

## Text Analytics and Transcription Technology for Quranic Arabic

Majdi Sawalha<sup>a</sup>, Claire Brierley<sup>b</sup>, Eric Atwell<sup>b</sup>, James Dickins<sup>c</sup>

<sup>a</sup>KASIT, Computer Information Systems, University of Jordan, Amman, Jordan

<sup>b</sup>School of Computing, University of Leeds, Leeds, UK

<sup>c</sup>AMES, School of Modern Languages and Cultures, University of Leeds, Leeds, UK

sawalha.majdi@gmail.com, C.Brierley@leeds.ac.uk, E.S.Atwell@leeds.ac.uk, J.Dickins@leeds.ac.uk

### Abstract

Natural Language Processing Working Together with Arabic and Islamic Studies is a 2-year project funded by the UK Engineering and Physical Sciences Research Council (EPSRC) to study prosodic-syntactic mark-up in the Quran (Atwell et al 2013). Tajwīd or correct Quranic recitation is very important in Islam. The original insight informing this project is to view tajwīd mark-up in the Quran as additional text-based data for computational analysis. This mark-up is already incorporated into Quranic Arabic script, and identifies phrase boundaries of different strengths, plus lengthened syllables denoting prosodically and semantically salient words. We have developed a grapheme-phoneme mapping scheme (Brierley et al 2016), plus state-of-the-art software (Sawalha et al 2014) for generating a stressed and syllabified phonemic transcription or citation form for each word in the entire text of the Quran, using the International Phonetic Alphabet (IPA). This canonical pronunciation tier for Classical Arabic is informed and evaluated by Arabic linguists, tajwīd scholars, and phoneticians, and published in an open-source Boundary-Annotated Quran corpus and machine learning dataset (ibid). We utilise statistical techniques such as keyword extraction to explore semiotic relationships between sound and meaning in the Quran, invoking a Saussurean-type view of the sign as ‘...a bi-unity of expression and content...’ (Dickins 2007). Our investigation entails: (i) text data mining for statistically significant phonemes, syllables, words, and correlates of rhythmic juncture; and (ii) interpretation of results from interdisciplinary perspectives: Corpus Linguistics; tajwīd science; Arabic Linguistics; and Phonetics and Phonology.

**Keywords:** tajwid; prosody; phonemic transcription; phrase boundary.

### 1. INTRODUCTION

We present an overview of work related to a two-year project (2013-2015) funded by the UK Engineering and Physical Sciences Research Council (EPSRC) (Atwell et al 2013). The project is entitled: Natural Language Processing Working Together with Arabic and Islamic Studies, and involves a core interdisciplinary and international research team from Computing and Arabic at the Universities of Leeds and Jordan, developing Islamic applications for the Arabic Quran as core text. The original insight informing this project is to view tajwīd or recitation mark-up in the Quran as additional text-based data for computational analysis. This mark-up is an integral part of the Quranic Arabic script, and identifies phrase boundaries of different strengths known as waqf, plus colour-coded

highlighting of various recitative effects denoting prosodically and semantically salient words.

## 2. The Boundary-Annotated Quran Corpus (version 1.0)

We are interested in phrasing strategies in Arabic and other languages, notably English, and to pursue this interest we use corpora annotated with phrase boundaries and other linguistic information. Widely-used speech corpora for British and American English are the Lancaster/IBM Spoken English Corpus or SEC (Taylor and Knowles 1988), and the Boston University Radio News Corpus (Ostendorf *et al* 1995) respectively. The British system for prosodic annotation uses a tripartite boundary annotation scheme of {major, minor, none}, while the American *Tones and Break Indices* (ToBI) annotation scheme (Beckman and Hirschberg 1994) identifies five levels of juncture between words: {0, 1, 2, 3, 4}. Such corpora then permit corpus-based studies on the linguistic correlates of phrase juncture (Brierley and Atwell 2011), and also machine learning experiments to predict phrase boundaries in unseen text.

The Arabic Quran, with fine-grained, *tajwīd* mark-up of *waqf*, is hitherto the only boundary-annotated speech corpus for Arabic. We have previously reported on our *Boundary-Annotated Quran* corpus and machine learning dataset (Brierley *et al* 2012) with multiple linguistic annotation tiers. In this corpus, we map the prosodic annotation scheme from *tajwīd* into the {major, minor, none} categories of the British system as these are more manageable for phrase break prediction. Version 1.0 of our corpus (*ibid*) aligns Arabic words in the Quran (in both the Othmani and Modern Standard Arabic script) to two different, coarse-grained levels of syntactic and prosodic categorisation (Table 1), and also identifies sentence terminals, namely: words immediately preceding compulsory and recommended stops. The alternative syntactic categories are noun, verb, or particle {N, V, P}, and a slightly expanded scheme of 10 labels after Dukes and Habash (2010). Alternative boundary types are a binary scheme of {break, nonbreak} and the tripartite scheme of {major, minor, none}.

Table 1: Aligned tiers in version 1.0 of the *Boundary-Annotated Quran* dataset

Othmani	MSA	Syntax: NVP	Syntax: 10 PoS	Verse ends	Boundaries: tripartite	Boundaries: binary	Sentences
بِسْمِ	بِسْمِ	N	NOUN	-	-	non-break	-
اللَّهِ	اللَّهِ	N	NOUN	-	-	non-break	-
الرَّحْمَنِ	الرَّحْمَنِ	N	NOMINAL	-	-	non-break	-
الرَّحِيمِ	الرَّحِيمِ	N	NOMINAL	⊗		break	terminal
الْحَمْدُ	الْحَمْدُ	N	NOUN	-	-	non-break	-
لِلَّهِ	لِلَّهِ	N	NOUN	-	-	non-break	-
رَبِّ	رَبِّ	N	NOUN	-	-	non-break	-
الْعَالَمِينَ	الْعَالَمِينَ	N	NOUN	⊗		break	terminal

### 3. Arabic Phrase Break Prediction

We are pioneering phrase break prediction for Arabic. In Sawalha *et al* (2012a, 2012b) we use trigram and HMM taggers from the *Natural Language Toolkit* (Bird *et al* 2009) to predict boundaries in a discrete Quran test set of 7318 words and 849 sentences, using both sets of syntactic features and break types in the *Boundary-Annotated Quran*. This test set comprises Quran chapters where Meccan/ Medinan provenance is disputed, and constitutes a fair test for a classifier trained on both styles. These preliminary experiments were conducted with respect to a challenging majority class baseline of 85.56% accuracy, and our most promising result is summarised in Figure 2. Our best score of 88.47% accuracy was achieved by a trigram tagger for binary phrase break classification (*i.e.* break, nonbreak), with words tagged simply as noun, verb or particle {N, V, P}. An important aspect of this work is evaluation via a range of metrics as recommended in Brierley (2011). Hence, though the HMM tagger performs below par in terms of accuracy (82.63%), it makes significant gains on both the baseline and the trigram tagger in terms of Balanced Classification Rate (BCR), namely: the average of correct predictions over each class (Table 2). Accuracy and BCR scores for predicting phrase breaks in Quranic Arabic were calculated from raw predictions: true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs).

Table 2: Trigram and HMM tagger scores for predicting phrase breaks in Quranic Arabic in terms of two evaluation metrics: accuracy and BCR

Tagger	Syntax: NVP	Boundaries: binary	Accuracy	BCR	TPs	TNs	FPs	FNs
Baseline	3	2	85.56%	0.50	0	6261	0	1057
Trigram	3	2	88.47%	0.67	380	6094	167	677
HMM	3	2	82.63%	0.72	601	5446	815	456

### 4. Arabic > IPA Transcription Technology

One of our objectives in the *Working Together* project is automated transcription of Arabic using the International Phonetic Alphabet (IPA). This alphabet was chosen in preference to one of the many romanization schemes for Arabic because it is an international standard. The output of our algorithm is a phonemic pronunciation and citation form for each Arabic word, similar to entries in the widely-used Oxford and Longman learner dictionaries for English. Thus our new transcription technology for Arabic targets Arabic dictionaries and Arabic language learning, plus Natural Language Engineering applications for Arabic involving speech recognition and speech synthesis. A full account of our Arabic > IPA mapping scheme, including its linguistic underpinning in Quranic *tajwīd*, medieval Arabic linguistics, and modern phonetics, is given in Brierley *et al* (2016).

Although Arabic spelling is largely a phonemic system, with one-to-one letter to sound correspondence, certain exceptions make automated Arabic > IPA transcription a non-trivial task. The principal innovation in our approach, as compared to existing schemes like the Speech Assessment Methods Phonetic Alphabet (SAMPA) for Arabic (Wells 2002), lies in its treatment of certain character sequences as compounds requiring transcription. This has the practical benefit of reducing the number of rules to resolve difficult cases. One such problematic and frequently occurring case is assimilation of /l/ in the definite article before coronal sounds or so-called “sun-letters”, such that السَّمَاءُ can be correctly transcribed in IPA as /ʔassama:ʔu/.

Our mapping algorithm has a 58-entry dictionary of mapped Arabic > IPA pairings which anticipate and document grapheme-phoneme transitions for: consonants; long and short vowels; diacritic marks; and compound character sequences such as the trigram (VCV) pattern for the Arabic diphthong اِيّ transcribed in IPA as /aj/. It is implemented in two stages: a pre-processing stage, and a rule-development stage; and the tokenization module of the SALMA tagger (Sawalha 2011) is used throughout to tokenize and preprocess input Arabic text. In the pre-processing stage, Arabic word letters are mapped to their IPA character equivalent on a *one-to-one* basis. The rule development stage then applies *phonetic rules* that modify the interim IPA string to produce a correct IPA transcription for the Arabic input word. Details of our mapping algorithm appear in Sawalha *et al* (2014), but a summary of main processing steps is given here in Table 3.

Table 3: Summary of main processing steps in our Arabic > IPA transcription tools

PRE-PROCESSING STAGE	RULE DEVELOPMENT STAGE
<ul style="list-style-type: none"> <li>• mapping Arabic consonant letters into one IPA alphabet such as ( ب، ت، ث، ...) &gt; (/b/, /t/, /θ/), or into two IPA alphabets such as ( ... ط، ض، ظ ) &gt; (/sʕ/, /dʕ/, /tʕ/)</li> <li>• long vowels such as ( ا، و، ي ) &gt; (/a:/, /u:/, /i:/), and short vowels such as ( َ ، ُ ، ِ ) &gt; (/a/, /u/, /i/)</li> <li>• <i>hamza<sup>h</sup></i> ( ء، أ، و، ئ ) , regardless of form or shape, is represented by the IPA character (ʔ)</li> <li>• <i>tanwīn</i> are defined such that ( ً ، ٌ ، ٍ ) &gt; (/an/, /un/, /in/)</li> <li>• <i>sukūn</i> is not mapped to any IPA character</li> </ul>	<p>Major challenges for the one-to-one mapping step are dealing with:</p> <ul style="list-style-type: none"> <li>• the definite article (<i>i.e.</i> whether the /l/ is pronounced or assimilated to the following coronal or “sun-letter” sound);</li> <li>• long vowels when they are pronounced as vowels;</li> <li>• <i>ʿalif</i> of the group ( أَلِف التَّفْرِيق ) which is not pronounced;</li> <li>• words with special pronunciations;</li> <li>• <i>hamza<sup>tu</sup> al-waṣl</i>;</li> <li>• <i>tanwīn</i>.</li> </ul>

## 5. The Boundary-Annotated Quran Corpus (version 2.0)

A further research objective in *Working Together* is stylistic analysis of the Quran, focusing in particular on the semiotics of sound, invoking a Saussurean-type view of the linguistic sign as ‘...a bi-unity of expression and content...’ (Dickins 2007). Hence, another motivation for our Arabic > IPA transcription technology was to generate a phonemic representation of the entire text of the Quran as a more reliable basis for obtaining frequency distributions for individual letters/sounds. This new phonemic representation of the text of the Quran is published in version 2.0 of our *Boundary-Annotated Quran* dataset (Sawalha *et al* 2014), which features Arabic words tagged with two alternative pausal, phonemic transcriptions in IPA (one *with* and one *without* short vowel case endings), plus a Buckwalter-style transliteration (reference), and an Arabic root where applicable. A snapshot of some of the aligned tiers in this dataset is given in Table 4.

Figure 4: Aligned tiers in version 2.0 of the *Boundary-Annotated Quran* dataset

Chapter & verse	Oth.	MSA			IPA (with case endings)	IPA (without case endings)	Buckwalter	Root
78 1	عَمَّ	عَمَّ	P	-	ʕamma	ʕamma	Eam~a	
78 1	يَتَسَاءَلُونَ	يَتَسَاءَلُونَ	V		ʃatasa:ʔalu:na	ʃatasa:ʔalu:n	yatasaA'aluwna	سأل
78 2	عَنِ	عَنِ	P	-	ʕani	ʕani	Eani	
78 2	النَّبَاِ	النَّبَاِ	N	-	ʔannabaʔi	nnabaʔi	Aln~aba<i	نبا
78 2	الْعَظِيْمِ	الْعَظِيْمِ	N		ʔalʕað'ijmi	lʕað'ijm	AloEaZiymi	عظم
78 3	الَّذِي	الَّذِي	N	-	ʔallaði:	ʔallaði:	Al~a*iy	
78 3	هُمْ	هُمْ	N	-	hum	Hum	humo	
78 3	فِيهِ	فِيهِ	P	-	fi:hi	fi:hi	fiyhi	
78 3	مُخْتَلِفُونَ	مُخْتَلِفُونَ	N		muxtalifu:na	muxtalifu:n	muxotalifuwna	خلف
78 4	كَلَّا	كَلَّا	P	-	kalla:	kalla:	kal~aA	
78 4	سَيَعْلَمُونَ	سَيَعْلَمُونَ	V		sajaʕlamu:na	sajaʕlamu:n	sayaEolamuwna	علم


## 6. Regular Expressions and Keyword Extraction

The rules of *tajwīd* recitation can be expressed algorithmically. In a recent paper, we have presented software for locating all *qalqalah* sites in the text of the *Quran* using our *Boundary-Annotated Quran* dataset (Brierley *et al* 2014). *Qalalah* is an emphatic articulation applied to a subset of Arabic letters { ق ط د ج ب } when they occur in weak prosodic positions. Our software is based on regular expression search patterns. These can be used to precisely define the contexts in which *qalqalah* (and other *tajwīd* effects) occur. Figure 5 gives our regular expression (RE) for capturing word-internal *qalqalah* events. It also

decomposes the RE to show ordering of constraints, and displays the desired result applied to *Quran* 85.12.

Words with *qalqalah* are not only frequent, but are also found to be statistically significant. Determining words of unusual frequency (or infrequency) in a test set relative to a suitable reference set is a statistical technique widely used in Corpus Linguistics; it is known as Keyword Extraction (KWE). We have previously used this technique, as implemented in the *Semantic Pathways* Visual Analytics toolkit, for exploring distinguishing features in British versus American English (Brierley *et al* 2013). In Brierley *et al* (2014), we report on *Semantic Pathways* for extracting keywords in the Arabic *Quran*. One such keyword is رَبِّ *rabbi lord*. It emerges as positively “key” in the Meccan versus Medinan corpus comparison. Since this word also implies *qalqalah* in its pausal form, we speculate that *qalqalah* is a *semantic* as well as a prosodic salience marker.

Table 5: Our algorithm captures the *qalqalah* site underlined in *Quran* 85.12

RE for word-internal <i>qalqalah</i> events		
u"[\u0621-\u0652]* [\u0642,\u0637,\u0628,\u062C,\u062F] \u0652 [\u0621-\u0652]+"		
1	zero or more occurrences of any Arabic letter/character	u"[\u0621-\u0652]*
2	one of the <i>qalqalah</i> set	[\u0642,\u0637,\u0628,\u062C,\u062F]
3	sukūn	\u0652
4	at least one Arabic letter/character	[\u0621-\u0652]+"
Quran 85.12		
	<p>Indeed the grip of your Lord is surely strong</p>	

## 7. Conclusions and Further Work

We have presented an overview of work undertaken so far in an international research project involving Natural Language Processing, Arabic, and Islamic Studies as interdisciplinary streams (Atwell *et al* 2013). Principal outcomes in Year 1 of the project have been the development of Arabic > IPA transcription technology, and the publication of version 2.0 of our *Boundary-Annotated Quran* dataset (Sawalha *et al* 2014). Work in progress explores the semantic function of *madd*, another *tajwīd* effect governing vowels. We will also extend our IPA transcription of the *Quran* to capture coarticulation and full *tajwīd* rules.

## 8. References

Atwell, E.S., Dickins, J. and Brierley, C. 2013. Natural Language Processing Working Together with Arabic and Islamic Studies. Engineering and Physical Sciences Research Council (EPSRC). EP/K015206/1. Online. Accessed: 29.06.2014. <http://gow.epsrc.ac.uk/NGBOViewGrant.aspx?GrantRef=EP/K015206/1>

- Beckman, M. and Hirschberg, J. 1994. The ToBI annotation conventions. The Ohio State University and AT&T Bell Laboratories, unpublished manuscript. Online. Accessed September 2011. <ftp://ftp.ling.ohio-state.edu/pub/phonetics/TOBI/ToBI/ToBI.6.html>.
- Bird, S., Klein, E. and Loper, E. 2009. *Natural Language Processing with Python*. Sebastopol, CA. O'Reilly Media, Inc.
- Brierley, C. 2011. *Prosody Resources and Symbolic Prosodic Features for Automated Phrase Break Prediction*. PhD Thesis. School of Computing. University of Leeds.
- Brierley, C. and Atwell, E. 2011. "Non-Traditional Prosodic Features for Automated Phrase-Break Prediction." In *Journal of Literary and Linguistic Computing*. (Digital Humanities 2010 Special Issue), doi: 10.1093/ljc/fqr023
- Brierley, C., Sawalha, M. and Atwell, E. 2012. 'Open-Source Boundary-Annotated Corpus for Arabic Speech and Language Processing'. In *Proceedings of the Language Resources and Evaluation Conference (LREC) 2012*. Istanbul, Turkey. Online. Accessed: 29.06.2014. <http://www.lrec-conf.org/proceedings/lrec2012/index.html>
- Brierley, C., Atwell, E., Rowland, C. and Anderson, J. 2013. *Semantic Pathways: a novel visualization of varieties of English*. In *Journal of International Computer Archive of Modern and Medieval English (ICAME)*. Volume 37.
- Brierley, C., Sawalha, M. and Atwell, E. 2014. *Tools for Arabic Natural Language Processing: a case study in qalqalah prosody*. To appear in *Proc. Language Resources and Evaluation Conference (LREC 2014)*, Reykjavik
- Brierley, C., Sawalha, M., Heselwood, B. and Atwell, E. (2016). *A Verified Arabic-IPA Mapping for Arabic Transcription Technology, Informed by Quranic Recitation, Traditional Arabic Linguistics, and Modern Phonetics*. *Journal of Semitic Studies* (1), 157-186.
- Dickins, J. 2007. *Sudanese Arabic: phonematics and syllable structure*. Wiesbaden: Otto Harrassowitz Verlag.
- Dukes, K. and Habash, N. 2010. 'Morphological Annotation of Qur'anic Arabic.' In *Proceedings of Language Resources and Evaluation Conference (LREC 2010)*, Valletta, Malta.
- Ostendorf, M., Price, P. and Shattuck-Hufnagel, S. 1996. *Boston University Radio Speech Corpus*. Philadelphia. Linguistic Data Consortium.
- Sawalha, Majdi. 2011. *Open-source Resources and Standards for Arabic Word Structure Analysis*. Leeds: University of Leeds PhD.
- Sawalha, M., Brierley, C., and Atwell, E. 2012a. 'Predicting Phrase Breaks in Classical and Modern Standard Arabic Text.' In *Proceedings of LREC 2012: Language Resources and Evaluation Conference*. Istanbul, Turkey. May 2012.
- Sawalha, M., Brierley, C., and Atwell, E. 2012b. "Open-Source Boundary-Annotated Qur'an Corpus for Arabic and Phrase Breaks Prediction in Classical and Modern Standard Arabic Text." In *Journal of Speech Sciences*, 2.2.
- Sawalha, M., Brierley, C. and Atwell, E. 2014. *Automatically generated, phonemic Arabic-IPA pronunciation tiers for the Boundary Annotated Qur'an Dataset for Machine Learning (version 2.0)*'. In *Proceedings of the Workshop on Language Resources and Evaluation for Religious Texts (LRE-Rel2) at LREC 2014*. Reykjavik, Iceland.
- Taylor, L.J. and Knowles, G. 1988. 'Manual of Information to Accompany the SEC Corpus: The machine readable corpus of spoken English.' Accessed: January 2010.
- Wells, J.C. 2002. *SAMPA for Arabic*. Online. Accessed: 25.04.2013. <http://www.phon.ucl.ac.uk/home/sampa/arabic.htm>