Automatically generated, phonemic Arabic-IPA pronunciation tiers for the Boundary Annotated Qur'an Dataset for Machine Learning (version 2.0)

Conference Paper · January 2014

DOI: 10.13140/2.1.2887.2640

CITATIONS
12

READS
1,344

3 authors, including:

Eric Atwell
University of Leeds
407 PUBLICATIONS **4,081** CITATIONS

SEE PROFILE

Majdi Sawalha
University of Jordan
51 PUBLICATIONS **524** CITATIONS

SEE PROFILE

Proceedings Paper:
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 LRE-Rel 2: 2nd Workshop on Language Resource and Evaluation for Religious Texts, LREC 2014 post-conference workshop 31st May 2014, Reykjavik, Iceland. LRE-Rel 2: 2nd Workshop on Language Resource and Evaluation for Religious Texts, LREC 2014 post-conference workshop, 31st May 2014, Harpa Conference Center, Reykjavik, Iceland. The University of Leeds , 42 - 47.

Automatically generated, phonemic Arabic-IPA pronunciation tiers for the *Boundary Annotated Qur'an Dataset for Machine Learning* (version 2.0)

Majdi Sawalha[1,2], Claire Brierley[2], Eric Atwell[2]
University of Jordan[1] and University of Leeds[2]
[1] Computer Information Systems Dept., King Abdullah II School for IT, University of Jordan, Amman 11942, Jordan
[2] School of Computing, University of Leeds, LS2 9JT, UK
E-mail: sawalha.majdi@gmail.com, C.Brierley@leeds.ac.uk, E.S.Atwell@leeds.ac.uk

Abstract

In this paper, we augment the *Boundary Annotated Qur'an* dataset published at LREC 2012 (Brierley *et al* 2012; Sawalha *et al* 2012a) with automatically generated phonemic transcriptions of Arabic words. We have developed and evaluated a comprehensive grapheme-phoneme mapping from Standard Arabic > IPA (Brierley *et al* under review), and implemented the mapping in Arabic transcription technology which achieves 100% accuracy as measured against two gold standards: one for Qur'anic or Classical Arabic, and one for Modern Standard Arabic (Sawalha *et al* [1]). Our mapping algorithm has also been used to generate a pronunciation guide for a subset of Qur'anic words with heightened prosody (Brierley *et al* 2014). This is funded research under the EPSRC "*Working Together*" theme.

Keywords: IPA phonemic transcription; SALMA Tagger; Arabic transcription technology

1. Introduction: the Boundary Annotated Qur'an dataset for machine learning

For LREC 2012, we reported on a Qur'an dataset for Arabic speech and language processing, with multiple annotation tiers stored in tab-separated format for speedy text extraction and ease of use (Brierley *et al* 2012). One novelty of this dataset is Arabic words mapped to a prosodic boundary scheme derived from traditional *tajwīd* (recitation) mark-up in the Qur'an as well as syntactic categories. Thus we used our dataset for experiments in Arabic phrase break prediction: a classification task that simulates human chunking strategies by assigning prosodic-syntactic boundaries (phrase breaks) to unseen text (Sawalha *et al* 2012a; 2012b). In this paper, we report on version 2.0 of this dataset with 4 new prosodic and linguistic annotation tiers. It features novel, fully automated transcriptions of each Arabic word using the International Phonetic Alphabet (IPA), with an IPA > Arabic mapping scheme based on Quranic recitation, traditional Arabic linguistics, and modern phonetics (Brierley *et al* under rveiew).

2. The *Boundary Annotated Qur'an* for Arabic phrase break prediction

Phrase break prediction is a classification task in supervised machine learning, where the classifier is trained on a substantive sample of "gold-standard" boundary-annotated text, and tested on a smaller, unseen sample from the same source *minus* the boundary annotations. The task equates to classifying words, or the junctures (*i.e.* whitespaces) between them, via a finite set of category labels: in most cases, a binary set of `breaks` versus `non-breaks`; and less commonly, a tripartite set of `{major; minor; none}`.

2.1 Boundary annotations

A boundary-annotated and part-of-speech tagged corpus is a prerequisite for developing phrase break classifiers. One novelty of our dataset is that we derived a coarse-grained, prosodic-syntactic boundary annotation scheme for Arabic from traditional recitation mark-up, known as *tajwīd*. Tajwīd boundary annotations are very fine-grained, delineating eight different boundary types, namely: three major boundary types, four minor boundary types, and one *prohibited* stop. For our initial experimental purposes, we collapsed these eight degrees of boundary strength into the familiar `{major, minor, none}` sets of British and American English speech corpora (Taylor and Knowles 1988; Beckman and Hirschberg 1994). Another novelty is that we used certain stop types (*i.e.* compulsory, recommended, and prohibited stops) to segment the text into 8230 sentences.

2.2 Syntactic annotations: the efficacy of traditional Arabic syntactic category labels

Phrase break prediction assumes part-of-speech (PoS) tagged input text as well as prior sentence segmentation, since syntax and punctuation are traditionally used as classificatory features. Traditional Arabic grammar (Wright, 1996; Ryding, 2005; Al-Ghalayyni, 2005) classifies words in terms of just three syntactic categories {`nouns; verbs; particles`}, and another novelty of our dataset is that we retained this traditional feature set as the default. Subsequently, we added a second syntactic annotation tier differentiating a limited selection of ten subcategories extracted from fully parsed sections of an early version of the *Quranic Arabic Corpus* (Dukes 2010). These comprise `{nouns; pronouns; nominals; adverbs; verbs; prepositions; ʻlām prefixes; conjunctions; particles; disconnected letters}`. However, our preliminary experiments with a trigram tagger for Arabic phrase break prediction report a significant improvement of 88.69% in respect of baseline

accuracy (85.56%), using the traditional, tripartite, syntactic feature set (Sawalha *et al* 2012a; Sawalha *et al* 2012b). A sample from version 1.0 of the multi-tiered *Boundary Annotated Qur'an* dataset is shown in Fig 1. The 8 columns of Fig 1 are: (1) Arabic words in Othmani script, (2) Arabic words in MSA script, (3) three POS tags of the word, (4) ten POS tags of the word, (5) verse ending symbol, (6) tripartite boundary annotation tag, (7) binary boundary annotation tag, and (8) sentence terminal.

OTH	MSA	Syntax: NVP	Syntax: 10 PoS	Verse ends	Boundaries: tripartite	Boundaries: binary	Sentences
بِسْمِ	بِسْمِ	N	NOUN	-	-	non-break	-
ٱللَّهِ	اللَّهِ	N	NOUN	-	-	non-break	-
ٱلرَّحْمَٰنِ	الرَّحْمَٰنِ	N	NOMINAL	-	-	non-break	-
ٱلرَّحِيمِ	الرَّحِيمِ	N	NOMINAL	۞	‖	break	terminal
ٱلْحَمْدُ	الْحَمْدُ	N	NOUN	-	-	non-break	-
لِلَّهِ	لِلَّهِ	N	NOUN	-	-	non-break	-
رَبِّ	رَبِّ	N	NOUN	-	-	non-break	-
ٱلْعَٰلَمِينَ	الْعَالَمِينَ	N	NOUN	۞	‖	break	terminal

Figure 1: Sample tiers from the *Boundary-Annotated Qur'an* dataset version 1.0

3. IPA transcription tiers for the *Boundary Annotated Qur'an* (version 2.0): rationale

One of our objectives in the EPSRC-funded *Working Together* project is to automate Arabic transcription using a carefully defined subset of the International Phonetic Alphabet (IPA). This task matches the definition of *transcription*, as opposed to *transliteration* and *romanisation* for Arabic, as described in Habash *et al* (2007). The output of our algorithm is a phonemic pronunciation for each Arabic word as an element of its citation form, similar to entries in the OALD [1] and LDOCE[2] for English, to enhance Arabic dictionaries, to facilitate Arabic language learning, and for Arabic natural language engineering applications involving speech recognition and speech synthesis.

A corpus of fully vowelised Arabic text was essential for developing and evaluating both the comprehensive grapheme-phoneme mapping from Standard Arabic > IPA, and the mapping algorithm itself. The Qur'an is an iconic text and an excellent gold standard for modeling and evaluating Arabic NLP, since it arguably subsumes other forms of Arabic, including regional dialects (Harrag and Mohamadi 2007), and MSA (Sharaf 2012). Hence we have developed our mapping and our mapping algorithm on the *Boundary-Annotated Qur'an* dataset, which includes the entire text of the *Qur'an* in fully vowelised MSA as well as traditional *Othmani* script.

A further research objective in *Working Together* is stylistic and stylometric analysis of the *Qur'an*, and a phonemic representation of the entire text via our MSA > IPA mapping will facilitate this. Version 2.0 of the *Boundary Annotated Qur'an* dataset has emerged as a result, featuring 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 and an Arabic root where applicable.

[1] Oxford Advanced Learner's Dictionary
[2] Longman Dictionary of Contemporary English

4. The Arabic > IPA mapping: linguistic underpinning

In general, Arabic spelling is a phonemic system with one-to-one letter to sound correspondence. Nevertheless, our mapping scheme is original due to its treatment of certain character sequences as *compounds* requiring transcription. This differentiates our scheme from the machine readable Speech Assessment Methods Phonetic Alphabet or SAMPA for Arabic (Wells 2002), where many more hand-crafted rules would need to be developed before implementing automatic Arabic > SAMPA transcription due to the sparseness of the scheme itself. Therefore, as well as the usual transcription of consonants, long and short vowels, and diacritic marks, we have compiled a dictionary of mapped MSA > IPA pairings that both anticipates and documents grapheme-phoneme relationships extending beyond a single letter to the immediate right-left context in fully vowelised Arabic text. For example, Arabic has two diphthongs which are each realised orthographically via the trigram sequence VCV, where V represents a short vowel or other diacritic mark and C is a consonant or semi-vowel (Fig. 2).

Arabic	Example	N-gram capture	IPA
◌َيْ	بَيْت	trigram: VCV	/aj/
◌َوْ	حَوْل	trigram: VCV	/aw/

Figure 2: Two diphthongs represented by a trigram character sequence

Our mapping is also original in that it draws on traditional Arabic linguistics for selecting the most appropriate subset of IPA symbols to represent the sound system of the language. A basic version of our Arabic > IPA mapping appears in Appendix I; the full version appears in Brierley *et al* (under review).

5. The Arabic > IPA mapping algorithm

The Arabic > IPA mapping algorithm automates phonetic transcription of Arabic words and outputs a phonemic pronunciation for each word. The algorithm has two stages: the pre-processing stage where Arabic word letters are mapped to their IPA character equivalent on a one-to-one basis; and a second stage which involves the development and application of phonetic rules that modify the IPA string produced in the first stage to produce the correct IPA transcription of the input Arabic word. The following subsections briefly describe the stages of the algorithm. The algorithm is explained in detail in (Sawalha *et al*, 2014a).

5.1 Pre-processing stage

During this project, a carefully defined subset of the International Phonetic Alphabet (IPA) for Arabic transcription was defined (Brierley *et al* under review). This includes mapping Arabic consonant letters into one IPA alphabet such as (... ب، ت، ث) > (/b/, /t/, /θ/), or into two IPA alphabets such as (... ص، ض، ط) > (/sˤ/, /dˤ/,/tˤ/). IPA alphabets for both long and short vowels are also defined, long vowels such as (ا، و، ي)> (/aː/,/uː/,/iː/), and short vowels such as (ِ ، ُ ، ◌َ)>(/a/,/u/,/i/). *hamza*^{h} (ء، أ، ؤ، ئ), regardless of form or shape, is represented by the IPA character (ʔ). IPA alphabets for *tanwīn* are defined such that (ً ، ٌ ، ٍ) > (/an/, /un/, /in/); *sukūn* is not mapped to any IPA character because *sukūn* represents silence. Using these carefully defined sets of IPA alphabets, a 58-entry dictionary was constructed to facilitate the Arabic > IPA one-to-one mapping. Appendix I shows a basic version of our Arabic > IPA mapping; the full version appears in Brierley *et al* (under review).

The tokenization module of the SALMA-Tagger (Sawalha, 2011; Sawalha and Atwell, 2010) was used to tokenize and preprocess the input Arabic text. The SALMA-Tokenizer preprocesses Arabic words by resolving gemination marked by (ّ) *šadda*^{h} into two similar letters: the first carries a *sukūn* diacritic and the second carries a short vowel similar to the short vowel of the original geminated letter. The tokenizer also replaces the prolongation letter (آ) *madd* into *hamza*^{h} followed by the long vowel *'alif*. The SALMA-Tokenizer has a spell checking and correction module which verifies the spelling of the Arabic word in terms of valid letter and diacritic combinations. It limits each letter of the processed word to only one diacritic. The output of the SALMA-Tokenizer is an Arabic word string which best suits the one-to-one mapping of Arabic letters into IPA alphabets.

As a first step in the mapping process, the one-to-one mapping module reads the processed Arabic word. For each letter it searches the dictionary for its equivalent IPA alphabet. The output is an IPA string representing the sequence of IPA alphabets equivalent to the Arabic letters and diacritics of the input word. For example the word يَتَسَاءَلُونَ *yatasā'alūna* "they are asking one another" is mapped into the IPA string /jatasaaːʔaluwna/. The accuracy of the preprocessing stage of the Arabic > IPA mapping algorithm showed that about 70% of Arabic words in the test sample were not mapped correctly. Therefore, mapping Arabic words into IPA using one-to-one mapping only is not accurate and a rendering stage of pronunciation is needed. The following subsection discusses the rule-based stage of the Arabic > IPA algorithm that renders the produced string and generates 100% accurate results.

5.2 Rule development

As shown in the previous section, a pronunciation rendering stage is needed to produce correct phonetic transcription of Arabic words. Traditional Arabic orthography includes silent letters, and ambiguous letters such as the letters (أ، و، ي) *ʾalif*, *wāw*, and *yāʾ* which can be consonants, semi-vowels or long vowels. Also, short vowels and diacritics necessary to convey the pronunciation reliably are usually absent. Some letters appear in the orthographic word but are not pronounced and some sounds are not presented in the orthographic word altogether. The major challenges for the one-to-one mapping step are: dealing with the (i) definite article (i.e. whether the *l* is pronounced or assimilated to the following sound becoming a geminate of it), (ii) long vowels when they are pronounced as vowels, (iii) *ʾalif* of the group (ألف التفريق) which is not pronounced, (iv) words with special pronunciations, (v) *hamza*^{tu} *al-waṣl* and (vi) *tanwīn*.

The second stage of the Arabic > IPA mapping algorithm is based on especially developed rules and regular expressions to deal with cases for which the one-to-one mapping fails to generate a correct phonetic transcription. Output from the previous step was studied for the purpose of finding patterns in the mistaken transcriptions. Around 50 rules were developed and ordered correctly so that algorithm could generate the correct IPA transcriptions of input Arabic words. For example, words ending in *tanwīn al-fatḥ* which were transcribed into /aaːan/ in the first stage, are rendered by the IPA string as /an/ as in the word مِهَادَاً *mihādan* "resting place" which is transcribed into /mihaːdan/. Other rules deal with the definite article (ال) when followed by a letter corresponding to coronal and non-coronal sounds[3]. If the definite article is followed by coronal sound then the IPA string /aːl/ representing the one-to-one mapping is replaced by /ʔa/ followed by a doubling of the coronal sound such as transcribing the word النَّبَإِ *an-naba'i* "news" as /ʔannabaʔi/. If it is followed by a non-coronal sound it is replaced by /ʔal/,

[3] Coronal consonants are consonants articulated with the flexible front part of the tongue. They are also known as solar or sun letters. Non-coronal consonants are known as lunar or moon letters.

such as the word الْحَقُّ *al-ḥaqqu* "the truth" transcribed in the IPA string as /ʔalħaqqu/. This process is non-trivial: Sawalha *et al* (2014a) has a detailed description of patterns of mistaken transcription and rules for correcting the output.

6. Evaluation

Evaluation of the *Boundary Annotated Qur'an* (version 2.0) focused on assessing the automatically generated Arabic > IPA transcription tiers of the BAQ corpus. The evaluation is performed using two methods: (i) measuring the accuracy of the algorithm by comparing the results against an especially constructed gold standard for evaluation; and (ii) generating a frequency list of Qur'an with automatically generated IPA transcriptions of the Qur'an word types and verifying these transcriptions by linguists specialized in tajwīd and phonology.

6.1 The Qur'an gold standard

A gold standard for evaluating the Arabic > IPA mapping algorithm was especially constructed. The gold standard consists of about 1000 words from the Qur'an, chapters 78 to 85. For each word in the gold standard, the IPA transcription was manually generated. Figure 3 shows a sample of the gold standard for evaluation. Evaluation of the output of the Arabic > IPA mapping algorithm showed an accuracy of 100%, in indication of what we had aspired for.

Chapter #	Verse #	Word #	Word	Pausal mapping - with case ending
78	1	1	عَمَّ	/ʕamma/
78	1	2	يَتَسَاءَلُونَ	/jatasaːʔaluːna/
78	2	1	عَنِ	/ʕani/
78	2	2	النَّبَإِ	/ʔannabaʔi/
78	2	3	الْعَظِيمِ	/ʔalʕaðˁijmi/

Figure 3: A sample of the Qur'an gold standard

6.2 Generation and verification of the IPA transcriptions of the Qur'an

The second method for evaluating the *Boundary Annotated Qur'an* (version 2.0) was to manually check and verify the automatically-generated IPA transcriptions of all words in the BAQ corpus. The BAQ corpus consists of 77,430 words and 17,606 word types. To reduce the time and effort of manual verification of the IPA transcription, word types were verified rather than words. A frequency list of the Qur'an was generated first, and then the IPA transcription for each word type of the BAQ corpus was verified by linguists who are specialized in both tajwīd and phonetics. This evaluation method is suitable for verifying the IPA transcriptions of the BAQ corpus words in their pausal forms while preserving case endings in the transcription. A sample of the frequency list for the first 50 word types is in Appendix II.

After manual verification of the Qur'an frequency list, we computed the accuracy of the Arabic>IPA transcription algorithm using the verified frequency list of the Qur'an. Only 91 errors in transcription were found in the Qur'an frequency list. These are 91 word types from a total of 17,606 word types in the list. Therefore, the accuracy of the automatic Arabic > IPA transcription algorithm is 99.48%. The 91 word type errors occur 347 times in the BAQ corpus which contains 77,430 words. Therefore, the computed accuracy of the automatic Arabic > IPA transcription algorithm is 99.55%.

On the other hand, contextual transcription of the BAQ corpus words is concerned with transcribing the words in context. They are transcribed so as to represent co-articulatory effects in continuous speech but with a definite pause at the end of each sentence. For example, the two sentences/verses in figure 3 " عَمَّ يَتَسَاءَلُونَ ﴿١﴾ عَنِ النَّبَإِ الْعَظِيمِ ﴿٢﴾" "About what are they asking one another? (1) About the great news - (2) " are transcribed contextually into /ʕamma jatasaːʔaluːn (1) ʕani nnabaʔi lʕaðˁijm (2)/. Verification of the contextual transcription tier of the BAQ corpus is done by checking these transcriptions sentence by sentence. The Quranic verses in the BAQ corpus are divided into sentences and pauses are defined as either major or minor. This information, which is already provided, makes our task simpler and more accurate.

78	1	1	1	عَمَّ	عَمَّ	P	-	ʕamma	ʕamma	Eam~a	
78	1	1	2	يَتَسَاءَلُونَ	يَتَسَاءَلُونَ	V	||	jatasaːʔaluːna	jatasaːʔaluːn	yatasaA'aluwna	سأل
78	2	1	1	عَنِ	عَنِ	P	-	ʕani	ʕani	Eani	
78	2	1	2	ٱلنَّبَإِ	النَّبَإِ	N	-	ʔannabaʔi	nnabaʔi	Aln~aba<i	نبأ
78	2	1	3	ٱلْعَظِيمِ	الْعَظِيمِ	N	||	ʔalʕaðˁijmi	lʕaðˁijm	AloEaZiymi	عظم
78	3	1	1	ٱلَّذِى	الَّذِي	N	-	ʔallaðiː	ʔallaðiː	Al~a*iy	
78	3	1	2	هُمْ	هُمْ	N	-	hum	Hum	humo	
78	3	1	3	فِيهِ	فِيهِ	P	-	fiːhi	fiːhi	fiyhi	
78	3	1	4	مُخْتَلِفُونَ	مُخْتَلِفُونَ	N	||	muxtalifuːna	muxtalifuːn	muxotalifuwna	خلف
78	4	1	1	كَلَّا	كَلَّا	P	-	kallaː	kallaː	kal~aA	
78	4	1	2	سَيَعْلَمُونَ	سَيَعْلَمُونَ	V	||	sajaʕlamuːna	sajaʕlamuːn	sayaEolamuwna	علم

Figure 4: shows a sample of the multi-tiered BAQ dataset version 2.0.

7. The multi-tiered *Boundary Annotated Qur'an* dataset: version 2.0

The *Boundary-Annotated Qur'an* dataset: version 1.0 contains 13 tiers, including: 2 tiers for the Arabic word, 2 tiers for part-of-speech, 2 tiers for boundary types, etc. Version 2.0 adds another 4 tiers for the BAQ dataset. These tiers are: (i) an IPA pausal transcription of the corpus words with case ending, (ii) an IPA contextual transcription tier, (iii) transliteration tier using Tim Buckwalter transliteration scheme[4], and (iv) root for each word in the dataset. Figure 4 shows a sample of the multi-tiered BAQ dataset version 2.0.

8. Conclusions

In this paper, we have extended the development of the Boundary Annotated Qur'an: a dataset for machine learning. Version 1.0 and 2.0 of the BAQ dataset contains multiple annotation tiers in a machine readable format. IPA phonetic transcriptions of the Qur'an are newly added tiers. Pausal phonemic transcriptions with and without case endings were automatically generated and added to the dataset. These transcriptions were then manually verified and corrected to reach 100% accurate dataset. A transliteration tier was added using Tim Buckwalter's transliteration scheme for MSA Arabic words. This shows the difference between 1-to-1 letter mapping and IPA phonetic transcriptions. Finally, the root of each word in the dataset was added.

9. References

Al-Ghalayyni. 2005. جامع الدروس العربية "Jami' Al-Duroos Al-Arabia" Saida - Lebanon: Al-Maktaba Al-Asriyiah "المكتبة العصرية".

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.

Brierley, C., Sawalha, M., Heselwood, B. and Atwell, E. (Under review). A Verified Arabic-IPA Mapping for Arabic Transcription Technology Informed by Quranic Recitation, Traditional Arabic Linguistics, and Modern Phonetics. Submitted to the *Journal of Semitic Studies*.

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., Atwell, E. 2012. "Open-Source Boundary-Annotated Corpus for Arabic Speech and Language Processing." In *Proceedings of Language Resources and Evaluation Conference (LREC)*, Istanbul, Turkey.

Habash, N., Soudi, A. and Buckwalter, T. 2007. On Arabic Transliteration. In Soudi, A., van den Bosch, A. and Neumann, G. (eds.), *Arabic Computational Morphology: Knowledge-based and Empirical Methods*.

Harrag, A. and Mohamadi, T. 2010. QSDAS: New Quranic Speech Database for Arabic Speaker Recognition. In *The Arabian Journal for Science and Engineering*. 35, 2C. 7-19.

Hassan, Z.M. and Heselwood, B. (eds.). 2011. *Instrumental Studies in Arabic Phonetics*. Amsterdam. John Benjamins Publishing Company.

Ryding, Karin C. 2005. *A Reference Grammar of Modern Standard Arabic*. Cambridge. Cambridge University Press.

Sawalha *et al* 2014a. (Forthcoming). IPA transcription technology for Classical and Modern Standard Arabic.

Sawalha, M., Brierley, C., and Atwell, E. 2012a. 'Predicting Phrase Breaks in Classical and Modern Standard Arabic Text.' In *Proceedings of LREC 2012,* Istanbul, Turkey.

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.

Sharaf, A.B. 2011. Automatic categorization of Qur'anic chapters. In *7th. International Computing Conference in Arabic (ICCA'11)*, Riyadh, KSA.

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

Wright, W. 1996. *A Grammar of the Arabic Language, Translated from the German of Caspari, and Edited with Numerous Additions and Corrections*. Beirut: Librairie du Liban.

[4] http://www.qamus.org/transliteration.htm

Appendix I: Arabic > IPA mapping

Arabic consonants	IPA symbol selection	Equivalent sound (if any) in English
ا	aː	bag
ب	b	big
ت	t	tin
ث	θ	thin
ج	ʤ	jam
ح	ħ	breathy ‘h’ as in hollow or whole
خ	x	loch
د	d	dog
ذ	ð	there
ر	r	rock
ز	z	zoo
س	s	sat
ش	ʃ	shall
ص	sˤ	a *bit* like the ‘so’ sound in ***so***ck
ض	dˤ	a *bit* like 'd' sound in '***d***uck', 'bu***d***', 'no***d***'
ط	tˤ	a *bit* like 't' sound in 'bough***t***', 'bo***tt***le'
ظ	ðˤ	no English equivalent but *voiced th*-like
ع	ʕ	no English equivalent
غ	ɣ	like the ‘r’ in the French word *rue*
ف	f	fun
ق	q	no English equivalent
ك	k	king
ل	l	lemon
م	m	man
ن	n	next
ه	h	house
و	w	will
ي	j	yellow
ء	ʔ	glottal stop as in Cockney *bo**tt**le*

Shaded cell: We are using /x/ for /χ/ for better readability of IPA transcriptions

Arabic short and long vowels	IPA	Equivalent sound (if any) in English
◌َ	a	short ‘a’ as in m*a*n
◌ِ	i	short ‘i’ as in h*i*m
◌ُ	u	short ‘u’ as in f*u*n
ا	aː	long ‘a’ as in c*a*r
ي	iː	long ‘i’ sound as in sh*ee*p
و	uː	long ‘u’ sound as in b*oo*t

Appendix II: A sample of the Qur'an frequency list with IPA transcriptions of pausal form with case ending

word type number	Word type frequency	Word type	Word type in IPA
1	1673	مِنْ	min
2	1185	فِي	fiː
3	1010	مَا	maː
4	828	اللَّهِ	ʔallaːhi
5	812	لَا	laː
6	810	الَّذِينَ	ʔallaðiːna
7	733	اللَّهُ	ʔallaːhu
8	691	مِنَ	mina
9	670	عَلَى	ʕalaː
10	662	إِلَّا	ʔillaː
11	658	وَلَا	walaː
12	646	وَمَا	wamaː
13	609	إِنَّ	ʔinna
14	592	اللَّهَ	ʔallaːha
15	519	أَنْ	ʔan
16	416	قَالَ	qaːla
17	405	إِلَى	ʔilaː
18	372	مَنْ	man
19	344	إِنْ	ʔin
20	337	ثُمَّ	θumma
21	327	بِهِ	bihi
22	325	لَهُمْ	lahum
23	323	كَانَ	kaːna
24	296	بِمَا	bimaː
25	294	لَكُمْ	lakum
26	280	ذَلِكَ	ðaːlika
27	275	لَهُ	lahu
28	268	الَّذِي	ʔallaðiː
29	265	هُوَ	huwa
30	264	أَوْ	ʔaw
31	263	قُلْ	qul
32	253	آمَنُوا	ʔaːmanuː
33	250	قَالُوا	qaːluː
34	241	فِيهَا	fiːhaː
35	239	وَاللَّهُ	wallaːhu
36	234	وَمَنْ	waman
37	229	كَانُوا	kaːnuː
38	219	الْأَرْضِ	ʔalʔardˤi
39	195	إِذَا	ʔiðaː
40	190	هَذَا	haːðaː