classical Arabic form, in case some of the meaning is lost in translation.

4. Arabic corpus linguistics research at Leeds University

Leeds University is unique in its range of departments and research units active in Arabic corpus linguistics research, including Languages Cultures and Societies, Computing, Arabic Islamic and Middle Eastern Studies (AIMES), Translation Studies, Linguistics and Phonetics, and Institute for Artificial Intelligence and Biological Systems (I-AIBS). This broad range of expertise has led to interdisciplinary collaboration and cross-fertilization of ideas, resulting in a range of Arabic corpus linguistics research projects.

4.1 ABC: Arabic By Computer

Arabic corpus linguistics research collaboration at Leeds University started with a project to collect a Modern Standard Arabic texts for use in computer-aided Arabic language teaching: the ABC Arabic By Computer project to develop an Arabic text database and glossary system for Arabic language students (Brockett et al 1989). At the time, a major challenge was editing and display of Arabic text, requiring specialist Apple Macintosh hardware and software. Another practical challenge was giving students access to the ABC resources: the University computing provision did not include language teaching, as computers were assumed to be only used for science and engineering research. This inspired us to investigate methods for free, open access to Arabic corpus linguistics resources for teaching and research.

4.2 Arabic corpus-trained chatbots

Researchers at Leeds have developed computing tools and resources for a range of languages. An unusual use of corpora was to train web-based machine-learning chatbot systems (Abu Shawar and Atwell 2005a), and using a corpus-trained web-based chatbot system as a tool to animate and explore the corpus (Abu Shawar and Atwell 2005b). Our system could be trained to chat in the language of a given training corpus; as an example, we trained the chatbot with the Quran, resulting in a web-based Arabic chatbot giving answers from the Quran (Abu Shawar and Atwell 2004). Another version of the chatbot was trained on an Arabic computing FAQ corpus, to provide answers to Frequently Asked Questions (Abu Shawar and Atwell 2009).

4.3 CCA Corpus of Contemporary Arabic

Machine Learning can be applied to adapt generic NLP techniques to Arabic. This required an Arabic text training set, so we developed the first freely downloadable million-word Corpus of Contemporary Arabic (Al-Sulaiti and Atwell 2005, 2006). The CCA was designed to be comparable to the million-word LOB and Brown corpora of Modern British English and American English published texts (Leech et al1983a); but rather than slavishly copying the exact set of text genres in LOB and Brown, we surveyed potential users in Arabic language teaching and Arabic text analytics, to identify user preferences for the distribution of genres to be included. The Corpus of Contemporary Arabic has been used for research, for example in learning Arabic spelling and vocabulary (Erradi et al 2012), Arabic lexical profiling (Attia et al 2011), the translation of culturally bound metaphors in the genre of popular science articles (Merakchi and Rogers 2013), lexical differences in world affairs and sports sections in Arabic newspapers (Abdul Razak 2011), and corpus-based sociolinguistics (Friginal and Hardy 2014).

4.4 aConCorde concordancer for Arabic

We realized that corpus concordance tools available at the time were not designed to handle the unusual properties of Arabic script, including: non-Roman character set, several rival encoding standards, some characters vary their shape depending on context, vowels are often omitted resulting in spelling variations, varying use of punctuation, text flows right-to-left not left-to-right, and concordance ‘before’ and ‘after’ windows need to be swapped. So, we developed aConCorde, a freely downloadable open-source extendable concordance program specifically for Arabic corpus linguistics (Roberts et al 2005, 2006).   A review of concordancing software at the time (Wiechmann and Fuhs 2006) praised aConcorde for ‘… providing comprehensive support for working with Arabic texts. This is reflected on several levels: the user interface can be switched to Arabic, the character encoding supports Unicode as well as specific Arabic fonts and text orientation can be mirrored vertically.’ A decade later, an evaluation of concordance tools for Arabic corpora (Alfaifi and Atwell 2014a) found that most other concordancers still do not handle Arabic text well; so aConCorde is still used for Arabic corpus linguistics research, for example exploration of common lexical patterns in Arabic text (Ali 2012), key words and phrases (El-Haj et al forthcoming), and to identify crime patterns from an Arabic crime news report corpus (Alruily 2012).

4.5 Corpus-based Arabic language teaching

We have been able to experiment with the use of web-based corpora, concordance and chatbots in the teaching of Arabic (Al-Sulaiti et al 2005, 2007), for example corpus-based vocabulary lists for language learners (Kilgarriff et al 2013). We had access to students as well as language teachers in the department of Arabic at Leeds University, keen to use Arabic corpus linguistics resources in learning and teaching. Also, the local Muslim community ran a Saturday school for children to learn Arabic to read and understand the Quran, and they enjoyed the novelty of a web-based corpus-trained chatbot giving answers in Quranic Arabic (Abu Shawar and Atwell 2004).

4.6 Arabic Web-as-Corpus

The British National Corpus (BNC) became an established gold standard for English corpus linguistics in the 1990s; but for other languages, funding and expertise were not available for a large general corpus of the size of BNC, 100 million words. But then the Web-as-Corpus approach was developed (Baroni and Bernardini 2004). In essence, this requires you to select a representative list of words in your target language; then use these in a corpus-harvester program, which sends subsets of the words as search-terms to Google, Yahoo, Bing or other web search engine, downloads the hit web-pages, scrapes the text and collates results in a Corpus. At Leeds University, (Sharoff 2006) used the Web-as-Corpus method to collect Internet corpora for Arabic, Chinese, French, German, Italian, Spanish, Polish and Russian, freely accessible via a concordance and collocation search interface at http://corpus.leeds.ac.uk/internet.html. This includes the 176-Million-word Arabic Internet Corpus, which was subsequently lemmatized using the SALMA morphological analysis toolkit (Sawalha and Atwell 2013a).

We later collected a smaller World Wide Arabic Corpus, analogous to the World Wide English Corpus (Atwell et al 2007), comprising 200,000-word subcorpora from each country, to capture country-by-country dialect variation. This was used to study 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). We also used the Web-as-Corpus approach for collecting a specialized ‘corpus’ of texts for university-level teaching about Islam (Atwell et al 2011).

4.7 Arabic corpus Part-of-Speech tagging and morphological analysis

I had worked on the LOB Corpus tagging project (Atwell 1982, Leech et al 1983b), and I wanted to apply my experience of English corpus tagging to Arabic (Atwell 2008, Atwell et al 2008), to develop Arabic corpus annotation software for part-of-speech tagging, morphological analysis, and more. A first step was a comparative evaluation of existing Arabic language morphological analyzers and stemmers, by hand-analysis of small ‘gold standard’ samples of the Quran and modern news text, to compare with and evaluate the outputs of automated analyzers (Sawalha and Atwell 2008). We also compared linguistically informed and corpus informed approaches to morphological analysis of Arabic (Sawalha and Atwell 2009). This guided the development of a new fine-grained morphological analyzer and Part-of-Speech tag-set and tagger software for Arabic text (Sawalha and Atwell 2010a): the SALMA tagger. The name SALMA could stand for either ‘Sawalha Atwell Leeds Morphological Analysis’ (Sawalha and Atwell 2013a), or ‘Standard Arabic Language Morphological Analysis’ (Sawalha et al 2013); but it originated as a suggested name for a granddaughter …

The SALMA tagger is part of a broader Arabic corpus analysis toolkit, including a standard tag-set expounding traditional morphological features for Arabic language part-of-speech tagging (Sawalha and Atwell 2013); this involved formal analysis of traditional Arabic grammarians’ theoretical research applied to PoS-tagging, giving a detailed and comprehensive ontology of established Arabic word structure theory. Several other Arabic PoS-tagsets have been developed for specific tasks, but generally are adapted from English models, and/or cover only a limited subset of traditional treatises on Arabic morphology. We provided an online benchmark for comparison and evaluation of task-specific PoS-tagsets. Arabic corpus linguistics research is booming but fragmented; this work will enable Arabic corpus PoS-tagging research to be grounded on established traditional Arabic linguistic theory. The SALMA tag-set and tagging software are complemented with a broad-coverage Arabic lexicon derived from open-source online lexical resources and traditional Arabic dictionaries (Sawalha and Atwell 2010b), and tools for visualization of Arabic morphology (Sawalha and Atwell 2012). To verify its robustness for processing large corpora, we applied the SALMA-tagger to the 176-million-word Arabic Internet Corpus (Sawalha and Atwell 2013b). The SALMA corpus analysis toolkit has been used for corpus linguistics research, for example in developing vocabulary lists for Arabic language learners (Kilgarriff et al 2014), learning Arabic spelling and vocabulary (Erradi et al 2012), Arabic grammatical analysis (Rabiee 2011), analysis of Arabic soclal media (El-Beltagy and Ali 2013).

4.8 Discourse Treebank for Modern Standard Arabic

A different sort of tagging is required for Arabic discourse analysis: Al-Saif and Markert (2010) developed the Arabic Discourse Treebank, a news corpus of 537 news texts where all 5651 discourse connectives are identified and annotated with the discourse relations they convey as well as with the two arguments they relate. This required the development of a discourse annotation tool for Arabic text, and a website for dissemination of the Arabic Discourse Treebank.

4.9 Arabic Learner Corpus

The Arabic Learner Corpus (ALC) is a freely downloadable resource for Arabic language teaching (Alfaifi and Atwell 2013, Alfaifi et al 2014), comprising a collection of written and spoken materials from learners of Arabic in Saudi Arabia. The ALC data was captured in 2012 and 2013. It includes 282,732 words, 1585 materials (written and spoken), produced by 942 students from 67 nationalities, and 66 different L1 backgrounds. Average length of a text is 178 words. The metadata information, in English and Arabic, enables researchers to identify characteristics of text and its producer in each transcription, which add more depth to the data analysis. The original hand-written sheets are also downloadable as scanned PDF files. MP3 audio recordings (3.5 hours in total) of learners who granted permission are also available to download. Corpus filenames indicate the key characteristics of the text; e.g. S038_T2_M_Pre_NNAS_W_C shows student identifier number, text number, author gender, level of study, nativeness, text mode, and place of text production. The ALC is downloadable from our ArabicLearnerCorpus.com website, and also searchable online via SketchEngine (Kilgarriff et al 2014).

We also developed a new error tag-set for error annotation of the Arabic Learner Corpus (Alfaifi et al 2013, Alfaifi and Atwell 2012, 2014b). This is informed by error tag-sets used in other Learner Corpus projects, but adapted to the specific types of errors made by learners of Arabic.
.

4.10 Arabic phonetic and prosodic tagging

Another sort of “tagging” is phonetic transcription of texts to be read out loud, such as the Quran. Phonetics researchers use the IPA International Phonetic Alphabet to transcribe spoken texts in any language; however, Arabic script is not entirely phonetic, and there is not a simple one-to-one mapping between Arabic written characters and IPA symbols. So, we developed a verified mapping between Arabic script and IPA International Phonetic Alphabet, for automated Arabic phonetic transcription; this was informed by analysis of Quranic recitation, traditional Arabic linguistics, and modern phonetics (Brierley et al forthcoming, Sawalha et al 2014). The Quran traditional source text also includes prosodic symbols denoting several types of pause or phrase boundary. Muslins are required to follow “Tajweed” when reading aloud verses from the Quran; Tajweed refers to the rules governing pronunciation and prosody during recitation of the Qur'an. The IPA mapping and prosodic symbols were used in phonetic and prosodic annotation of the Quran, to produce the Boundary-annotated Quran Corpus (Brierley et al 2012a,b). This prosody-tagged Arabic corpus can help non-Arabic-speakers to learn correct Quran recitation; it can also be used to train a prosody tagger for other Arabic texts, including modern standard Arabic (Sawalha et al 2012a,b),

4.11 Corpus-based comparison of English and Arabic

Given my background in English corpus linguistics (eg Leech et al 1983a,b), I am interested in corpus-based comparisons of English and Arabic. These include: Arabic influences on Arab English, the variety of English in use in the Arab world (Atwell et al 2009); comparing morphological and Part-of-Speech tag-sets for English and Arabic (Atwell 2008, Sawalha and Atwell 2013c); and visualization of prosody in English and Arabic corpora (Brierley et al 2012c)

4.12 Quranic Arabic Corpus

Since the open-access release of the Corpus of Contemporary Arabic, a growing number and variety of open-access modern Arabic corpora have appeared. However, the Quran and Classical Arabic have attracted much less interest, at least among corpus linguists. The best known Classical Arabic project is the Quranic Arabic Corpus (Dukes et al 2013), a collaboratively constructed linguistic resource initiated at the University of Leeds, with multiple layers of annotation including part-of-speech tagging, 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 of Quranic concepts. The motivation behind this work is to produce a resource that enables further analysis and understanding of the Quran. This project contrasts with other Arabic treebanks by providing a deep linguistic model based on the historical traditional grammar known as i'rāb. By adopting this well-known canon of Quranic grammar, it is possible to encourage online annotation by Arabic linguists and Quranic experts. This new approach to linguistic annotation of an Arabic corpus is via automatic rule-based tagging, initial manual verification, and online supervised collaborative proofreading. The Quranic Arabic Corpus morphological tagging project relied on approximately 100 unpaid volunteer annotators each suggesting corrections to existing linguistic tagging. A small number of expert annotators had a supervisory role, allowing them to review or veto suggestions made by other collaborators. The Quran also benefits from a large body of existing historical grammatical analysis, in traditional commentaries on the Quran by Islamic scholars. The challenges of annotating Quranic Arabic online required a custom-built linguistic software platform to aid collaborative annotation: 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 Classical Arabic, including Arabic word stemming (Yusof et al 2010), Arabic grammatical analysis (Mohammed and Omar 2011, Rabiee 2011), Arabic stylometrics (Alqurneh et al forthcoming), coherence analysis in Arabic translation studies (Tabrizi and Mahmud 2013), summarization (El-Haj et al Forthcoming), oral-formulaic analysis (Bannister 2014), and has also had significant social impact: the million visits a year include non-Arabic-speakers, gaining a deeper insight into the original Classical Arabic text through the linguistic annotations.

4.13 QurAna: Quran pronoun anaphoric co-reference

QurAna (Sharaf and Atwell 2012a) is a large-scale annotation of the Quran as corpus, where each personal pronoun is tagged with its antecedent, the word or phrase it refers to, in the preceding (or occasionally following) text; and also with its “meaning”, the person, entity or concept that pronoun and antecedent stand for, in a Quran ontology or set of people, entities and concepts. QurAna has a comparatively large number of pronouns tagged with antecedent information, over 24,500 pronouns; and an ontology of over a thousand persons, entities and concepts, all nouns or phrases which are referred to by personal pronouns. Deciding on the reference of a personal pronoun is not always straightforward; but for the Quran, we can use Tafsirs or scholarly commentaries to guide the annotator. In uncertain cases, we followed the co-reference analysis in the Tafsir of Ibn Kathir, a highly trusted and acclaimed Islamic reference work; we have at least as much confidence in this as in the alternative of relying on inter-annotator agreement between two casual-worker annotators. The pronoun anaphoric reference annotation is first freely downloadable corpus tagging of its kind for Classical Arabic as well as Modern Arabic.

4.14 QurSim: Quran verse similarity

QurSim (Sharaf and Atwell 2012b) is another layer of linguistic annotation on the original Quranic text, where semantically similar or related verses are linked together. This corpus is a freely downloadable resource for corpus linguists investigating similarity, relatedness and paraphrasing in short texts. In our QurSim related-verse dataset, we relied on the Tafsir or Quranic commentary work of Ibn Kathir, a highly-reputed Quranic scholar: his commentary on each verse pointed out the related verses, so we text-mined the Tafsir to extract cross-references, producing over 7,600 pairs of related verses. The QurSim dataset is incorporated into a website TextMiningTheQuran.com where users can visualize for a given verse a network of all directly and indirectly related verses. Experiments showed that only 33% of related verse pairs shared word roots, indicating that “relatedness” goes beyond lexical matching, and involves semantics and domain knowledge. QurSim can be used for extraction and visualization of topics in the Quran (Panju 2014). Ibn Kathir’s commentary is on the Classical Arabic source text of the Quran, but the verse-relations can also apply to translations: two verses should be “related” in any language. Hence, QurSim is potentially a corpus resource for research on textual similarity and relatedness in any language that has a Quran translation.

4.15 Qurany: Quran annotated with verse topics.

Qurany (Abbas 2009, Abbas and Atwell 2013) is a bilingual (English/Arabic) web-based search tool for the Quran that enhances recall and precision when searching for concepts. This is achieved by a combination of corpus annotations. Each verse in the Quran is annotated with semantic conceptual information, extracted from Mushaf Al Tajweed, a respected Quran commentary which includes an index of nearly 1100 concepts or topics. The Mushaf Al Tajweed index, showing the verses each concept or topic appears in, was transformed into an ontology tree data-structure; on the Qurany website, users can navigate the ontology tree to find their chosen concept, then follow the link to a list of verses tagged with this concept. Each Arabic verse is also annotated with 8 alternative English translations from popular published sources; thus, a verse can be found via English keyword-search if any of the translations contains the keyword(s). Also, the user can opt to see WordNet synonyms of keywords, to broaden the search-terms and hence improve recall. The Qurany dataset is also accessible and downloadable as a website of separate HTML files, one per Quran verse, including Arabic source, 8 English translations, and list of Mushaf Al Tajweed concepts relating to the verse. This HTML format is compatible with the standard Google search interface if you append the site: operator. For example, a Google search for “sex site:http://www.comp.leeds.ac.uk/nora/html” finds all Quran verses containing the word “sex” in at least one of the English translations or Mushaf Al Tajweed concepts; and this in turn allows you to see a range of alternative English translations for these verses, along with the Arabic source text.

4.16 KSUCCA King Saud University Corpus of Classical Arabic

Lexical patterns in the Quran can be studied using an Arabic-friendly concordance program such as aConCorde. However, linguists and lexicographers generally need much larger corpora to study collocations and concordance patterns: the British National Corpus, initially developed for British English dictionary research, is 100 million words, while the Quran comprises only about 50,000 words (depending on how word-boundaries are counted). For research on distributional lexical semantics, we need a reasonably large number of examples of each word or collocation to be studied; but many words and phrases in the Quran occur only a handful of times. So, we collaborated with researchers at King Saud University to collect a 50-million word corpus of Classical Arabic texts from the same period as the Quran, the King Saud University Corpus of Classical Arabic (Alrabiah et al 2013, 2014a,b). This allows us to select a word from the Quran and then find many more examples of its use in context in Classical Arabic. The KSUCCA corpus is downloadable from the KSUCCA website, and also searchable online via SketchEngine (Kilgarriff et al 2014). KSUCCA is a key to corpus-based study of Arabic historical linguistics (Alrabiah et al 2014a), and distributional lexical semantics of the language of the Quran (Alrabiah et al 2014b).

4.18 Ontologies and semantic tagging for the Quran

We have added linguistic tagging to the Quran at several levels: Part-of-Speech tags, morphology, anaphoric references, phonetic transcription, prosodic phrase boundaries, syntactic phrase structure and dependency structure. We also have several types of annotation representing the “knowledge” in the Quran: ontology of nominal entities referred to by personal pronouns; verse topics and verse similarities; translations into English at word and verse levels. We aim to combine or unify these linguistic annotations and ontologies for the Quran (Abbas et al 2013, Abbas and Atwell 2013), to produce a knowledge-representation formalism for semantic tagging of the Quran (Sharaf and Atwell 2009, Alrehaili and Atwell 2013, 2014).

5. Conclusions and ideas for further research

Natural Language Processing is a subfield of Computer Science and Artificial Intelligence which uses corpora for computational modeling; and this computational focus is not always attractive to linguists: ‘NLP / computational linguistics has come into the field like a schoolyard bully, forcing everything that's not computational into submission, collusion or the margins.’ (Kilgarriff 2007). However, we believe Arabic Natural Language Processing can be harnessed to produce useful resources for Arabic Corpus Linguistics. Researchers at Leeds University have developed a series of online datasets and software to use in Arabic corpus linguistics research. Our resources are open-source and accessible via the web, rather than commercial; we hope this will help make them widely re-used, compared to resources kept ‘in-house’ by other Arabic computational linguistics research groups. Our resources include various types of Modern Arabic: the Corpus of Contemporary Arabic, the Arabic Internet Corpus, the World Wide Arabic Corpus, the Arabic Discourse Treebank, the Arabic Learner Corpus; and also Classical Arabic of the Quran, and the King Saud University Corpus of Classical Arabic. We have tackled various levels of linguistic analysis of Arabic corpora, including: Part-of-Speech tags, morphology, anaphoric references, named entities, learner errors and error-types, phonetic transcription, prosodic phrase boundaries, syntactic phrase structure, dependency structure, discourse connectives and relations, parallel English translations, topics, distributional lexical semantics. We have developed a range of software tools for Arabic corpus linguistics research, including tagging and annotation tools for these layers of analysis, and also Arabic corpus exploration via search, concordance, chatbot, and visualization tools.

A challenging area for further research is how to tag and represent semantics and ‘knowledge’ in Arabic texts, and in particular the Quran and other religious texts. Atwell et al (2010) proposed ‘Understanding the Quran’ as a Grand Challenge for Computer Science and Artificial Intelligence research, with twin objectives: to represent Quran “knowledge” in an Artificial Intelligence formalism, as an extra layer of semantic tagging in the Quran corpus; and to use computing and AI to help Muslims and non-Muslims to better understand the lessons of the Quran.

Understanding Islam is a major societal issue. Western schools, universities and the general public need an objective, impartial online Quran Knowledge-based Corpus to learn about Islam. Non-Arabic-speaking Muslims also want to better understand the deeper meanings in the Quran, beyond oral recitation. Current systems can search for words, and even answer basic factual questions eg ‘are angels male?’; but we need a new religious knowledge representation tag-set and formalism capable of capturing complex, subtle religious knowledge encoded in the Quran.

Machine Learning research needs a Gold Standard corpus where each text is classified and marked up by experts, so the learning algorithm can learn the classification. The Quran is an excellent Gold Standard, since many expert analyses already exist, so we can use these to train Machine Learning. Quranic scholars can verify the results of Machine Learning to ensure that Knowledge Based Systems based on the Quran are logically consistent and correct. Huge worldwide interest in the Quran means we can harness volunteers for ‘crowd-sourcing’ analysis, following the approach used successfully in the Quranic Arabic Corpus: initial automatic analysis, then proofreading and correction by many volunteers.

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, and for online collaboration. Understanding the Quran is a grand challenge for Arabic corpus linguistics.

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.

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.

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.

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, 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.

Alqurneh, A, A Mustapha, M Murad, and N Sharef. Forthcoming. ‘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.

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

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, 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, Riyadh.

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.

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.

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.

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

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

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. 2012. ‘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. 2012. ‘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. 2012. ‘Visualisation of Prosody in English and Arabic Speech Corpora.’ Proceedinds of AVML’2012 Advances in Visual Methods for Linguistics.

Brierley, C, M Sawalha, B Heselwood, and E Atwell. forthcoming. ‘A verified Arabic-IPA mapping for Arabic transcription technology, informed by Quranic recitation, traditional Arabic linguistics, and modern phonetics.’ Journal of Semitic Studies.

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.

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

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. Forthcoming. Creating language resources for under-resourced languages: methodologies, and experiments with Arabic. Language Resources and Evaluation journal.

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.

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.

Google. 2014. Definition of ‘Artificial Intelligence’.

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

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

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.

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.

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.

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.

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.

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.

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

Sawalha, M, C Brierley, and E Atwell. 2012b. ’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. 2014. ‘Automatically generated, phonemic Arabic-IPA pronunciation tiers for the boundary annotated Qur’an dataset for machine learning.’ Proceedings of LRE-Rel’2: 2nd Workshop on Language Resource and Evaluation for Religious Texts.

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

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.

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’

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.

Zolfagharifard, E. 2009. ‘Anti-terror technology tool uses human logic’. The Engineer, 23/11/09.