3
TRANSLATING SACRED SOUNDS

Encoding *tajwīd* rules in automatically generated IPA transcriptions of Quranic Arabic

*Claire Brierley, Majdi Sawalha and Hanem El-Farahaty*

**Introduction**

*Tajwīd* (تَجْوِيد, *elocution*) is a long-standing disciplinary sub-field of Islamic Studies and defines theory and practice in the oral tradition of Quranic recitation (Denny, 1989). The word القرآن *qurʾān* itself is a verbal noun derived from the verb قرأ *qara* with a literal meaning of reciting or reading (aloud); and the word *tajwīd* from the triliteral root ن-م-ذ *j-w-d* has wider connotations of ‘making correct’ and ‘beautifying’ (Gade, 2004: 485). The motivation for this rule-set is therefore to avoid potential ambiguity and misunderstanding due to poor delivery and hence reception of the divine message.

We have developed transcription technology for Quranic and Modern Standard Arabic that outputs phonemic transcriptions of stand-alone Arabic words using the International Phonetic Alphabet or IPA (Sawalha *et al.*, 2014). The Arabic-IPA mapping scheme underlying our transcription technology is an important contribution of our work (Brierley *et al.*, 2016). It is informed by *tajwīd* theory, by the work of ancient Arabic linguists such as Sibawayh (d.c. 181/798), and by modern phonetics (Brierley *et al.*, 2016). Another major contribution is the dataset used to develop and evaluate our transcription algorithms. This is the *Boundary-Annotated Qurʾān* (BAQ) dataset for machine learning which maps each word in the Qurʾān to a set of linguistic labels or annotations (Sawalha *et al.*, 2014; Brierley *et al.*, 2012). This dataset or corpus is described in the section “The *Boundary-annotated Qurʾān* dataset”.

Our objectives in automating Arabic-IPA transcription are twofold: (1) to generate a pronunciation guide in a standard character set (IPA) over the entire text of the Qurʾān for the benefit of non-native Arabic-speakers; and (2) to generate canonical pronunciation or citation forms for entries in Arabic learner dictionaries. While mastering part of the IPA is a challenge for learners, this alphabet was chosen because it enables representation of an equivalent phoneme across the world’s languages via the same symbolic form (e.g. the symbol /ʃ/ denoting the “s” in “sugar”; the “ch” in “Chirac”; and realization of the letters “щ”, “ש”, “щ” in Russian, Arabic and Hebrew respectively). Moreover, many Arabic phonemes are represented in familiar “roman” script in IPA, so it is not a case of learning an entire new alphabet (see Arabic-IPA chart in Appendix 3.1). Our objectives and approach are different, therefore, from recent work in automatic speech recognition for Arabic, where the system outputs...
Translating sacred sounds

phonetic (rather than phonemic) transcriptions of Arabic words to match the pronunciation of individual speakers (Ramsay et al., 2014). Guidance in the form of IPA phonemic transcriptions is an integral part of learner dictionaries for English, such as the Longman Dictionary of Contemporary English (2014). IPA transcriptions are also planned for bilingual dictionaries of colloquial Arabic dialects; these transcriptions will again differ from ours by displaying allophonic variations characteristic of the region (Graff and Maamouri, 2012).

In this chapter we document our most recent work, which involves: (1) automatic assignment of primary stress, following syllabification, over the existing phonemic transcription tier in our Boundary-Annotated Qurʾān corpus; and (2) automatic capture and encoding of pronunciation rules governing long vowels and the consonantal subset \{ﺏ ﺩ ﺝ ﺕ ﻁ ﻕ\}. This meticulous characterization and prescription of Arabic phonemes as sacred sounds in tajwīd theory and practice lends itself to algorithmic formulation. Thus, we construe automated Arabic-IPA transcription as a form of translation, with computation as a translation intermediary. Our software outputs Arabic word forms represented as stressed and syllabified IPA sequences, with further “annotation” of tajwīd prosody, for the benefit of non-native Arabic-speaking language learners and students of the Qurʾān.

The Boundary-Annotated Qurʾān dataset

We have previously reported on our Boundary-Annotated Qurʾān (BAQ) dataset where Arabic words in both Othmani and Modern Standard Arabic (henceforth MSA) scripts are mapped to multiple information tiers of linguistic and symbolic information (Sawalha et al., 2014; Brierley et al., 2012). An example would be the MSA form ﺍﻟْﻤُﺴْﺘَﻘِﻴﻢَ (the-straight) mapped to its part-of-speech (nominal), its verse-end boundary status as a major break, and its basic phonemic transcription (ʔ almustaqiː ma)—but see Figure 3.4 for more examples. The rationale behind this corpus, as distinct from other Qurʾān-based projects such as Corpus Coranicum and the Quranic Arabic Corpus, is that it was originally designed for machine learning experimentation at the prosody-syntax interface in Arabic, and development and evaluation of Arabic chunking algorithms (Sawalha et al., 2012). Corpus Coranicum is an ongoing research project from the Berlin-Brandenburg Academy of Sciences and Humanities, and is accessible online (cf. Neuwirth, 2010). Their aim is to document the historical development of the Qurʾān, both as an oral tradition and as a manuscript, and in terms of Quranic exegesis. The Quranic Arabic Corpus is an online resource detailing and visualizing the Arabic grammar (known as أعراب iʿrāb), syntax and morphology of each word in the Qurʾān rendered in Othmani script (Dukes, 2014). Spelling variations between the Othmani and MSA scripts are discussed in the section “Tokenization in the Boundary-Annotated Qurʾān”; and differences between phonemic transcriptions in the Boundary-Annotated Qurʾān, and romanization and transliteration schemes used in Corpus Coranicum and the Quranic Arabic Corpus are discussed in the section “Automating stress assignment in pause form pronunciation”.

Prosodic speech corpora

Our precursor for the Boundary-Annotated Qurʾān is the Prosody and Part-of-Speech Spoken English Corpus (ProPOSEC) dataset (Brierley, 2011; Brierley and Atwell, 2010) where word tokens in a version of the Machine Readable Spoken English Corpus (MARSEC) have been automatically tagged with an array of prosodic and syntactic annotations which can readily be adapted as input features to a phrase break classifier. Figure 3.1 shows an extract from ProPOSEC for the phrase a similar flight. Each word token is mapped to its part-of-speech
Claire Brierley et al.

(POS) plus symbolic prosodic information as follows: the Speech Assessment Methods Phonetic Alphabet (SAMPA) phonemic transcription; syllable count; lexical stress pattern; a stressed and syllabified phonemic transcription using the DISC character set; and syllable segments assigned to primary stress weightings.

**Boundary annotation tiers in the Boundary-Annotated Qurʾān (version 1.0)**

A corpus annotated with prosodic boundaries as well as POS is a prerequisite for phrase break prediction. The *Boundary-Annotated Qurʾān* is unique in reconciling a verified Arabic phrase break annotation scheme derived from Ḥafṣ bin ʿĀṣim, a traditional and widely used Quranic recitation style, with an established benchmark for British English inherited from the *Lancaster/IBM Spoken English Corpus* or SEC (Taylor and Knowles, 1988). *Tajwīd* editions of the Qurʾān are marked up with prescriptive boundary annotations, constituting a kind of punctuation to help speakers parse the text during recitation. Figure 3.2 shows the set of boundary annotations and their denotation in Ḥafṣ bin ʿĀṣim, as used throughout the Gulf and the Levant. While some pauses during recitation will be speaker-dependent (e.g. accidental pauses), implementing the Ḥafṣ bin ʿĀṣim schema throughout is a linguistically sound solution to producing a normalized version of the corpus with authentic boundary mark-up.

In Brierley et al. (2012), we collapsed the eight degrees of boundary strength in Ḥafṣ (i.e. three major boundary types, four minor boundary types, and one prohibited stop) into the \{major, minor, none\} set familiar from SEC and MARSEC. An additional novelty was that we then segmented the text into notional sentences based on the compulsory, recommended and prohibited stops in *tajwīd* mark-up. As well as delineating grammatical units, these sentences are often realized as sequences of intonation units which resemble mainstream sentences in their ‘feeling of closure’ (Croft, 1995: 841). A selection of data from the *Boundary Annotated Qurʾān*, including boundary and sentence mark-up, appears in Figure 3.3 in the next section.

**Tokenization in the Boundary-Annotated Qurʾān**

As reported in Brierley et al. (2012), the *Boundary-Annotated Qurʾān* was initially compiled by merging morphological tags extracted from an early version of the *Quranic Arabic Corpus* (Dukes et al., 2010), with a machine readable version of the text rendered in MSA script from the *Tanzil* project (Zarrabi-Zadeh, 2007–2019). Readers will note that the *Quranic Arabic Corpus* uses the Othmani script throughout.

The word count for the Qurʾān in Othmani script is 77,430 words while the word count for the Qurʾān in MSA script is 77,797 words when tokenizing the text on whitespace (Sawalha, 2011). This difference in word count largely stems from spelling variations between the two scripts. For example, the vocative particle یَا یَا is affixed to its dependent noun in Othmani script, but originally appears as a stand-alone token in the MSA version (e.g. the word یَا یَا یَا).
Translating sacred sounds

| End of verse symbol and compulsory major break |
| Major and compulsory verse-medial break completing the meaning of a phrase |
| Major break: a break is allowed and preferable |
| Minor break: the reader can continue without pausing, but a pause is preferable |
| Minor break permitted: readers can pause if they wish, but it is better not to |
| Minor break signified by dropping the short vowel on the pre-boundary letter |
| Alternative minor boundary sites in the same phrase: either is allowed but not both |
| Non-break: pausing is not permitted as it would change the meaning of the verse |

Figure 3.2 Symbolic tajwîd annotation of pauses (waqf) in Hafs and incorporated in prosodic boundary mark-up in the Boundary-Annotated Qur’ān

<table>
<thead>
<tr>
<th>Othmani</th>
<th>Modern Standard Arabic spelling</th>
<th>“Grouped” MSA form in BAQ</th>
<th>English interlinear</th>
</tr>
</thead>
<tbody>
<tr>
<td>yāmūsā</td>
<td>yā mūsā</td>
<td>yāmūsā</td>
<td>O Musa (Moses)!</td>
</tr>
<tr>
<td>yā’ahla</td>
<td>yā ahla</td>
<td>yā’ahla</td>
<td>O people of</td>
</tr>
<tr>
<td>yālaytanī</td>
<td>yā laytanī</td>
<td>yālaytanī</td>
<td>I wish if I had</td>
</tr>
<tr>
<td>wa’allaw</td>
<td>wa’n law</td>
<td>wa’allaw</td>
<td>And if not</td>
</tr>
<tr>
<td>yā’tsā</td>
<td>yā tsā</td>
<td>yā’tsā</td>
<td>O Isa (Jesus)!</td>
</tr>
<tr>
<td>yāqawm</td>
<td>yā qawm</td>
<td>yāqawm</td>
<td>O people</td>
</tr>
</tbody>
</table>

Figure 3.3 Examples of spelling/tokenization variations between the Othmani script and MSA script

yāmūsā “O Mūsā, ‘Moses’!” in Othmani script versus yā mūsā in MSA). Figure 3.3 gives further examples of spelling variations between the two scripts.

While one-to-one Othmani-MSA mapping was accomplished automatically, spelling variations between the two writing schemes were resolved manually. Importantly, two MSA tokens (as in yā mūsā) were grouped together where appropriate to match a single Othmani token (as in yāmūsā) to preserve alignment with morphological tags extracted from the Quranic Arabic Corpus. Hence yā mūsā is grouped as yāmūsā in our MSA corpus tier. These morphological tags were then processed and mapped automatically onto two coarse-grained syntactic annotation schemes (3 and 10 POS categories) to mine syntactic correlates of prosodic boundaries in the corpus for Arabic phrase break prediction (Sawalha et al., 2012; Brierley et al., 2012).

Arabic-IPA phonemic transcription in the Boundary-Annotated Qurʾān (version 2.0)

Phonemic transcriptions are an important element of prosodic datasets and speech corpora for English, for example: ProPOSEC, (Aix)-MARSEC and SEC for British English; and the Boston University Radio News Corpus for American English (Ostendorf et al., 1996). The Boundary-Annotated Qurʾān (version 2.0) uses the IPA for automated Arabic transcription rather than the Buckwalter transliteration scheme (as in the Quranic Arabic Corpus), or one of the many romanization alphabets for Arabic (as in Corpus Coranicum). This is another major distinction between our work and the aforementioned Qurʾān projects; this is discussed in more detail in the section “Automating stress assignment in pause form pronunciation”.

49
Figure 3.4 shows selected information tiers in the *Boundary-Annotated Qurʾān* where entries span Q.70.10 and a fragment of Q.70.11. It is important to note that our transcription algorithm operates over the Qurʾān rendered in MSA rather than Othmani script, although the latter is also represented as a separate tier in our corpus.

The column headed “MSA script” in Figure 3.4 identifies the individual word tokens and character sequences that need to be decoded and processed in order to generate the equivalent IPA transcriptions. The Buckwalter scheme (rightmost column) transliterates every Arabic grapheme, including any diacritics, into an ASCII character to ensure correct display via a typical western keyboard and text editor; it is only incidentally (and not intentionally) readable for persons brought up on “romanized” alphabets.

### The Arabic-IPA mapping algorithm and transcription tool: overview

Automated Arabic-IPA transcription is not a trivial task. Our first step was development and verification of an Arabic-IPA mapping scheme; this is described in detail in Brierley *et al.* (2016) so will not be covered here. Instead we will focus on our Arabic-IPA mapping algorithm. This has two stages:

1. Pre-processing: Arabic word letters are mapped to their IPA character equivalent on a one-to-one basis using the tokenization module of the Standard Arabic Language Morphological Analysis (SALMA) Tagger (Sawalha, 2011).
2. Rule-development: linguistic rules are extracted and applied to modify the one-to-one mapping and produce the correct IPA transcription of the input Arabic word.

### The SALMA Tagger

The aforementioned SALMA Tagger is a suite of Arabic language processing tools for automatic, fine-grained morphological analysis of Arabic text (Sawalha and Atwell, 2013; Sawalha, 2011). The Tagger comprises a series of modules to be used sequentially for the following functions: tokenization; lemmatization and stemming; pattern generation; vowelization; and morphological tagging. The system draws on dedicated language resources, notably the fine-grained SALMA tag set with 22 morphological features, plus various lexica such as the SALMA ABC Lexicon (Sawalha and Atwell, 2010). Modules such as the Tokenizer can
Translating sacred sounds

also be used individually, as illustrated in the section “Further challenges and examples in full-form Arabic-IPA transcription”.

The two-stage mapping algorithm: SALMA pre-processing

In the pre-processing stage, Arabic word letters are mapped to their IPA character equivalent on a one-to-one basis using the tokenization module of the SALMA Tagger (Sawalha, 2011). A 58-entry dictionary was constructed to facilitate this one-to-one mapping from Arabic to IPA. The dictionary contains entries where an Arabic characters are mapped to single IPA characters (e.g. ٣ bā mapped to /b/), or to double IPA characters (e.g. ٣ alif mapped to /aː/).

The SALMA Tokenizer resolves gemination (٣ šaddah and the prolongation letter (٣) ً into their originals. It also limits each letter of the processed word to only one diacritic. The output is an Arabic word string which best suits the one-to-one mapping of Arabic letters and diacritics to the IPA alphabet. For example, the word أَمْنَى َـْ ثُمَّ َـْنَى (we believed) is pre-processed by the Tokenizer into ثُمَّ َـْنَى which is then transcribed on a one-to-one basis by replacing each Arabic character with its equivalent IPA symbol(s) using our constructed 58-entry dictionary. The output of this one-to-one mapping stage for our example is /ʔaːmnnáː/. This example has one error as highlighted in bold, namely: there is orthographic duplication of the vowel sound and any mapping against the preceding short vowel diacritic needs to be deleted.

The two-stage mapping algorithm: the rule-development stage

Around 50 specialized rules were developed and then ordered to resolve failures (such as /ʔaːmnnáː/ instead of /ʔaːmnnáː/ for أَمْنَى) resulting from the one-to-one mapping stage in order to produce the correct Arabic-IPA transcription. One such rule applies to words ending in مَثَأ نَثَأ which is one-to-one mapped into /aaːn/. For example, the word أَبْنَى abwāban (doors) is initially transcribed as /ʔabwːā baaːn/, and then corrected by rule to output /ʔabwːa baaːn/.

Another rule deals with the definite article by correcting the one-to-one mapping from /aːl/ to /ʔal/ in full form transcription before non-coronal consonants, namely: the set of so-called lunar letters {٠، ١، ٣، ٤، ٥، ٦، ٧، ٨، ٩}. Figures 3.5 and 3.6 illustrate the application of staged letter-to-sound rules in our algorithm for accurate transcription of the sky, and أنتما were. In each figure, the first two rows show SALMA tokenization of the original Arabic script, where readers should note that the rightmost character (i.e. the first character in the Arabic word) is processed first but that processing is carried out from left to right. The third row in each figure shows intermediate results after the one-to-one mapping stage; and the fourth row identifies issues to be resolved.

Further challenges and examples in full-form Arabic-IPA transcription

Our mapping automates Arabic-IPA transcription and outputs (in the first instance) a full-form phonemic transcription of each Arabic word. Challenges in the pre-processing stage include: resolution of hamza as /ʔ/ , regardless of form or shape {٠، ١، ٢، ٣}; and mapping pharyngeals such as ص and ض into two IPA characters: /sʕ/ , /dʕ/.

Transcription of حَمَى and حَمْيَة in Figure 3.3 as /hamːiː mun/ and /hamːiː man/ respectively raises two further issues. One is associated with resolution of تَمَنَّ as /un/, /an/, or /in/.

The second is associated with elimination of redundant grapheme-phoneme transformations inherited from the pre-processing stage: a literal, one-to-one mapping will output /hamːiː mun/ and /hamːiː maːn/ for حَمْيَة and حَمَى unless corrected by rule.

Figures 3.5 and 3.6 illustrate the application of staged letter-to-sound rules in our algorithm for accurate transcription of the sky, and أنتما were. In each figure, the first two rows show SALMA tokenization of the original Arabic script, where readers should note that the rightmost character (i.e. the first character in the Arabic word) is processed first but that processing is carried out from left to right. The third row in each figure shows intermediate results after the one-to-one mapping stage; and the fourth row identifies issues to be resolved.

51
automatically by rule (and highlighted as X). The final row shows the resulting full-form IPA transcription: /ʔasamaʔu/ (Figure 3.5) and /kaːnuː/ (Figure 3.6).

Issues addressed in Figure 3.5 are: assimilation associated with the definite article in certain contexts, and removal of duplication (due to a redundant short vowel diacritic). Issues addressed in Figure 3.6 are: removal of duplication (again due to a redundant short vowel diacritic), and discrimination via context of dual-purpose Arabic graphemes: here the waw needs to be transcribed as a long vowel, not a consonant.

Automatically generated syllabified IPA transcriptions of Arabic words

The ProPOSEC dataset dedicates several annotation tiers to symbolic representation of syllable structure and syllable weight for each entry (see the section “Prosodic speech corpora”). Such symbolic representations of prosody, as distinct from continuous variables such as fundamental frequency and duration, can be incorporated as features into a phrase break classifier as language model (Brierley, 2011; Ostendorf, 2010). For example, a rule-based chunker from Atterer and Klein (2002) first identifies function word groups, and then bundles these into intonational phrases (IPs) by limiting the number of syllables in an IP to a variable threshold figure default (setting is 13 syllables per IP) in the absence of intervening punctuation.

One application for the Boundary-Annotated Qurʾān is developing and evaluating Arabic chunking algorithms (see the section “The Boundary-Annotated Qurʾān dataset”); hence syllable information is included as an annotation tier. A detailed account of our algorithm for automatic syllabification of Arabic is intended for a future publication. We therefore summarize the process in Figure 3.7, which shows syllabification over the IPA form for بُنْورُهم (their light). First, a CV pattern is extracted from the IPA transcription, where C = consonant; V = vowel; and VV = long vowel or diphthong. Then CV patterns are used to identify syllable boundaries, where the set of possible syllable types for Arabic is defined as {CV, CVV, CVC, CVVC, CVCC}. All syllables begin with a consonant, but no syllable begins with a consonant cluster (Ryding, 2014). Hence doubled shaddah letters will tend to appear either side of a syllable boundary.
Automatic assignment of primary stress

A further refinement of the full form IPA transcription for ﺑُﻠُوْرُ ﺔْـ in Figure 3.7 would be: /bi-ˈnuː-ri-him/, where primary stress has been assigned to the second syllable by prefixing it with a small vertical modifier (ˈ) in the usual way. We have automated primary stress assignment over the whole text of the Qur’ān in version 3.0 of our corpus.

The rules for assigning lexical stress in full-form pronunciation of Arabic lend themselves to algorithmic formulation. We adopt the definition of full-form pronunciation as given in Ryding (2014), namely: Arabic words pronounced out-of-context as independent units, including pronunciation of all desinential inflexion markers. Lexical stress in Arabic is more predictable than in English; it is non-phonemic and determined by number and weight of syllables. The set of Arabic syllable types in the section “Automatically generated syllabified IPA transcriptions of Arabic words” contains one light syllable {CV}, two heavy syllables {CVV, CVC}, and two “super-heavy” syllables {CVVC, CVCC}; and readers will note that all syllables begin with a consonant. Primary stress is then assigned following the constraints defined in Ryding (2014), and in the stress algorithm for Classical Arabic in Watson (2011) as follows:

- it always falls on the first syllable of bi-syllabic words (e.g. ﺑُهْ /bi-hi/, in it; ﺭَزْقًا /ˈriz-qan/, livelihood);
- for words of three or more syllables, it falls on the penultimate syllable if it is heavy (e.g. ﺍَمْـ /ʔaː-ˈman-naː/, we believed, CVV-CVC-CVV); otherwise it falls on the antepenult as in: ﺗَنُؤُمَ /nuʔ-mi-naː/, we believe, CVC-CV-CV;
- for monosyllables, a primary stress mark is inserted at the beginning of the word as in: ﻓِ /fiː/, CVV.

Automating stress assignment in pause form pronunciation

Ryding (2014) and Watson (2011) also specify another rule governing primary stress when words are pronounced in pause form (i.e. without desinential inflexion marks). This rule states that primary stress falls on the final syllable when it is super-heavy {CVVC, CVCC}. The most recent version of our corpus now includes a new, automatically generated phonemic transcription tier which implements several transformations over the existing CV and IPA tiers, including transformation of verse terminal and/or pre-boundary words. This again differentiates our work from Corpus Coranicum and the Quranic Arabic Corpus. Figure 3.8 compares transcription of Q.1.3 in all three corpora, abbreviated as CC (Corpus Coranicum), QAC (Quranic Arabic Corpus) and BAQ (Boundary- Annotated Qur’ān).
QAC does not implement any rules for assimilation of lām when the definite article precedes “sun” letters or coronal consonants, namely, the set: {ﺕ ﺙ ﺫ ﺩ ﺭ ﺯ ﺱ ﺵ ﺹ ﺽ ﻁ ﻅ ﻝ ﻥ} which are produced with the flexible front part of the tongue. CC assimilates the lām of the definite article before coronals in every case but does not consider refining this rule when the definite article is in phrase-initial position. According to the rules of tajwīd governing al-idghām (assimilation), BAQ assimilates the lām of the definite article before coronals but retains the sound of the alif in (ال) when the definite article begins the verse/phrase: /ʔarrahmaːni/ (BAQ) versus r-raḥīmi (CC). BAQ also assumes a pause in respect of tajwīd boundary annotations when a word is in verse/phrase final position. This entails dropping the case mark and associated short vowel sound on the final word where appropriate: /rra ḥiːm/ (BAQ) rather than r-raḥ ˈ mi (CC). Readers will note the reduction in syllable count here. It also entails substitution of the long vowel alif for tanwīn fatha in pre-boundary words. Thus, our target transcription for /خَمِيْسَ مُحِمِّيْنَ (they will be put in sight of each other …’)/ would need to be transcribed as a discrete sense unit if the reader/reciter pauses here.

A moot point is resolution of geminate coronals resulting from assimilation of the definite article. Syllables in Arabic never start with a consonant cluster; hence we must attach the first /r/ in /rraḥīmi/ to the preceding syllable or drop it altogether in the final transcription. A further refinement which has bearing when generating target stressed and syllabified contextual transcriptions is to observe verse-internal tajwīd pause marks. One example is the recommended in-verse pause represented by the superscript /ۡدُ/) in Figure 3.3; this is an important prosodic and syntactic marker, rather like conventional punctuation. Hence the word ‘يَصِرُّونُهُمْ (‘… they will be put in sight of each other …’)/ would need to be transcribed as a discrete sense unit if the reader/reciter pauses here.

The tajwīd concept of prolongation or madd

There is a subset of tajwīd recitation rules governing vowel durations known as madd or prolongation. The natural madd letters are the long vowels alif, wāw and yāʾ (ا و ي). These are conceived as being twice as long as short vowels and in the first instance, are held for two “counts” during recitation. In phonological terms, counts are synonymous with morae, units
denoting syllable weight. Hence normal prolongation also applies to the Arabic diphthongs *aw*, *aw*, and *ay*. Their natural value or default setting is again two counts.

In certain contexts, however, the duration of long vowels and diphthongs is prolonged for four, five or six counts during Quranic recitation. This is termed abnormal prolongation. The full set of *tajwīd* rules defining *madd* need not concern us here; we will restrict the discussion to one category of particular interest: *madd* before pause. In this context, the *madd* letter (or diphthong) can be prolonged for up to six morae as long as the pause in question is a genuine verse/phrase boundary and not simply a disfluency.

In *tajwīd* editions of the Qur’ān, it is customary to use colour coding to annotate and thus highlight all sites where *tajwīd* rules apply, including the explicit association and significations of boundaries with final syllable lengthening (i.e. *madd* before pause). Thus, the theory and practice of Quranic *tajwīd* provides authoritative evidence from the oral tradition for the dominant stress rule for pause form pronunciation discussed in Ryding (2014) and Watson (2011) (see the section “Automating stress assignment in pause form pronunciation”), where a super-heavy final (i.e. pre-pausal) syllable of type {CVVC, CVCC} automatically attracts primary stress. We have already identified one such case when comparing transcriptions for Q.1.3 in the QAC, CC and BAQ corpora (cf. Figures 3.8 and 3.9).

Another interesting example is كِنَّ آلِ ذِئَبِينَ, literally the plan of the betrayers (Q.12.52), where is verse terminal. In a reputable *tajwīd* edition of the Qur’ān, *madd* annotation specifies prolongation of *alif* before hamza (four or five counts), and prolongation of *ya* in the final syllable (up to six counts), such that the target running IPA transcription in the Boundary-Annotated Qur’ān would be: /ˈkaj-dal-xaː-ʔi-ˈniːn/. Readers will note this is a stressed and syllabified transcription which includes assimilation. Moreover, if we were to adopt the usual colour-coding system for *tajwīd* annotation, we could distinguish the two different categories of “abnormal” *madd* here as in: /ˈkaj-dal-xaː-ʔi-ˈniːn/. This phrase then carries three main stresses or beats: (1) on the diphthong (a case of natural *madd* so not colour-coded); (2) on the long vowel before hamza; (3) on the final super-heavy syllable {CVVC}. These beats increase in intensity (i.e. duration) over the phrase.

**Evaluation of automatic stress assignment over Quranic Arabic**

Prosody is inherently variable. There is no single ‘gold standard’ prosodic realization or representation of a given utterance, and there is variation in the prosodic performance of individuals as well as across different speakers (Hirschberg, 1999: 7). Human judgement is therefore as essential part of any evaluation procedure for automatically generated prosodic annotation to test the acceptability of annotations in terms of accuracy and naturalness (Viana, 2003).

Lexical stress in Arabic is not fixed, as in languages such as Icelandic, where stress usually falls on the first syllable of a word, or Polish where it usually falls on the penultimate syllable. Nor is it varied as in English. In Arabic, stress placement is said to be regular (Ryding, 2014). Though stress in Arabic may fall on different syllables, this follows predictable patterns which can be captured by rule, and which have previously been discussed in the section “Automatic assignment of primary stress”. In this section, we focus on outputs from automated stress assignment over Quranic Arabic, and expound particularly challenging examples for human evaluation; a detailed account of the algorithm itself is reserved for another paper.

**Measuring inter-annotator agreement**

We undertook two small inter-annotator agreement studies to ensure that our principal annotator understood the annotation task of checking automatically assigned primary stress
weightings in our dataset. Prior to this, our annotator had been thoroughly acquainted with our Arabic-IPA mapping and the rules for allocating primary stress over full form Arabic pronunciation forms as defined in Ryding (2014) and Watson (2011). The first study was conducted over an unseen 2,000-word sample by two human annotators, one of whom was a native Arabic-speaking linguist, and the other a computational linguist. The annotation task entailed inspection of automatically assigned primary stress weightings for all words in the sample, and marking agreement or disagreement with the program output in each case. A similar study could not be carried out between human versus machine annotator in this case because the Kappa measure is not recommended when agreement is rare for one category but not with another (Viera and Garrett, 2005).

The Kappa statistic is generally used to measure concordance in categorical sorting between two or more annotators. This statistic is a ratio which expresses: (1) the excess of observed over expected concordant items, versus (2) the chance expected number of non-concordant items (i.e. observed over expected accuracy/random chance). It is calculated via the formula: \((\text{observed accuracy} – \text{expected accuracy}) / (1 – \text{expected accuracy})\). The results of our first study appear in Table 3.1 and represent a fair level of inter-annotator agreement, with a Kappa coefficient of 0.38 and a 95 per cent confidence interval. This can be interpreted as 38 per cent agreement (concordance) in excess of coincidental or chance agreement.

In this first study, our two annotators agreed on 1,987 instances (1,983 Yes and 4 No) and the observed accuracy is therefore: \(0.9935\) (i.e. \(1,987/2,000\)). The calculation for expected accuracy is less straightforward, and involves computation over marginal frequencies for each annotator in respect of each class. The value for expected accuracy in our case is \(0.9896\), calculated as: \(((1,992*1,987) / 2,000) + ((13*8) / 2,000)) / 2,000\). This value is then inserted into the final equation for computing Kappa: \((0.9935 – 0.9896) / (1 – 0.9896) = 0.375\), rounded up to 0.38.

As previously stated, the level of agreement on the first study is rated as fair for Kappa values in the range: 0.21 to 0.40 (Viera and Garrett, 2005). Measuring inter-annotator agreement is often an iterative process, and we wanted to achieve a more congruent score. Having

Table 3.1 Our two annotators agreed on 1,987 instances, and disagreed on 13 instances

<table>
<thead>
<tr>
<th>Annotator 1</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotator 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>1,983</td>
<td>4</td>
<td>1,987</td>
</tr>
<tr>
<td>No</td>
<td>9</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td>Total</td>
<td>1,992</td>
<td>8</td>
<td>2,000</td>
</tr>
</tbody>
</table>

Table 3.2 Our two annotators agreed on 994 instances, and disagreed on 6 instances

<table>
<thead>
<tr>
<th>Annotator 1</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotator 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>988</td>
<td>6</td>
<td>994</td>
</tr>
<tr>
<td>No</td>
<td>0</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>988</td>
<td>12</td>
<td>1,000</td>
</tr>
</tbody>
</table>
discussed all 13 instances where one or both annotators queried system outputs in the first study, we conducted a second inter-annotator agreement study on a new 1,000-word sample. The results from this second study appear in Table 3.2 and now represent substantial agreement, with a Kappa coefficient of 0.66 (falling in the range of 0.60 to 0.80 (ibid)) and a 95 per cent confidence interval.

**Overall accuracy**

We can report 97 per cent accuracy for our stress-assignment algorithm over canonical, phonemic pronunciation forms as verified manually by our native Arabic-speaking linguist. The task of manual verification was made feasible by generating a list of Quranic Arabic word types, arranged in descending order of frequency, and mapped to their stressed and syllabified IPA transcriptions. This reduced the number of items to be considered from 77,430 (the total number of word tokens in our corpus) to 17,606 (the total number of word types). We consider this a suitable approach for verifying full-form, phonemic pronunciation forms where case endings are retained. In all, 508 transcriptions were queried by our principal annotator, hence our reported overall accuracy of 97 per cent: ((508/17,606)*100).

**Difficult cases highlighted in evaluation of primary stress assignment**

In this section, we focus on outputs from automated stress assignment over Quranic Arabic, and expound particularly challenging examples for evaluation; a full account of the algorithm, including further resolution of transcription errors, is reserved for another paper. A principal area for discussion here is word-initial particles which necessitate modification of our algorithm following expert human judgement of algorithmic outputs. These particles can be further divided into bound versus unbound morphemes.

Arabic follows a root-and-pattern morphology which is commonly described by using examples of derivational verbs (Watson, 2002). Watson elaborates more on this in the following quotation:

> in Standard Arabic, the imperfect is formed by changing the quality of the rightmost stem vowel and adding an imperfect prefix (ʔu- {first person singular}, yu- {third person masculine}, tu- {second/third feminine singular}, nu- {first person plural}). The passive is formed by a change in the vocalic melody. (2002: 124)

Arabic distinguishes between inflected bound morphemes marking definiteness, case, mood and gender, and uninflceted morphemes which include, among others, conjunctions, prepositions and various particles such as interrogatives and negatives (Badawi et al., 2003). Unbound or free morphemes include the vocative بِالأَمْزِيْنُ /ja:/ as in يَا بِالأَمْزِيْنُ /ja:- 'mar-ja-mu/, O Mary. Bound morphemes can be attached to the word as explained by Watson above or they can be uninflceted particles, such as the coordinating conjunctions وَا (and), and فَا (and so) or subordinating conjunctions such as the particle of cause فَ as in فَلَنَا /fa- 'qul-na:/ ([so] we said).

**Word-initial wa-**

The morphemes وَا and فَا have both posed challenges for automatic stress assignment. An illustrative example would be the word وَا and not وَا from Figure 3.4, where stress is initially
allocated to the first syllable via algorithmic rule: /ˈwa-la/. However, even though this word token is bisyllabic, stress should not fall on the first syllable as per rule, but on the second syllable, such that the correct stressed and syllabified transcription will in fact be: /wa-ˈla/. Another example is the combination of coordinating conjunction wa- with the emphaser /qad/ in ʿabd and indeed which is automatically transcribed as /wa-qad/, but where the second syllable should carry primary stress: /wa-ˈqad/. A further example would be ʿāl and when, where the bound morpheme wa- attracts primary stress according to our unmodified transcription rules, resulting in: /ˈwa-ʔi-dī/; but once again primary stress should fall on the second syllable: /wa-ˈʔi-dī/ and when. The same applies to ʿān and that, correctly transcribed as /wa-ʔa-ni/ not /wa-ʔa-ni/. These and similar cases necessitate modification of our stress-assignment algorithm to incorporate (coarse-grained) morphological analysis as a pre-processing step to identify instances of bound morphemes. Our solution implements morphological analysis via the SALMA Tagger (Sawalha, 2011) which decomposes Arabic word tokens into five parts: proclitics, prefixes, stem or root, suffixes and enclitics. Bound morphemes are recognized in SALMA as proclitics. Thus for words that start with bound morphemes (i.e. proclitics) this initial element will be discounted in the stress-assignment rules.

**Word-initial fa-**

The coordinating and subordinating bound morpheme ʿfa- requires a similar solution. For example, it is initially allocated primary stress in ʿalī /fa-ʔi-da:/ then when, whereas expert human judgement assigns it to the penultimate syllable: /fa-ˈʔi-dā/,. Another example would be ʿfa- attached to a monosyllabic particle as in ʿ affidavit /fa-lam/ and not; where primary stress should be assigned to the second syllable: /fa-ˈlam/. These examples indicate that bound morphemes attached to bisyllabic and monosyllabic words should not attract primary stress. A further interesting point to make about particles here is that one cannot pause after them if they do not end with sukūn; this means that they need to be pronounced in combination with syllables that follow.

**Problems caused by alif**

Another problem for transcription and automatic stress assignment when working with MSA script is the case of dagger alif, the alif which is pronounced but not written in certain words. Although it was successfully captured in transcription of ʿāf /ha-ːdāː / this, it was not captured when the same word appeared with three prefixed morphemes as in: ʿaf ēdd / fa-ˈba-ʾ bi- ha-ːdāː / This correct transcription for this word is ʾa-bi-ʾ ha-ːdāː: primary stress should fall on the penultimate long syllable {CVV}. Another example of the missed dagger alif is the word ʾa-li-ki/ as well where the long vowel was initially not captured. The correct transcription of this word is ʾa-ʾli-ki/. Our approach to resolving dagger alif scenarios when working with MSA script is to group all such special cases along with their correct, full-form IPA transcription, in a transcription lexicon as input to the algorithm. At runtime, the modified algorithm cross-checks each Arabic word token against mapped pairs of word and manually verified stressed and syllabified phonemic transcription in the dictionary; any matches return a bespoke transcription for that word before the ordered set of rules are systematically implemented.

Other examples of technical issues surrounding alif include words where the alif is written but not pronounced as in ʾaː two hundred, originally transcribed as /ˈmīːtaːtaːjīn/.
Translating sacred sounds

version of this word should be /mi-ʔa-ˈtaj-nil/, with primary stress on the penultimate syllable {CVV}. A related example is the word مائتا hundred which should correctly be transcribed and pronounced as /ˈmi-ʔa-tin/ (not /ˈmia:ʔatin/), with primary stress on the first syllable.

A final issue is transcription of the long vowel alif in cases such as: وَالدَّيْنَةُ وَالدَّيْنَةُ. Although stress was assigned to the correct syllable via algorithmic rules, expert judgement identified transcription errors surrounding alif: /ˈwa-li-dun/, /wa-ˈli-di-hi/, /wa-li-ˈda-ti-ka/ instead of /ˈwa-li-dun/, /wa-ˈli-di-hi/, /wa-ˈli-da-ti-ka/. Figure 3.10 gives further examples of words with special orthography in Othmani and MSA scripts and their correct Arabic-IPA transcription stored in the transcription lexicon, ready for look-up as a preliminary algorithmic step when automating stressed and syllabified phonemic transcription in MSA.

Rule-based capture and encoding of the tajwīd effect of qalqalah

As we have seen, tajwīd editions of the Qurʾān include fine-grained boundary annotations as a kind of punctuation (see the section “Prosodic speech corpora”), and use colour-coding to highlight: co-articulatory effects (see the section “Automating stress assignment in pause form pronunciation”); and normal and abnormal prolongation of long vowels in specific contexts (see the section “The tajwīd concept of prolongation or madd”). Yet another branch of tajwīd is concerned with precise phonemic realization of the consonantal subset {ﺏ، ﺩ، ﺝ، ﺨ، ﻕ}. This is the effect of qalqalah or “vibration”.

The qalqalah consonants are pronounced with a definite ‘punch’ under certain constraints to ensure correct pronunciation and correct perception (Watson and Heselwood, 2014). In tajwīd, members of the set {ﺏ، ﺩ، ﺝ، ﻁ، ﻕ} are given an emphatic delivery in weak prosodic positions, namely: (1) at the end of a word; or (2) within a word and immediately preceding another consonant.

We have reported on rule-based capture of all qalqalah sites in the Qurʾān in a previous paper (Brierley et al., 2014). Here we summarize our algorithm before demonstrating the potential benefits to Quranic Arabic language learners of encoding qalqalah and other tajwīd effects in automated Arabic-IPA transcription.

Language processing for Arabic poses challenges above and beyond those encountered for languages with romanized alphabets, not the least being that when programming, the researcher does not engage directly with the Arabic script but with Unicode strings instead. An example is given in Figure 3.9, which also highlights a verse-medial (and word-internal) qalqalah site in Q.85.12. The word in question is بَطَّنُ grip, which consists of six graphemes, including one instance of sukūn. The Unicode equivalent also contains six elements demarcated by the backslash character (\), where the fourth one along from the left in Figure 3.11 represents sukūn: \u0652\.

Our search algorithm for qalqalah first returns all Quranic verses where any member of the set {ﺏ، ﺩ، ﺝ، ﻁ، ﻕ} occurs. A second search over the returned list of verses is then conducted via Regular Expressions (REs) or search patterns. Figure 3.12 decodes the RE used in our algorithm to locate word-internal qalqalah sites. Rule-ordering is important here: the RE operates over word tokens in each verse string to determine whether or not each Arabic consonant belongs to the qalqalah set, is associated with sukūn, and is word-internal.

Q.85.12 is an interesting case because it furnishes a second instance of qalqalah which must be realized in pause form pronunciation. The first instance is in بَطَّنُ as we have seen.
The second instance is in the verse-terminal word ﷺاﷲـا لـاهلُ surely strong. In full form pronunciation, ﷺاﷲـا لـاهلُ would be transcribed as /la-ʃ adiː dun/. However, in pause form pronunciation, the tanwīn will be dropped so that what was originally the penultimate syllable in /la-ʃ a-ˈdiː dun/ becomes a super-heavy final syllable: /la-ʃ a-ˈdiːd/. This transformation then jeopardizes the final consonantal phoneme (and member of the qalqalah set)—which might be swallowed in pronunciation unless given that distinctive qalqalah “bounce”. In future work we will visualize this tajwīd logic as a learning aid for students of Arabic and the Qurʾān, as illustrated in Figure 3.11. Readers will note that we have also encoded the category of “madd before pause” in our target output (cf. Figure 3.13).

Conclusions

The Boundary-Annotated Qurʾān is an evolving database for the study of Arabic prosody, where each entry (i.e. Arabic word form or word token) is tagged with an informative array of linguistic information. In the latest version of our corpus, this array includes: coarse-grained syntactic category; break-type (i.e. whether the word immediately precedes a boundary or not); CV pattern; and stressed and syllabified IPA transcription. Features extracted from these annotations can be incorporated into phrase break models for Arabic. Moreover, these annotations encode different aspects of Arabic prosody and can be further processed to output canonical pronunciation forms with full tajwīd for non-native Arabic speakers and students of the Qurʾān.
Language processing for Arabic entails extra transformations or “translations” to cope with the Arabic script itself. We have discussed how each Arabic grapheme is mapped to its discrete Unicode symbol, and how researchers and programmers in Arabic are working directly with disembodied Unicode sequences rather than actual Arabic words. We have also discussed how another layer of abstraction is superimposed over Unicode sequences to conduct searches, namely: the formulaic language of Regular Expressions (see the section “Rule-based capture and encoding of the \textit{tajwīd} effect of \textit{qalqalah}”).

IPA characters constitute another alphabet that requires “translation” into Unicode for language processing and correct screen/printer display. Ostensibly, there are two different mappings underpinning our IPA transcriptions of Quranic Arabic: (1) from Arabic grapheme to IPA character (cf. Brierley \textit{et al.}, 2016); and (2) from IPA character to its Unicode equivalent. However, we have also shown that the first step (i.e. from Arabic grapheme to IPA character) entails additional transformations to modify literal one-to-one grapheme–phoneme pairings and output correct transcriptions (see the section “Arabic-IPA phonemic transcription in the \textit{Boundary-Annotated Qurān} (version 2.0)”). For example, the different surface forms of hamza must be mapped to one IPA character on a many-to-one basis.

Finally, we have “translated” or encoded linguistic rules for Arabic in further algorithmic manipulation of intermediate, abstract forms (i.e. Unicode sequences) to annotate our IPA transcriptions with canonical prosody. This has two dimensions, the first being codification of linguistic rules for syllabification and primary stress, where both features interact with prosodic-syntactic boundaries in full-form versus pause-form pronunciation. The second challenge has been codification of Quranic recitation rules (i.e. \textit{tajwīd}). This entails formulaic capture of target contexts via regular expressions.

To conclude, we have presented a novel and versatile algorithm that outputs a variety of IPA phonemic transcriptions as an aid for Arabic language learners and/or Islamic scholars. Our transcriptions range from basic, full-form pronunciation of stand-alone words (as in: /ʔīnna/ /batʕa/ /rabbika/ /laʃādi:duːn/), to concatenated intonational phrases displaying interconnected syllables, beats (i.e. primary stress), and recitative enhancements: /ʔin- na- batʕ- fa- rab- bi- ka- la- fa- diː d/.
Notes

1  www.corpuscoranicum.de/index/index/sure/1/vers/1.
4  http://tanzil.net/download/.
8  www.bwgriffin.com/gsu/courses/edur9131/content/InterRater_nominal_codes.pdf.

Further reading


References


Translating sacred sounds


Watson, J. and Heselwood, B. (2014). Can spoken Southern Arabic and Modern South Arabian inform research into Qur’anic tajwīd? Arabic and Middle Eastern Studies Talks (5 March 2014), the School of Modern Languages, Cultures and Societies, University of Leeds, UK.

## Appendix 3.1

### Summary chart for Arabic >IPA mapping scheme in BAQ

<table>
<thead>
<tr>
<th>Arabic consonant</th>
<th>IPA symbol</th>
<th>Illustrative equivalent in English</th>
<th>Arabic consonant</th>
<th>IPA symbol</th>
<th>Illustrative equivalent in English</th>
</tr>
</thead>
</table>
| ﺏ | a: | farm | ﺕ | t | None – but emphatic; a bit like “t” in *star*
| ﺏ | b | bang | ﺕ | ٌ | None – but pharyngealized/velarized voiced fricative
| ﺕ | t | time | ﺭ | ّ | -
| ﺛ | ﺧ | thing | ﺭ | ٪ | None – but realized as voiced velar/uvular fricative e.g. French *rester* (to stay)
| ﺝ | ﺝ | jump | ﺟ | f | family
| ﺡ | ﻩ | homily | ﺡ | q | None – but a bit like “c” in *scar*
| ﺝ | ﻩ | loch | ﻩ | k | king
| ﺝ | ﻩ | dynamite | ﻩ | l | lamb
| ﺝ | ﻩ | though | ﻩ | m | man
| ﺝ | ﻩ | ring | ﻩ | n | nut
| ﺝ | ﻩ | zen | ﻩ | h | hip
| ﺝ | ﻩ | sun | ﻩ | w | went
| ﺝ | ﻩ | show | ﻩ | j | yellow
| ﺝ | ﻩ | None – but a bit like “s” in *salt* | ﻩ | ء | glottal stop as in Cockney *motor*
| ﺝ | ﻩ | d | None – but voiced and pharyngealized

**Shaded cell:** We use /x/ instead of /χ/ for better readability of IPA transcriptions.

### Arabic short and long vowels, plus diphthongs

<table>
<thead>
<tr>
<th>Arabic short and long vowels, plus diphthongs</th>
<th>IPA symbol</th>
<th>Illustrative equivalent in English</th>
</tr>
</thead>
<tbody>
<tr>
<td>ﺏ</td>
<td>a</td>
<td>bat</td>
</tr>
<tr>
<td>ﺝ</td>
<td>i</td>
<td>ink</td>
</tr>
<tr>
<td>ﺝ</td>
<td>u</td>
<td>bun</td>
</tr>
<tr>
<td>ﺝ</td>
<td>a:</td>
<td>(similar to) farm</td>
</tr>
<tr>
<td>ﺝ</td>
<td>i:</td>
<td>freeze</td>
</tr>
<tr>
<td>ﺝ</td>
<td>u:</td>
<td>blue</td>
</tr>
<tr>
<td>ﺝ</td>
<td>ء</td>
<td>aj</td>
</tr>
<tr>
<td>ﺝ</td>
<td>ء</td>
<td>aw</td>
</tr>
</tbody>
</table>