Journal of Semitic Studies LXI/1 Spring 2016 doi: 10.1093/jss/fgv035

A VERIFIED ARABIC-IPA MAPPING FOR ARABIC TRANSCRIPTION TECHNOLOGY, INFORMED BY QURANIC RECITATION, TRADITIONAL ARABIC LINGUISTICS, AND MODERN PHONETICS

CLARE BRIERLEY, MAJDI SAWALHA, BARRY HESELWOOD and ERIC ATWELL

UNIVERSITY OF LEEDS UNIVERSITY OF JORDAN

Abstract

In this paper, we present a detailed mapping from the graphemes of Modern Standard Arabic (MSA) to symbols from the International Phonetic Alphabet (IPA) for automated transcription of Arabic text. This mapping is distinctive in several ways. First, the corpus used in rule development is the full text of the Qur'ān rendered in fully pointed MSA. Second, we validate our scheme via automatically-generated frequency distributions of Arabic letters and diacritics over the whole corpus to anticipate and disambiguate non-trivial, compound grapheme-to-phoneme *events*, thus reducing the number of letter-to-sound rules. Such difficult cases include: the definite article; the letters *alif*, *wāw*, and *yā'*; the variant forms of *hamza*; the *tanwīn* case mark; and words with special pronunciations. Finally, our mapping scheme is informed by theory and practice from medieval Arabic linguistics and traditional Quranic recitation or *tajwīd*; we make a novel contribution with new translations for ancient terms which incorporate concepts familiar to modern phoneticians. Our principal objective in automating Arabic-IPA transcription is to generate phonemic citation forms of Arabic words to enhance Arabic dictionaries, to facilitate Arabic language learning, and for natural language engineering applications.

1. Introduction

In this paper, we present a mapping from the graphemes of Modern Standard Arabic (MSA) to symbols from the International Phonetic Alphabet (IPA) customized specifically for automated transcription of Arabic text. This mapping is detailed and distinctive since the corpus used in development is the full text of the Qur'ān rendered in fully

pointed MSA (Brierley et al. 2012), and because it is validated by automatically-generated frequency distributions of Arabic letters and diacritics over the whole corpus to ensure comprehensive coverage, and to anticipate and disambiguate grapheme-to-phoneme *events* that are potentially challenging for computational processing. Examples of such events, which defy one-to-one letter to sound transformations, are given in Table 1 for immediate reference, and are discussed in more detail in Section 6, and in our parallel paper (Sawalha et al. forthcoming), which implements the mapping in transcription technology for the Qur'ān and contemporary Arabic texts. Table 1 juxtaposes a literal and erroneous one-to-one mapping, where each Arabic letter/diacritic is replaced by an IPA character, with the target transcription (IPA Gold), and logs the number of errors to be resolved by carefully-ordered rules, demonstrating that automated Arabic > IPA transcription is a non-trivial task.

Table 1. One-to-one letter-to-sound transformations need further rule-based resolution and our mapping reduces the number of rules for automated Arabic > IPA transcription.

Word	Romanization	Meaning	IPA 1-to-1	IPA Gold	Errors
يَتَسَاءَلُونَ	*yatasā'lūna*	they are asking one another	jatas**aa**ːʔal**uw**na	jatasaːʔaluːna	2
كَانُوا	*kānū*	they were	k**aa**ːn**uw**aː	kaːnuː	2
الاتِّصَالَاتِ	*al-'itiṣālāti*	the communications	**ʔala**ːttisˤaːlaːti	**ʔali**ttisˤaːlaːti	1
كُوزْمُوبُولِيْتَانْ	*kuzmubulitan*	cosmopolitan	k**u**ːzm**u**ːb**u**ːliːtaːn	k**u**zm**u**b**u**lit**a**n	5

The mapping scheme is applications-oriented. One of our objectives in automating Arabic-IPA transcription is to generate a canonical pronunciation or citation form for each Arabic word in our corpus as a model for Arabic learner dictionaries. Hence we have used an international standard (the IPA) rather than adding to the proliferation of romanized transcription schemes for Arabic.

Another motivation for the mapping is stylistic and stylometric analysis of the Qur'ān. Studies of a numerological nature, based on letter counts which purport to unlock the secrets of the Quranic text, are popular but somewhat dubious (cf. Philips 1987). Reputable publications (both ancient and modern) on the *characteristics* of individual Arabic letters and sounds can be found in printed (Surty 2000) and e-learning materials on *tajwīd* defined as the principles of correct Quranic recitation (§4); and *tajwīd* editions of the Qur'ān (cf. Dar-al-Maarifah 2008) make explicit use of prosodic mark-up to highlight letters accorded special recitative effects. There is, therefore,

legitimate interest in the semiotics of letters and sounds in the Qur'ān. As part of our planned research, we will customize statistical techniques like keyword extraction (cf. Scott 2012) to explore significant *segmental* features in chapters of the Qur'ān (cf. Brierley and Atwell 2010 for similar work for English), and a phonemic representation of the text from our Arabic > IPA mapping will facilitate this.

This research is funded by the Engineering and Physical Sciences Research Council (EPSRC) for a project entitled: *Natural Language Processing Working Together with Arabic and Islamic Studies* (Atwell et al. 2013). Project deliverables include tools and resources for Arabic transcription and Text Analytics, plus new interdisciplinary approaches to the study of religious and other literary texts.

This paper is structured as follows: the Arabic language and its uniqueness (§2); ancient analyses of Arabic phonetics in traditional Arabic Linguistics (§3); comparative categorization of Arabic phonemes in Quranic recitation, the Arabic linguistic tradition, modern Phonetics, and Visual Text-to-Speech (§4); comparison of Arabic transcription, transliteration and romanization (§5); a detailed presentation of the Arabic > IPA mapping (§6); the *Boundary-Annotated Qur'ān Dataset for Machine Learning* (Sawalha et al. 2014; Brierley et al. 2012) with Quranic verses transcribed in IPA (§7); conclusions (§8).

The following typographic conventions are followed in this paper: in-text romanization of Arabic words will be conducted as per the recommended scheme for this journal, and the same scheme will be used when referring to letters of the Arabic alphabet; the automatically generated IPA transcriptions of Arabic words are phonemic and are therefore encased in slashes //.

2. The Sounds of Modern Standard Arabic: an Overview

Arabic has twenty-eight consonants, three long vowels, three short vowels and two diphthongs. The count for consonants includes *wāw, yā'* and *hamza*, but not *alif,* as in ancient writings on Arabic phonetics (§3–4). For standard vowels, the long and short forms /iː/ and /i/, and /uː/ and /u/ are close front (palatal) and close back (velar) vowels respectively, and /aː/ and /a/ are open central vowels (al-Khatib 2003; Newman 2005; Damien et al. 2009). Newman and Verhoeven (2002) note the similarity of this inventory of Arabic vowels to tongue positions for the same described in medieval texts (§4.1). The

diphthongs /ay/ and /aw/ are approximate equivalents to the complex vowels in *night* and *now* for English.

Arabic linguists dating as far back as the eighth century have highlighted the distinctive characteristics of Arabic which are associated with its pharyngealized sounds and emphatic consonants, especially ض (*ḍād*) or the phoneme /dˤ/. The *ḍād* is described by the early Arab grammarians as a voiced, pharyngealized, *lateral fricative*, not the voiced, pharyngealized, apico-alveolar stop of many modern dialects (e.g. Egyptian, Sudanese and Levantine Arabic); and discussions as to its status are still ongoing (§3). Al-Khatib (2003) suggests that the ancient form of *ḍād* was a complex consonant that began with occlusion, and ended with sibilance, meaning it was an affricate (see Heselwood 1996: 22). Both ض and ظ (*ẓā'*) — also classed as a voiced, pharyngealized fricative, but produced in the interdental region — are very rare among the world's languages, and as far as is known, their geminate or doubled counterparts are unique to Arabic (Newman 2002; 2005). The total of twenty-eight Arabic consonants (irrespective of their geminate forms) is unusually high, while the number and quality of vowel sounds is well below average, representing only the fundamental or most common vowels in the world's languages (ibid.). In phonetics, consonant segments are categorized and differentiated in terms of their place (e.g. pharyngeal, labial) and manner (e.g. fricative, stop) of articulation, namely: *where constriction is narrowest* in the vocal tract, and the *degree* of constriction. Arabic is also unusual in that it has a particularly rich set of obstruent phonemes many of which group into voiced-voiceless-emphatic triads (Cantineau 1960: 184).

3. Traditional Arabic Linguistics

The official *Othmanic Recension* of the Qur'ān in the eighth century is regarded as the catalyst for early scholarly interest in the Arabic language, centred on the various *qirā'āt* قراءات or Quranic readings/recitations (Bohas et al. 1990; Shah 2003a), and within the rapidly evolving schools of linguistic science at Basra and Kufa. Shah (ibid.) ascribes the genesis of Arabic linguistic thought to the *qurrā'* القراء or Quranic readers (albeit a minority of them) whose linguistic curiosity superseded their core preoccupation with Quranic diction and its preservation (e.g. functional treatises on orthography and diacritics). He contends that indigenous forms of grammatical analysis used to authenticate variant readings, and originating in particular from the Hijaz, may explain the provenance of linguistic terminology used in

later and more sophisticated branches of Islamic scholarship (ibid.), and that Meccan readers mentored Basran linguists (Shah 2003b).

The foremost Basran linguist was Sibawayh سيبويه (d. 177/798), whose pioneering work on Arabic linguistics included a detailed ordering and description of Arabic consonants in terms of *makhārij al-ḥurūf* مخارج الحروف and *ṣifāt al-ḥurūf* صفات الحروف: the *exits* and *characteristics* of the *letters*, concepts which have much in common with *place* and *manner* of *articulation* in modern phonetics (Heselwood and Hassan 2011; Dickins 1999; Semaan 1968). Sibawayh's precise categorization of Arabic letters in terms of their place of articulation is cross-referenced in Table 2. As an example, he stipulates that *ḍād* is produced between: '…the beginning of the edge of the tongue and the adjoining part of the molars…' (Dickins 1999). This agrees with contemporary thinking on the lateral articulation of *ḍād* (Heselwood and Hassan 2011) and its preservation in modern South Arabian languages such as Mehri (Watson and al-Azraqi 2011). Furthermore, the classification *lateral* and *fricative* is also taught as the correct aspect and manner of articulation for *ḍād* in Quranic recitation (Mansour 2013).

On manner of articulation, Sibawayh divides the obstruents into stops which he calls *al-shidda* الشدة (*intensity*), and fricatives which he calls *al-rikhāwa* الرخاوة (*laxity*). The distinction *majhūra* مجهورة (*unbreathed*) versus *mahmūsa* مهموسة (*breathed*) divides all the phonemes into two classes based on glottal settings (Heselwood and Maghrabi 2015). He also distinguishes the set of emphatic consonants {ظ ط ض ص} in his discussion of *lidding* or emphasis (Dickins 1999). These terms are cross referenced in Table 3. Sibawayh's comprehensive inventory of Arabic consonants is still current (Semaan 1968: 56).

Other significant contributors to the nascent field of Arabic phonetics were: al-Sakkaki السَّكَّاكي (d. 340/951) who produced the first known vocal tract diagram, again compatible with today's account (Heselwood and Hassan 2011); Ibn Jinni ابن جنّي (d.392/1002) who wrote the first book exclusively devoted to Arabic phonetics and phonology: the *Khasā'is* (Bohas et al. 1990); and Ibn Sina ابن سينا (d.429/1037) who documented existing knowledge of speech acoustics and physiology (Heselwood and Hassan 2011; Bohas et al. 1990; Semaan 1968). Shah (2003b) contends that these revolutionary approaches and insights were triggered by a willingness, pioneered by the Basran school, to extend the investigative corpus of Arabic beyond scripture to profane texts (i.e. poetry) and data gathered from fieldwork. Throughout, these early grammarians and philologists

developed an empirical approach, recording and attempting to explain mismatches between theory and practice using specific examples from Bedouin speech.

4. The Islamic Studies Sub-Field of *Tajwīd*

Tajwīd (i.e. the principles of correct recitation of the Qur'ān), is an intense subject of study in its own right, and any introduction to the subject will undoubtedly include the enigmatic statement that *every Arabic letter has its right and its due* (Harrag and Mohamadi 2007). The *dues* denote temporary characteristics such as contextual allophonic variation, plus accenting or prolongation, and special articulatory effects realized under certain constraints. The *rights* are said to denote inherent characteristics which we can construe to mean the identity of a phoneme in terms of its oppositional relations with other phonemes in the language.

4.1 Teaching Materials for Tajwīd

Tajwīd theory also resonates with other phonetics-based concepts such as place and manner of articulation. *Tajwīd* points of articulation (*makhārij al-ḥurūf*) are more precisely translated as sound *exits*. There are eight of these in total, as they appear on a reputable *tajwīd* website (http://www.quran1.net) developed by a leading expert in Quranic recitation: Professor Mansour of the Faculty of Sharīʿa (Islamic Studies) at the University of Jordan. The same sound exits are listed in a widely-used textbook on *tajwīd* recitation written for native Arabic speakers (Shukri et al. 2013: 62–3); and *makhārij al-ḥurūf* as a concept also resurfaces in the United Nations (UN) romanization scheme for MSA (Khraish 2007) where sound *exits* are paraphrased as *ways out*. Finally, a detailed illustration of the vocal tract showing the phonation of individual sounds and sets of sounds feature in a *tajwīd* edition of the Qur'ān (Dar-al-Maarifah 2008).

The parallels between all these sources and the fine-grained ordering of Arabic letters and their places of articulation in Sibawayh's account are evident. They are presented in Table 2, which compares sound exits and their respective sets of letters in both ancient traditions. For Sibawayh, material has been drawn from an idiomatic English translation by Dickins (1999), which suggests that what Sibawayh meant by places of articulation was the space between the articulators *where the sound comes out* — hence sound *exits*. For the latter, English translations of Arabic terms used in the *tajwīd* website have been

approximated as textbook labels from a typical vocal tract diagram. As is common, both passive and active articulators have been used for clarification in some instances; this applies in particular to the sets {ر ل ن} and {ز ص س}, where the Arabic terms suggest some distinction between *dhulq* ذلق (tongue blade) and *asal* أسل (tongue tip) respectively (Dickins 2013). The phoneme *ḍād* is grouped with the palatal fricatives {ج ش ي} as *nearest neighbours* (Mansour 2013) in a diagram of the vocal tract on the *tajwīd* website, but elsewhere on the same website, it is treated as a special case: it is the sole phoneme assigned the property of *al-istiṭālah* الإستطالة (*elongation*). The vocal tract diagram in the *tajwīd* Qur'ān (Dar-al-Maarifah 2008) clearly identifies *ḍād* as a lateral.

Table 2. Parallels in the categorisation of Arabic consonants in traditional Arabic linguistics and Quranic recitation are clearly apparent in online teaching materials for Islamic Studies.[1]

Sibawayh (eighth to ninth century)		**Adapted from Mansour (2013)**		
Exits	**Phoneme sets**	**Exits**	مخارج الحروف	**Phoneme sets**
Back of throat, furthest from mouth	ء هـ ا	Guttural	حلق *ḥalq* *throat*	ء هـ ع ح غ خ
Middle of throat	ع ح			
Upper region of throat nearest to mouth	غ خ			
Back part of tongue and upper palate region (with ق further back than ك)	ق ك	Uvular/Velar	لهاة *lahā* *uvula/velum*	ق ك
Area between middle parts of tongue and upper palate	ج ش ي	Palatal	شجر *shajar* *palate*	ض ج ش ي
Between edge of tongue and molars	ض			
Area between lowest edge and mid tip of tongue and upper palate	ل	Apico-alveolar (tongue blade)	ذلق *dhulq* *tongue blade*	ل ن ر
Area between tongue tip and upper palate	ن			
Similar to ن but further in towards underside of tongue	ر			

[1] Many thanks to James Dickins in Arabic, Islamic and Middle Eastern Studies at the University of Leeds for his patience and help in translating the Arabic terms in Table 2.

Sibawayh (eighth to ninth century)		Adapted from Mansour (2013)		
Exits	**Phoneme sets**	**Exits**	مخارج الحروف	**Phoneme sets**
Area between tip of tongue and roots of middle incisors	ط د ت	Alveolar/Dental	نطع *niṭ*ʿ *hard palate*	ط د ت
Area between tip of tongue and just above middle incisors	ص س ز	Apico-dental (tongue tip)	أسل *asal* *tongue tip*	ص س ز
Area between tip of tongue and tips of middle incisors	ظ ذ ث	Interdental	لثة *litha* *gum*	ظ ذ ث
Area between inside of lower lip and tips of upper incisors	ف	Labial	شفة *shafa* *lip*	ب و م ف
Area between lips	ب و م			

4.2 Manner of Articulation or the Characteristics of the Letters: ṣifāt al-ḥurūf

On the aforementioned *tajwīd* website, manner of articulation is meticulously described in traditional terminology. Thus we have the distinction, familiar from Sibawayh, between *mahmūsa/majhūra* or breathed/unbreathed (Heselwood and Maghrebi 2015); and *al-shidda/al-rikhāwa* or *intensity/laxity* (Bohas et al. 1990). This categorization is corroborated in a reputable textbook on *tajwīd* recitation, this time written for native English speakers (Surty 2000: 67–8); and both sources attest additional concepts and groupings from traditional Arabic linguistics. Interestingly, both sources assign missing letters from the *al-shidda/al-rikhāwa* contrast to an intermediate set of *moderate* letters.

Tajwīd systematizes the classes of *manner of articulation* rigorously and meticulously, but the terminology used is not immediately transparent for modern Phonetics. We make a novel contribution to the field in Table 3 by offering a practical translation for each of the twenty traditional *tajwīd* categories via concepts familiar to modern phoneticians as well as giving a literal translation.

Table 3. Common *tajwīd* terms appearing in reputable teaching materials, and their literal and technical translations.[2]

***Tajwīd* terms in the form of *abstract nouns*: (Mansour 2013)**	**Romanized forms (mostly *adjectival* and mostly from Surty 2000)**	**Literal translation of Arabic *tajwīd* terms (with some reference to Surty 2000)**	**Equivalent phoneme sets and concepts in modern phonetics**
الخفاء	*al-khafā'*	Invisibility	The letters /h/, /w/ and /j/ plus the long vowels: /a:/ and /u:/
اللين	*al-līn*	Softness	Arabic diphthongs: /aw/, /aj/
الانفتاح	*al-munfatiḥah*	Openness	All the *plain* consonants/ letters
الاستفال	*al-mustafilah*	Lowness	All consonants which are *not* members of the set {/sˤ/, /dˤ/, /tˤ/, /ðˤ/, /ʕ/, /ɣ/, /q/} where the back of the tongue is raised
الرخاوة	*al-rikhāwa*	Laxity	Fricatives
الهمس	*al-mahmūsah*	Breath running freely	Aspiration, i.e. produced with an open glottis
التوسط	*al-mutawaṣṣiṭ*	Standing between	Sonorants plus /ʕ/
الإذلاق	*al-mudhlaqah*	Light pronunciation	Sonorants plus labials /b/, /f/
الإصمات	*al-muṣmatah*	Hard pronunciation	All consonants/letters except /b/, /r/, /f/, /l/, /m/, /n/
الغنة	*al-ghunnah*	Nasalization	Nasalized sounds
التكرير	*al-takrīr*	Repetition	Trilling: /r/
الإستطالة	*al-istiṭālah*	Elongation	The consonant /dˤ/
التفشي	*al-tafashshā*	Distribution	The consonant /ʃ/
الانحراف	*al-inḥirāf*	Lateral and swerving movement	Liquids: /r/, /l/
الصفير	*al-ṣafīr*	Sibilance	Sibilants
القلقلة	*al-qalqalah*	Vibration	'qalqalah'

[2] Thanks are due to Khaled Mansour in Islamic Studies at the University of Jordan, and Janet Watson in Linguistics and Phonetics at the University of Leeds for their help in defining the *tajwīd* terms in Table 3.

Tajwīd **terms in the form of** *abstract nouns*: **(Mansour 2013)**	**Romanized forms (mostly** *adjectival* **and mostly from Surty 2000)**	**Literal translation of Arabic** *tajwīd* **terms (with some reference to Surty 2000)**	**Equivalent phoneme sets and concepts in modern phonetics**
الإطباق	*al-muṭbaqah*	Emphasis	The emphatic consonants
الاستعلاء	*al-musta'liyyah*	Raising	All consonants in the set: {/sˁ/, /dˁ/, /tˁ/, /ðˁ/, /ʕ/, /ɣ/, /q/} where the back of the tongue is raised
الشدّة	*al-shadīdah*	Intensity	Plosives
الجهر	*al-majhūrah*	Clear pronunciation	Unbreathed, i.e. produced with a narrowed, vibrating or closed glottis

4.3. Tajwīd *and Visual Speech*

Categorization of Arabic phonemes for state-of-the-art *visual* speech recognition and synthesis systems, involving photo-realistic models or talking heads, is also reminiscent of traditional groupings and their preoccupation with sound *exits*. In such systems, phonemes are visualized and differentiated via trademark lip shapes and jaw movements, though it is recognized that several phonemes can map to a single viseme (Pelachaud 2002; Al-Salmi in preparation). Currently, there is no standard set of lip shapes or visemes (i.e. visual phonemes) for any language (Damien et al. 2009); but methodologies for classifying visemes focus on different degrees of displacement from a neutral or still position for a given set of macroparameters (Lande and Quaglia 2003). Displacements such as mouth aperture (i.e. vertical distance between lips and also mouth width) and lip protrusion are re-cast as units of facial animation to be applied to the model. More recently, improvements in photorealism during synthesised speech have been achieved through better modelling of internal articulators and tongue dynamics, involving manual fine-tuning of tongue visemes based on a corpus of single speaker articulatory data (Dey 2012). Table 4 displays similarities for place of articulation in received wisdom and as determined via parameter measurements for MSA visual speech (Damien 2011; 2009). There is substantial agreement here; and discrepancies arise largely because visemes focus on facial deformation: the *appearance* of each sound as it is produced. Hence, labials {ب و م

ف} are distributed over 3 discrete viseme sets to reflect labio-dental articulation for {ف}, bilabial articulation for {ب م} and lip protrusion for {و}. Similarly, the inclusion of *ḍād* in the set {ض ل ن ر ط د ت} is probably due to similar measurements for the 'dark zone' parameter used in the study to detect phonemes produced via contact between the tongue and the teeth. This measures the inner mouth area between the tongue and the top incisors/upper lip which is in shadow. Other discrepancies arise from theoretical underpinning of the viseme set: the classification of { غ خ ك} as velar, for example.

Table 4. Parallels in classification of Arabic phonemes in terms of place of articulation in traditional Arabic linguistics, *tajwīd* studies, and visual speech applications.

Sibawayh (eighth–ninth century)	**Mansour (2013)**	**Damien (2011)**
Traditional Phoneme sets	***Tajwīd* Phoneme sets**	**Viseme Sets**
ء هـ ا	ء هـ ع ح غ خ	ء هـ ع ح ق
ع ح		
غ خ		
ق ك	ق ك	غ خ ك
ض	ض ج ش ي	ي
ج ش ي		ج ش
ل ن ر	ل ن ر	ل ن ر ط د ت ض
ط د ت	ط د ت	
ص س ز	ص س ز	ص س ز
ظ ذ ث	ظ ذ ث	ظ ذ ث
ف	ب و م ف	ف
ب و م		ب م
		و

5. Arabic Transcription, Transliteration and Romanization

An early objective in our aforementioned EPSRC-funded *Working Together* project is to automate phonemic and pausal Arabic-IPA transcription, using fully vocalized MSA text for rule development. This transcription will focus on pronunciation but not allophonic variations in pronunciation. We envisage that the output of our

algorithm will be a phonemic transcription for each Arabic word as an element of its citation form, similar to entries in the OALD and LDOCE for English, to enhance Arabic dictionaries, to facilitate Arabic language learning, and for natural language engineering applications involving speech recognition and speech synthesis. This task matches the definition of *transcription*, as opposed to *transliteration* and *romanization* for Arabic, in Beesley (1997), Habash et al. (2007), and the National Geospatial-Intelligence Agency or NGA: '…transcription [is] the process of recording the phonological and/or morphological elements of a language in terms of a specific writing system…' (NGA, 2013). Our task also presupposes fully vowelized Arabic text, as in Quranic and Classical Arabic. MSA is only partially vowelized in its written form, since it is not customary to include short vowel and other diacritic marks. Hence we are developing the mapping algorithm on our Boundary-Annotated Qur'ān Corpus — henceforward BAQ — (Brierley et al. 2012), which includes the entire text of the Qur'ān in fully vowelized MSA as well as the traditional *Othmani* script. This corpus has multiple linguistic annotation tiers, notably: Arabic words mapped to two different sets of coarse-grained syntactic annotations, plus binary and tripartite prosodic-syntactic boundary annotations. The Qur'ān is an iconic text and exemplifies *al-lugha al-fusḥā* اللغة الفصحى (most eloquent or refined speech): the best that Arabic can do, creatively and stylistically, as a medium of expression. Hence, we can justify its use as a gold standard for computational modelling and evaluation in Arabic Natural Language Processing (NLP) in that it subsumes other forms of Arabic, including MSA (Sharaf 2012), and constitutes a referent for the language and a common element among all the regional dialects (Harrag and Mohamadi 2007).

5.1 Arabic Transliteration

There is some confusion even in academic papers over terminology (i.e. transcription versus transliteration, see discussion in Heselwood 2013: 29–31) when recasting traditional Arabic script for computational processing via a typical western keyboard and text editor, and/or for intelligibility and readability for non-Arabic speakers (Beesley, 1997; Halpern 2007); and this is not helped by the proliferation of Arabic transliteration and romanization schemes. The most commonly used Arabic transliteration scheme for inputting, storing and displaying Arabic text is Buckwalter's (2002), developed at Xerox in the 1990s. This is a strict transliteration in that it adheres to

conventions in the orthography of the language but replaces Arabic characters with symbols from the American Standard Code for Information Interchange (ASCII) in a fully reversible, one-to-one mapping. Beesley (1997) makes the point that the notion of transliteration actually applies to machine-readable text for any language, and that there is a genuine need for exact transliteration in commercial Arabic NLP systems. Buckwalter's scheme has been adopted and extended by Habash et al. (2007) for the standard textbook on Arabic NLP, and by Dukes (2013) in the Quranic Arabic Corpus website.

5.2 Romanization for Arabic

Romanization, as the name suggests, substitutes standardized Roman-script spellings for languages with alternative scripts like Arabic, or for Latin alphabets that contain special characters like Turkish (NGA 2013), and has traditionally been used to visualize Arabic and clarify pronunciation in bi-lingual dictionaries such as Wehr's (2008). However, romanization does not eradicate the need for diacritics to differentiate manner and/or place of articulation for sounds that might otherwise be mapped to the same symbol, for example t and ṭ to represent the unvoiced and emphatic consonants ت and ط in Arabic. Furthermore, there are numerous different approaches, with individual journals, textbooks, research studies (e.g. Hassan and Heselwood 2011) and institutions each carrying their own romanization scheme for Arabic. Two such standards are: DIN 31635 as used in Table 4 in this paper, and BGN/PCGN 1956 from the US Board of Geographic Names (1946) and the UK Permanent Committee of Geographical Names (1956) for systematic romanization of place names in all regions of the Arab world based on fully pointed MSA. The former is used in the *Encyclopedia of Arabic Language and Linguistics* (Edzard et al. 2013), and the latter is still the approved scheme of the US National Geospatial Intelligence Agency or NGA (2013). There is also the Deutschen Morgenländischen Gesellschaft (DMG-Umschrift) system.

6. The MSA > IPA mapping

In this section, we present a comprehensive mapping from the graphemes of MSA to IPA as an essential step in algorithm development for phoneticizing Arabic. We also summarize our procedure for verifying the completeness of this mapping via automatic collection and

manual inspection of frequency distributions of letters and/or diacritics (unigrams) and longer consonant-vowel (CV) patterns (n-grams). Readers are referred to our related papers (Sawalha et al. 2014; Sawalha et al. forthcoming) for a more detailed account of this.

We are interested in automatic generation of phonemic pronunciation forms in a standard alphabet for use in Arabic learner dictionaries, and have selected a carefully researched subset of IPA labels for Arabic. In general, Arabic spelling is a phonemic system with one-to-one letter to sound correspondence; but there are notable exceptions to deal with. Difficult cases during processing and transcription of Quranic Arabic are: the definite article; the letters *alif*, *wāw*, and *yā'*; the variant forms of *hamza*; the *tanwīn* case mark; and words with special pronunciations. For MSA texts, further complexity arises when transcribing relative nouns; *hamzat al-waṣl*; and borrowed or Arabized words.

Table 1 gives two challenging examples from Quranic Arabic and two from our MSA sample; we will deal with these in order.

1. When transcribing يَتَسَاءَلُونَ *yatasā'lūna* (*they are asking one another*), the one-to-one mapping algorithm is not able to differentiate between the two behaviours of *alif* and *wāw*, and treats them as consonants. Therefore, additional rules must be applied to discover and render instances of *alif* and *wāw* (and similarly *yā'*) behaving as vowels. Thus *yatasā'lūna* is initially transcribed as /jatas**aa**ːʔal**uw**na/and then correctly rendered as /jatas**a**ːʔal**u**ːna/, with *alif* and *wāw* transcribed as /aː/ and /uː/ respectively.
2. When transcribing كَانُوا *kānū* (*they were*), we encounter an *alif of separation,* appearing after the masculine plural third person suffixed pronoun و. The latter is pronounced as **/uː/**, while the trailing *ʿalif* is not pronounced. Thus the Arabic word كَانُوا is initially transcribed as /k**aa**ːn**uwa**ː/ and then corrected via a specific rule to /ka**ː**nu**ː**/.
3. In الاتِّصَالَات *al-'itiṣālāti* (the communications), we encounter *hamzat al-waṣl* preceded by the definite article, and in this instance, it is pronounced as the short vowel *kasra* **/i/**. Othmani script distinguishes between *hamzat al-waṣl*, written as ٱ, and ا (ordinary *alif*), but this is not the case with MSA. Since we are working with MSA, the one-to-one mapping algorithm cannot differentiate between *hamzat al-waṣl* and ordinary *alif* and yields the erroneous transcription **/aː/**. The Arabic word الاتِّصَالَات is initially transcribed as /**ʔala**ːttisˁaːlaːti/, and then righted via a one-step rule to produce the correct transcription /**ʔali**ttisˁaːlaːti/.

4. Finally, borrowed or Arabized words are treated as special cases and stored in a look-up dictionary used by the Arabic > IPA transcription algorithm. The word كوزْمُوبُولِيْتَانْ *kuzmubulitan* (cosmopolitan) is written using Arabic characters, with the sounds /a/, /u/ and /i/ represented by long vowels: ي , و ,ا. The one-to-one transcription maps these long vowels into /**a:**/, /**u:**/, /**i:**/ respectively and produces the transcription /k**u:**zm**u:**b**u:**li:ta:n/. The correct transcription for this word /k**u**zm**u**b**u**lit**a**n/ is stored in the dictionary.

As demonstrated, automating the mapping process is non-trivial. Hence our challenge has been to design a mapping that reduces the number of letter-to-sound rules; and the originality of our scheme lies in our 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. The same distinction applies to our scheme versus DIN 31635, though the latter does achieve a neat and economical one-to-one grapheme-phoneme mapping for all consonants via selective use of pointing. One-to-one grapheme-to-phoneme transcription is available for German, Dutch and English via the DISC transcription sets used in CELEX (Baayen 1996) and this aids computational processing. We have considered and abandoned the idea of adapting the DISC set for Arabic because all ASCII symbols have already been exhausted in representing the range of vowel sounds in these European languages, and it was not considered good practice to re-allocate keyboard characters to Arabic phonemes under the same scheme.

6.1 Unigram Mapping of Arabic Letters to IPA Symbols: Consonant Segments

Table 5 presents our final mapping from grapheme to IPA for Arabic consonants arranged in a standard alphabetic sequence (Brustad et al. 2010), with equivalent symbols from the JSS, DIN 31635, and SAMPA sets cross-referenced for comparison. In addition to the IPA transcriptions, raw frequencies for each letter from our BAQ dataset appear in the rightmost column. The dataset comprises 691020 characters in total, distributed over 43 unigram characters (Sawalha et al. forthcoming). Shaded rows identify cases where there are one-to-many grapheme-phoneme transcriptions available depending on how

the sound is categorized in terms of place and/or manner of articulation (*cf.* the SAMPA transcriptions for the letters ع and ج), and where our IPA symbol selections result in the following:

- inclusion of {ظ ط ض ص ع ح} in the set of pharyngealized/pharyngeal sounds;
- inclusion of {غ خ } in the set of velar sounds;
- the letter {ج} realized as an affricate;
- the letter {ر} realized as a tap (Kalaldeh 2013), however it is often a trill [r] when geminated, and we prefer the symbol /r/ for reasons of typographical harmony as is the practice in phonemic transcription of English despite the most common realisation being the approximant [ɹ] in British English and [ɻ] in American English.
- omission of any velarized form for the letter {ل} since this only applies to the word *allah* and its morphological variants, and appropriate usage can easily be implemented via a simple rule.

Readers will note that we have taken one liberty in transcribing {خ} with a small /x/ instead of /χ/, and {غ } with /ɣ/ instead of /ʁ/; this is because we have judged them to harmonize better with the rest of the roman-based symbols and should not be taken to mean that we are asserting that these phonemes are velar rather than uvular. Similarly, use of diacritics indicating pharyngealization in our IPA transcriptions for the emphatic consonants {ظ ط ض ص} is not intended to exclude the possibility that the secondary articulation may be uvularization (see e.g. Zawaydeh 1997). Unicode hexadecimal numbers specifying each IPA symbol used in our scheme are recorded in our parallel paper (Sawalha et al. forthcoming). Finally, readers should note that the glottal stop /ʔ/ is discussed in more detail in Section 6.6.

Table 5. Unigram mapping of Arabic letters to IPA and other alphabets, with raw frequencies for each Arabic grapheme in the Qur'ān.

Arabic	JSS	DIN 31635	SAMPA	IPA symbol selection	BAQ frequencies
ا	ā	ā	a:	aː	45047
ب	b	b	b	b	12766
ت	t	t	t	t	11219
ث	th	t̲	T	θ	1427
ج	j	ǧ	g or Z	ʤ	3434
ح	ḥ	ḥ	X\	ħ	4150

Arabic	JSS	DIN 31635	SAMPA	IPA symbol selection	BAQ frequencies
خ	kh	ḫ	x	x	2543
د	d	d	d	d	6398
ذ	dh	ḏ	D	ð	5216
ر	r	r	r	r	12822
ز	z	z	z	z	1716
س	s	s	s	s	6099
ش	sh	š	S	ʃ	2181
ص	ṣ	ṣ	s\	sˤ	2133
ض	ḍ	ḍ	d\	dˤ	1726
ط	ṭ	ṭ	t\	tˤ	1319
ظ	ẓ	ẓ	D	ðˤ	862
ع	ʿ	ʿ	?or ?\	ʕ	9412
غ	gh	ġ	G	ɣ	1221
ف	f	f	f	f	8896
ق	q	q	q	q	7339
ك	k	k	k	k	10691
ل	l	l	l	l	41546
م	m	m	m	m	27786
ن	n	n	n	n	30530
ه	h	h	h	h	14890
و	w	w	w	w	25042
ي	y	y	j	j	22992
ء	’	’	?	ʔ	3089

6.2 Unigram Mapping of Arabic Letters to IPA Symbols: Vowel Segments

Table 6 presents our final mapping from orthographic character to IPA symbol for short and long vowels in MSA. We do not include contextual allophonic variants; and MSA diphthongs are presented in a later section (§6.4), which outlines our automated method of verifying their compound graphemic representations. Readers will note the many-to-one mapping for the long vowel /aː/ in Table 6:

when ا occurs at the end of a word, it is usually written as ى which is a spelling variant dating back to the recension of the Qur'ān (Brustad et al. 2010: 212).

Table 6. Unigram mapping of Arabic vowels to IPA and other alphabets, with raw frequencies for each Arabic grapheme in the Qur'ān.

Arabic	JSS	DIN 31635	SAMPA	IPA	BAQ frequencies
◌َ	a	a	a	a	121429
◌ِ	i	i	i	i	45962
◌ُ	u	u	u	u	57431
ا	ā	ā	a:	a:	45047
ى	ā	ā	a:	a:	2592
ي	ī	ī	i:	i:	22992
و	ū	ū	u:	u:	25042

6.3 Transcribing Other Diacritic Marks

Certain schemes (e.g. BGN/PCGN 1956) consider other diacritic marks. The *sukūn,* سكون does not need to be transcribed but is useful for syllabification and word stress: word internal letters carrying this diacritic signify syllable boundaries, unless *sukūn* occurs as part of a diphthong (cf. Table 7), and the syllable carrying *sukūn* will attract primary stress. Similarly, the *shadda* شدة needs no direct transcription: for our algorithm, we will follow the usual practice of doubling the letter for geminates. The nunation symbol *tanwīn* تنوين is a case marker for indefinite nouns and has been transcribed in each of its three forms, depending on its associated short vowel. In contextual (i.e. non-pausal) transcription (§7), the *tanwīn* signifies that the vowel itself should be pronounced and followed by the sound /n/. The suffix *tā' marbūṭa,* تاء مربوطة is a grammatical marker for feminine nouns and adjectives and only appears at the end of a word. It is pronounced as /t/ before *tanwīn* or short vowels and otherwise as /h/; it has therefore been assigned two different transcriptions. For example, *tā' marbūṭa* of the word *laʿibra,* لَعِبْرَةً (*surely* [*is*] *a lesson*) is pronounced as /h/ if it is read independently of its context, namely: /laʕibɾah/. However, *tā' marbūṭa* of the same word is pronounced as /t/ when it is read in context, as part of the sentence in which it appears: إِنَّ فِي ذَٰلِكَ لَعِبْرَةً لِّمَن يَخْشَىٰ *'inna fi dhālika laʿibratan liman yakhshā* (*In this there is a lesson for those who have fear of God*) namely:

/ʔinna fiː ðaːlika laʕibratan liman jaxʃaː/ (Qur'ān 79:26). Certain diacritics and glyphs, namely ◌ٰ ٱ (*dagger alif* and *hamzat al-waṣl*) are usually associated with the classical *Othmani* script. While the former will not be considered in this paper, *hamzat al-waṣl* as a *concept* is present in the mapping and will be transcribed as /ʔi/, /ʔu/, or /ʔa/ (Sawalha et al. forthcoming), since our automatic transcription algorithm must distinguish between *'alif, hamzat al-waṣl,* and *hamzat al-qaṭiʿ* (ordinary *hamza*). Finally, no transcription is given for the ligature ﻻ (*lā*) since it represents two independent letters (*lam* and *alif*), irrespective of its compositional appearance on Arabic keyboards.

Table 7. Unigram and/or one-to-many mapping of other Arabic diacritics and glyphs to IPA for contextual rule development.

Diacritics and glyphs	**Arabic term** (if applicable)	**Transcription** (if applicable)		
◌ْ	سكون	-		
◌ّ	شدة	-		
◌ً	تنوين فتح	an		aː
◌ٍ	تنوين كسر	in		-
◌ٌ	تنوين ضم	un		-
ة	تاء مربوطة	t		h
ا / ٱ	همزة الوصل	ʔi	ʔu	ʔa

6.4 Extensions for N-grams Mapped to IPA Symbols

Our objective is a comprehensive grapheme-phoneme mapping that reduces the need for hand-crafted rules and text pre-processing before automating phonemic transcription of Arabic. Therefore, we need a dictionary of mapped MSA > IPA pairs that anticipates and documents grapheme-phoneme relationships extending beyond a single letter to the immediate right-left context, assuming fully vowelized MSA as input text. For example, Arabic has two diphthongs (similar to *night* and *now* in English) which are each realized orthographically via the trigram sequence VCV, where V represents a short vowel or other diacritic mark and C is a consonant or semi-vowel. Moreover, long vowels frequently occur in text immediately preceded by their (largely redundant) superscript form as diacritic. To capture a number of common and not-so-common *non-unigram* events, and to resolve them prior to automated transcription in our dictionary, we have used the SALMA toolkit (Sawalha 2011; Sawalha and Atwell

2010) to analyse our BAQ corpus and to generate different n-gram frequency lists based on letters and diacritics rather than words (Sawalha et al. forthcoming). The program outputs: (i) unigram frequencies for each independent consonant and diacritic mark; (ii) bigram counts for *consonant + single diacritic* pairs; (iii) two different trigram formations to capture immediate right-left contexts: the first pattern being VCV *(previous diacritic + (consonant + diacritic))*, and the second pattern being CVC *((consonant + diacritic) + ensuing consonant)*; (iv) four-gram patterns of CVCV *((consonant + diacritic) + (consonant + diacritic))*; and even 5-grams VCVCV *(previous diacritic + (consonant + diacritic) + (consonant + diacritic))*. These n-gram counts have been manually inspected to verify and extend our MSA > IPA mapping. They have been useful in capturing orthographic representations of diphthongs and long vowels in Table 8, and they have been essential in generating the multi-faceted orthographic representations of *hamza* (thus resolving their IPA transcriptions) in Table 9. Though not represented in the Arabic orthography for long vowels and geminates in Table 8, readers should assume a trailing short vowel congruent with the n-gram capture listed (i.e. VCV or VCVCV), and should also note that BAQ frequencies represent total counts, subsuming all trailing short vowel variants associated with a given cluster and found in the corpus.

Table 8. Many-to-one and many-to-many mapping of Arabic graphemic events to IPA.

Arabic	Type	BAQ frequencies	N-gram capture	IPA
◌َيْ	diphthong	3972	trigram: VCV	aj
◌َوْ	diphthong	2605	trigram: VCV	aw
◌َا	long vowel	25472	trigram: VCV	a:
◌ُو	long vowel	10075	trigram: VCV	u:
◌ِي	long vowel	9917	trigram: VCV	i:
◌َى	long vowel	2499	trigram: VCV	a:
◌ِيّ	geminate	325	5-gram: VCVCV	ijj | i:
◌ُوّ	geminate	71	5-gram: VCVCV	uww

6.5 The SALMA Toolkit and SALMA Tokenizer

The SALMA Toolkit is a collection of open-source standards (the SALMA Tagset), tools (the SALMA Tagger) and resources (the

SALMA ABCLexicon) that widen the scope of Arabic word structure analysis — particularly morphological analysis — to process Arabic text corpora of different domains, formats and genres, for both vowelized and non-vowelized text. The modular approach adopted in developing the Tagger means that modules can be used independently to perform a specific task, or they can be used in sequence to produce full detailed morphological analyses of Arabic words.

For automated IPA transcription of Quranic Arabic rendered in fully pointed MSA, we have used a single class within the Tagger, namely the SALMA Tokenizer. This fulfils three main functions. The *tokenization* part deals with the input text files, determines what is considered an Arabic word, and stores the Arabic word in a unified format that enables the other components to deal with the word whether it is fully vowelized, partially vowelized or non-vowelized. The *spelling errors detection* and *correction* part checks the spelling of the tokenized words, and corrects the spelling if the word letters do not match certain patterns. The *word segmentation* part is responsible for generating all possible variant morpheme tokenizations of the analysed word, and mainly depends on matching the affixes and clitics of the analysed word with comprehensive lists of affixes and clitics within the Toolkit.

The Tokenizer program uses a regular expression tokenizer from the Natural Language ToolKit (Bird et al. 2009), plus regular expression patterns customized for Arabic, to tokenize input text into Arabic words, punctuation marks, currency tokens, numbers, words written in the Roman alphabet, and HTML/XML tags. The tokenizer then processes the extracted Arabic words by resolving any doubled letters *al-ḥurūf al-muḍaʿʿafa* الحروف المضعَّفة and any extensions *al-madd* المدّ. Any letter marked with *shadda* is replaced by a pair of consonants where the first is unreleased and marked with *sukūn*, and the second is vowelized with the same short vowel that appears on the original letter. This substitution can best be illustrated using Buckwalter's one-to-one mapping between Arabic letters and diacritics and their equivalent ASCII characters. For example, the word *waṣṣā,* وَصَّى (*he enjoined*) has the doubled letter ص and is transliterated as: waS~aY in Buckwalter. After SALMA processing it will take the form وَصْصَى waSoSaY where the ص has been doubled and the initial ص is marked with *sukūn*. The extension آ (*al-madd*) is replaced by two characters, *hamza* and *alif*, as in the word *āmanū*, آمنوا — transliterated as |manuwA in Buckwalter — (*they believed*) which becomes ءامنوا 'AmanuwA.

A constraint imposed by the Tokenizer is that only one short vowel can be associated with any letter of the word. Based on this fact, each Arabic word is decomposed into a list of (consonant + diacritic) tuples within a nested data structure of chapters, verses, and words. Each tuple stores the letter in the first position and the short vowel (if it is present) in the second position — and so on for all the letters and short vowels in the word. This unified format is represented as [(C,V), (C,V), …,(C,V)], where C represents a consonant and V represents a short vowel or other diacritic mark. Table 9 shows the data structure storing the words وَصَّى *waṣṣā* waS~aY and آمنوا *'amanu* |manuwA

Table 9. CV decomposition and tokenization in SALMA's word data structure.

Position	**0**		**1**		**2**		**3**		**4**		**5**	
وَصْصَى	و	َ	ص	ْ	ص	َ	ى	None				
waSoSaY	w	a	S	o	S	a	Y	None				
ءامَنُوا	ء	None	ا	None	م	َ	ن	ُ	و	None	ا	None
'AmanuwA	'	None	A	None	m	a	n	u	w	None	A	None

6.6 Transcribing Hamza

As mentioned, *hamza* has many different shapes or spellings in Arabic script depending on its vocalic context (Habash 2007). Table 10 is a comprehensive mapping of IPA symbols and sequences for all *hamza* variants appearing in our Qur'ān corpus. Readers should note that firstly, despite appearances, the forms {ؤ ئ إ أ ء} are all collected as *unigrams* by the SALMA Tagger (Sawalha et al. forthcoming) and are transcribed identically as /ʔ/. Secondly, readers should note that all further variants in Table 10 are collected as bigrams and this is mostly apparent from their IPA transcriptions. Cases where it is not apparent involve bigram instances displaying the *tanwīn* diacritic: these are transcribed as sequences of three IPA symbols (e.g. ٱً transcribed as /ʔun/). Thirdly, all variant *hamza* combinations are presented in descending order of raw counts for the same in our corpus. Finally, readers should note Table 11 as addendum to Table 10: this includes *al-madd* or lengthening, which denotes *hamza* followed by *alif* at the beginning or in the middle of a word {ءا}. It also includes infrequent but problematic grapheme sequences involving *hamza* and collected as higher order n-grams.

Table 10. SALMA analysis generates 'ready-made' IPA transcriptions for instances of *hamza* in the Qur'ān.

Hamza[b] unigram	*Hamza*[b] bigram variants	IPA	Frequency	*Hamza*[b] unigram	*Hamza*[b] bigram variants	IPA	Frequency
أ /ʔ/ 9119	أَ	ʔa	7725	ئ /ʔ/ 2811	ئِ	ʔi	768
	أُ	ʔu	878		ئْ	ʔ	121
	أْ	ʔ	509		ئَ	ʔa	110
	أً	ʔan	4		ئُ	ʔu	90
	أٌ	ʔun	3		ئ	ʔ	88
إ /ʔ/ 8015	إِ	ʔi	5099		ئٍ	ʔin	5
	إٍ	ʔin	9	ؤ /ʔ/ 376	ؤْ	ʔ	524
ء /ʔ/ 9803	ء	ʔ	1520		ؤُ	ʔu	110
	ءَ	ʔa	597		ؤَ	ʔa	33
	ءُ	ʔu	327		ؤ	ʔ	3
	ء	ʔi	283		ؤٌ	ʔun	2
	ءِ	ʔin	231		ؤِ	ʔi	1
	ءً	ʔan	78				
	ءٌ	ʔun	53				

Table 11. Three real (and one notional) graphemic event(s) in the Qur'ān corpus which are computationally challenging for automated transcription.

اء	ʔa:	1520
ْيِإ	ʔi:	75
ْوُأ	ʔu:	353
ْوُؤ	ʔu:	0

7. Implementing the Mapping: Gold Standard Transcription of Quranic Arabic

We have incorporated the MSA > IPA mapping presented in this paper in a mapping algorithm for automated phonemic transcription of Classical/Quranic Arabic and fully vowelized MSA (Sawalha et al. 2014; Sawalha et al. forthcoming); gold standard output

transcriptions are sampled in Table 12 for verses 19-21 in Chapter 78 of the Qur'ān: *The Tidings*, سورة النبأ. Supplementary information in this table covers the following: (i) partial but informative morpho-syntactic analysis of each Arabic word from a selection of outputs from the SALMA Tagger (Sawalha 2011); (ii) an interlinear translation for each Arabic word from the *Quranic Arabic Corpus* (Dukes 2013) — but see section 7.2; (iii) a syllable count for each Arabic word.

The transcriptions deviate slightly from a strictly pausal approach for Quranic text so as to convey its mesmerizing rhythm. Thus, to preserve naturalistic link-up (cf. Mortimer 1985) between verse-internal words, we have retained and transcribed suffixes such as *tā' marbūṭa* (e.g. /wasujji'rati/ in this extract) plus short vowel case endings for all save verse terminal items. This has a knock-on effect on syllable counts. For verse terminals, we have opted for a sense of finality at the end of each verse so the final syllable in these transcriptions ends on a consonant. This takes account of the nunation symbol or *tanwīn* over *alif* as case mark for indefinite nouns (e.g. /sa'ra:ban/ in this extract). Our gold standard transcription of the entire text of the Qur'ān, augmented with *tanwīn* and case endings, is published in our new version of the *Boundary Annotated Qur'ān Dataset for Machine Learning* (Sawalha et al. 2014; Brierley et al. 2012).

Table 12. Implementing the Arabic-IPA Mapping in Automatic Phonemic Transcription of Quranic Text

MSA	SALMA ANALYSIS	IPA	Interlinear	ID	Syllable Count
وَفُتِحَتِ	perfect verb; passive	wafuti'ħati	and is opened	78.19	5
السَّمَاءُ	generic noun; nom; fem; sing	ʔassa'ma:ʔu	the sky		4
فَكَانَتْ	perfect verb; active	fa'ka:nat	and (it) becomes		3
أَبْوَاباً	generic noun; acc; masc; broken plural	ʔab'wa:ban	gateways		3
وَسُيِّرَتِ	perfect verb; passive	wasujji'rati	and are moved	78.20	5
الْجِبَالُ	generic noun; nom; masc; broken plural	ʔalʤi'ba:lu	the mountains		4
فَكَانَتْ	perfect verb; active	fa'ka:nat	and (it) becomes		3
سَرَاباً	generic noun; acc; fem; sing	sa'ra:ban	a mirage		3

MSA	SALMA ANALYSIS	IPA	Interlinear	ID	Syllable Count
إِنَّ	subjunctive-governing particle	'ʔinna	indeed	78.21	2
جَهَنَّمَ	generic noun; acc; fem; sing	ʤa'hannama	hell		4
كَانَتْ	perfect verb; active	'ka:nat	is		2
مِرْصَاداً	noun of place; acc; fem; sing	mir'sˤa:dan	the way		3

7.1 Saj‘*-type Patterning in Quranic Verses*

Inspired by the work of Neuwirth (2013) and Stewart (1990), we are interested in text data mining the unique properties of Quranic *saj‘* or rhymed prose, using our gold standard, stressed and syllabified IPA transcription of the whole text of the Qur’ān. Phase one of this project, namely a pausal transcription, augmented with *tanwīn* and case endings, is published in a new version of the *Boundary Annotated Qur’ān Dataset for Machine Learning* (Sawalha et al. 2014). Later phases will implement a contextual transcription with full *tajwīd* rules, governing coarticulation and highlighting salient words, for the widely used *Hafs* recitation style.

The sample in Table 12 demonstrates several qualities of *saj‘*: end rhyme and rhythmic iteration reinforce syntactic parallelism across short verses construed as intonational phrases. There is end rhyme in /ʔab’wa:ban/, / sa’ra:ban/, / mir’sˤa:dan/; verses 19–20 each comprise four words and are rhythmically identical in their sequence of syllable counts and the location of primary stresses; these verses have the same word order and syntactic structure. Verse 78.21 clings to the same end rhyme — /fa’ka:nat ʔab’wa:ban/ (78.19), /fa’ka:nat sa’ʕa:ban/ (78.20), /’ka:nat mir’sˤa:dan/ (78.21) — but the changes in rhythm (fewer syllables) and disruption in syntax (insertion of an extra particle: إِنَّ *indeed*) seem to match the semantic content of the verse: not the sún, but the *road to perdition* revealed through a break in the clouds.

7.2 Translating مِرْصَاداً

The interlinear translation for Qur’ān 78.21 is given as ‘…*indeed hell is lying in wait*…’ on the *Quranic Arabic Corpus* website (Dukes 2013). This follows mainstream translations for this verse where hell

is depicted as a place of *ambush* (cf. Pickthall, Yusuf Ali and Arberry available from the same website), but there is a difference in syntax: the predicate *lying in wait* is a departure from the original Arabic, where end rhyme is achieved because *mirṣādan* is a noun (i.e. noun of place) just like *abwāban* and *sarāban*. We looked up the meaning of *mirṣādan,* مِرْصَادًا by entering its root *r-ṣ-d* رصد in an online root meaning search tool (Sawalha and Atwell 2010) which enables user access to a database and corpus of 23 Arabic dictionaries spanning over 1200 years. This search provided our preferred translation for *mirṣādan,* مِرْصَادًا as: *the way (ahead)*.

8. Conclusions

In this paper, we have presented the comprehensive, applications-oriented grapheme-phoneme mapping that underpins our new Arabic transcription technology. This technology generates phonemic citation forms for Arabic words in the IPA as standard alphabet for use in Arabic learner dictionaries. We have used n-gram counts at different levels of granularity over our chosen corpus to ensure completeness, and to reduce the number of rules at an early stage in automated transcription by including computationally-challenging, non-unigram events in the grapheme-phoneme mapping dictionary. This makes our scheme novel and unique.

The mapping has been carefully researched, and this process has highlighted the importance of cross-referencing and reconciling conceptual inventories for the sound system of Arabic in three scholarly traditions: (i) *tajwīd* or the principles of correct Quranic recitation; (ii) the Arabic linguistics tradition, dating at least as far back as the eighth century; and (iii) modern Phonetics. This will also benefit new research and application areas such as visual speech.

We have made a novel contribution to Arabic linguistics by translating and reconciling concepts used in *tajwīd* studies in/with concepts in modern phonetics (Table 3). The technical terminology of *tajwīd* was invented by the early Arab grammarians such as Sibawayh to describe and analyse the Arabic language and its usage, initially (but not solely) in accepted styles of Quranic recitation exemplified by *qurrā'* (readers) trained in the oral tradition. This terminology is still current and resonates with fundamental concepts in modern phonetics such as place and manner of articulation. *Tajwīd* teachings also shed light on topical issues: diachronic studies of the phoneme *ḍād* and its variant pronunciations in modern Arabic dialects, some of

which preserve an older realisation of *ḍād* as a lateral fricative (Al-Azraqi 2010), the canonical pronunciation of *ḍād* in the most widely-used recitation style.

Another application of our MSA > IPA mapping, and one of the deliverables in our EPSRC-funded project, is stylistic and stylometric analysis of the Qur'ān. We plan to customize statistical techniques like keyword extraction to explore significant segmental features in chapters of the Qur'ān and a phonemic representation of the text will facilitate this. New language resources produced via algorithmic implementation of this mapping include an updated version of 77430-word *Boundary Annotated Qur'ān Dataset for Machine Learning* (Sawalha et al. 2014; Brierley et al. 2012) and a pronunciation guide for the recitative effect of *qalqalah* or vibration (Brierley et al. 2014). Both resources display Arabic words transcribed in IPA and mapped to their chapter-verse ID.

Address for correspondence: C.Brierley@leeds.ac.uk

REFERENCES

Al-Azraqi, M. 2010. 'The Ancient Ḍād in Southwest Saudi Arabia', *Arabica* 57, 57–67

Al-Khatib, S.E. 2003. 'The Formal Notation of Some Phonological Processes in the Holy Quran', *Journal of King Abdulaziz University: Educ. Sci.* 16. 25–37

Atwell, E.S., J. Dickins and C. Brierley. 2013. *Natural Language Processing Working Together with Arabic and Islamic Studies*. EPSRC: EP/K015206/1. Online. Accessed: 09.08.2013. http://gow.epsrc.ac.uk/NGBOViewGrant.aspx?GrantRef=EP/K015206/1

Baayen, R.H., R. Piepenbrock, and L. Gulikers. 1996. *CELEX-2*. (Philadelphia)

Beesley, K.R. 1997. *Romanization, Transcription and Transliteration*. Online. Accessed: 09.08.2013. http://open.xerox.com/Services/arabic-morphology/Pages/romanization

Bird, S., E. Loper and E. Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Bohas, G., J.P. Guillaume and D.E. Kouloughli. 1990. *The Arabic Linguistic Tradition*. (London)

Brierley, C. and E. Atwell 2010. 'Holy Smoke: Vocalic Precursors of Phrase Breaks in Milton's *Paradise Lost*', *Literary and Linguistic Computing* 25.1

Brierley, C., M. Sawalha and E. Atwell. 2012. 'Open-Source Boundary-Annotated Corpus for Arabic Speech and Language Processing', in *Proceedings of Language Resources and Evaluation Conference (LREC) 2012*. Istanbul. Online. Accessed: 09.08.2013. http://www.lrec-conf.org/proceedings/lrec2012/index.html

—— 2014. 'Tools for Arabic Natural Language Processing: A Case Study in *qalalah* Prosody'. To appear in *Proceedings of LREC* 2014. (Reykjavik)

Brustad, K., M. Al-Batal and A. Al-Tonsi. 2010. *Alif Baa: Introduction to Arabic Letters and Sounds*.[3] (Washington)

Buckwalter, T. 2002. *Arabic Transliteration*. Online. Accessed: 01.11.2013. http://www.qamus.org/transliteration.htm

Cantineau, J. 1960. *Études de linguistique arabe*. (Paris)

Damien, P. 2011. 'Visual Speech Recognition of Modern Classic Arabic Language', *Proceedings of the International Symposium on Humanities, Science and Engineering Research (SHUSER) 2011*. Online. Accessed: 09.08.2013. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6008499

Damien, P., N. Wakim and M. Egéa. 2009. 'Phoneme-Viseme Mapping for Modern Classical Arabic Language', *Proceedings of the International Conference on Advances in Computational Tools for Engineering Applications* (ACTEA) 2009. 547–52

Dar-al-Maarifah. 2008. *Tajweed Qur'an with Meaning Translation in English*[3]. (Damascus)

Dey, P. 2012. *Visual Speech in Technology-Enhanced Learning*. Ph.D. Thesis, University of Sheffield

Dickins, J. 2013. Personal communication

—— 1999. *Literal (Lit) and idiomatic (Id) translations of Sibawayh's account of the 'letters' (حروف) of Arabic*. Unpublished manuscript

Dukes, K. 2013. *The Quranic Arabic Corpus*. Online. Accessed: 05.11.2013

Edzard, L. et al. (eds) 2013. *Encyclopedia of Arabic Language and Linguistics Online*. Accessed: 05.11.2013

Habash, N., A. Soudi and T. Buckwalter. 2007. 'On Arabic Transliteration', in A.Soudi, A. van den Bosch and G. Neumann (eds), *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. e-Book

Halpern, J. 2007. 'The Challenges and Pitfalls of Arabic Romanization and Arabization', *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Language*

Harrag, A. and T. Mohamadi. 2010. 'QSDAS: New Quranic Speech Database for Arabic Speaker Recognition', *The Arabian Journal for Science and Engineering*. 35, 2C. 7–19

Heselwood, B. 1996. 'Glottal States and Emphasis in Baghdadi and Cairene Arabic: Synchronic and Diachronic Aspects', in J. Dickin (ed.) *Three Topics in Arabic Phonology* (Centre for Middle Eastern and Islamic Studies Occasional Paper 53, Durham)

—— 2013. *Phonetic Transcription in Theory and Practice*. (Edinburgh)

Heselwood, B. and Z.M. Hassan. 2011. 'Introduction', in Z.M. Hassan and B. Heselwood (eds), *Instrumental Studies in Arabic Phonetics*. (Amsterdam)

Heselwood, B. and R. Maghrabi. 2015. 'An Instrumental-Phonetic Justification for Sībawayh's Classification of *ṭā', qāf* and *hamza* as *majhūr* Consonants', *JSS* 15:1, 131–75

Kalaldeh, R.A. 2013. Email correspondence: Re: P.S. to Draft Mapping: Arabic > IPA. 01.05.2013

Khraish, P.S.C.M. 2007. 'Arabic Phonetic Alphabet', in *Proceedings of 9th United Nations Conference on the Standardization of Geographical Names*. (New York)

Lande, C. and M. Quaglia, 2003. *Method of and Apparatus for Animation, Driven by an Audio Signal, of a Synthesized Model of a Human Face*. Patent: US 6665643 B1. Online. Accessed: 07.05.2013. http://patents.justia.com/patent/6665643

Mansour, M.K. Faculty of Sharia, University of Jordan. 2013. Personal communication

Mortimer, C. 1985. *Elements of Pronunciation*. (Cambridge)
National Geospatial-Intelligence Agency (NGA). *2013. Romanization Systems and Policies*. Online. Accessed: 09.08.2013. http://earth-info.nga.mil/gns/html/romanization.html
Neuwirth, A. 2013. 'Form and Structure of the Qur'ān', in McAuliffe (ed.) *Encyclopedia of the Qur'ān*. Online. Accessed: 26.11.2013
Newman, D. 2002. 'The Phonetic Status of Arabic within the World's Languages: The Uniqueness of the *lughat al-daad*', *Antwerp Papers in Linguistics* 100, 65–75
—— 2005. 'Contrastive Analysis of the Segments of French and Arabic', in A. Elgibali (ed.), *Investigating Arabic: Current Parameters in Analysis and Learning*. (Leiden)
Newman, D. and J. Verhoeven, 2002. 'Frequency Analysis of Arabic vowels in Connected Speech', *Antwerp Papers in Linguistics* 100, 77–86. Online. Accessed: 09.08.2013. http://dro.dur.ac.uk/4419/1/35758.pdf
Pelachaud, C. 2002. 'Visual Text-to-Speech', in I.S. Pandzic and R. Forchheimer, (eds), *MPEG-4 Facial Animation: The Standard, Implementation And Applications*. (Chichester)
Philips, A.A.B. 1987. *The Qur'an's Numerical Miracle: Hoax and Heresy*. Jeddah. Online. Accessed: 09.08.2013. http://islamic-replies.ucoz.com/2/1987_Qurans_Numerical_Miracle_19_Hoax_Heresy.pdf
Sawalha, M. 2011. 'Open-Source Resources and Standards for Arabic Word Structure Analysis: Fine Grained Morphological Analysis of Arabic Text Corpora'. Ph.D. Thesis, University of Leeds
Sawalha, M. and E. Atwell. 2010. 'Fine-Grain Morphological Analyzer and Part-of-Speech Tagger for Arabic Text', *Proceedings of LREC 2010*. (Valetta)
Sawalha, M., C. Brierley and E. Atwell. (Forthcoming). 'IPA Transcription Technology for Classical and Modern Standard Arabic'
—— 2014. 'Automatically Generated, Phonemic Arabic-IPA Pronunciation Tiers for the *Boundary Annotated Qur'ān Dataset for Machine Learning* (version 2.0)'. Submitted to the *Workshop on Language Resources and Evaluation for Religious Texts* at *LREC* 2014
Scott, M. 2012. *WordSmith Tools* version 6. (Liverpool)
Semaan, K.I. 1968. *Linguistics in the Middle Ages: Phonetic Studies in Early Islam*. (Leiden)
Shah, M. 2003a. 'Exploring the Genesis of Early Arabic Linguistic Thought: Qur'anic Readers and Grammarians of the Kūfan Tradition (Part I)'/تطور الدراسات اللغوية بين القراء والنحاة الكوفيين (القسم الأول), *Journal of Qur'anic Studies* 5:1. 47–78
—— 2003a. 'Exploring the Genesis of Early Arabic Linguistic Thought: Qur'anic Readers and Grammarians of the Baṣran Tradition (Part II)'/تطور الدراسات اللغوية بين القراء والنحاة البصريين (القسم الثاني), *Journal of Qur'anic Studies* 5:2. 1–47
Sharaf, A. 2012. 'Annotation of Conceptual Co-reference and Text Mining the Qur'an'. Ph.D. Thesis, University of Leeds
Shukri, A. et al. 2013. *al-munīr fī 'aḥkām at-tajwīd*[8]. المنير في أحكام التجويد (Amman)
Stewart, D.J. 1990. 'Saj' in the Qur'ān: Prosody and Structure', *Journal of Arabic Literature*. 21:2, 101–39
Surty, M.I. 2000. *A Course in 'Ilm Al-Tajwid: The Science of Reciting The Qur'an*. (Leicester)

Watson, J. and M. Al-Azraqi. 2011. 'Lateral Fricatives and Lateral Emphatics in Southern Saudi Arabia and Mehri', *Proceedings of the Seminar for Arabian Studies.* 41: 425–32

Wehr, H. 2008. *A Dictionary of Modern Written Arabic: Arabic - English*. Edited by J. Milton-Cowan, J. (Beirut)

Wells, J.C. 2002. *SAMPA for Arabic*. Online. Accessed: 25.04.2013. http://www.phon.ucl.ac.uk/home/sampa/arabic.htm

Zawaydeh, B.A. 1997. 'An Acoustic Analysis of Uvularization Spread in Ammani-Jordanian Arabic', *Studies in the Linguistic Sciences* 24, 185–200