Tools for Arabic Natural Language Processing: a case study in qalqalah prosody

Claire Brierley1, Majdi Sawalha2, Eric Atwell1

1 School of Computing, University of Leeds, LS2 9JT, UK
2 Computer Information Systems Dept., King Abdullah II School of IT, University of Jordan, Amman 11942, Jordan

E-mail: C.Brierley@leeds.ac.uk, sawalha.majdi@gmail.com, E.S.Atwell@leeds.ac.uk

Abstract

In this paper, we focus on the prosodic effect of qalqalah or “vibration” applied to a subset of Arabic consonants under certain constraints during correct Qur'anic recitation or taǧwīd, using our Boundary-Annotated Qur'an dataset of 77,430 words (Brierley et al 2012; Sawalha et al 2014). These qalqalah events are rule-governed and are signified orthographically in the Arabic script. Hence they can be given abstract definition in the form of regular expressions and thus located and collected automatically. High frequency qalqalah content words are also found to be statistically significant discriminators or keywords when comparing Meccan and Medinan chapters in the Qur’an using a state-of-the-art Visual Analytics toolkit: Semantic Pathways. Thus we hypothesise that qalqalah prosody is one way of highlighting salient items in the text. Finally, we implement Arabic transcription technology (Brierley et al under review; Sawalha et al forthcoming) to create a qalqalah pronunciation guide where each word is transcribed phonetically in IPA and mapped to its chapter-verse ID. This is funded research under the EPSRC “Working Together” theme.

Keywords: Qur’anic recitation; qalqalah prosody; regular expressions

1. Introduction

The theory and practice of taǧwīd or correct recitation of the Qur’an has developed over time to help believers achieve clearly articulated recitation. We have discussed an important aspect of this in previous LREC papers (Brierley et al 2012; Sawalha et al 2012), namely: fine-grained annotation of prosodic boundaries or stops and starts mark-up (waqf wa ibtida‘). In this paper, we focus on the prosodic effect of qalqalah or “vibration” applied to a subset of Arabic consonants, namely: ق ط د ج ب (q t d j b) under certain constraints during taǧwīd recitation. These constraints can be expressed algorithmically, so we present software for locating all instances of qalqalah in the text of the Qur’an, using our purpose-built dataset of 77,430 words: the Boundary-Annnotated Qur'an Dataset for Machine Learning (Brierley et al 2012). We have also generated a qalqalah frequency list, further sorted into raw frequencies for three different “strengths” or categories of qalqalah. Next, we have divided the Qur’an dataset into Meccan versus Medinan chapters, and used the Semantic Pathways toolkit (cf. Brierley et al 2013) and keyword extraction techniques to identify statistically significant high frequency qalqalah content words in each sub-corpus. Finally, we provide a qalqalah pronunciation guide with each word mapped to an automatically-generated, canonical, IPA1 transcription, plus its chapter-verse ID. This is EPSRC-funded research under the “Working Together” theme. Our project is entitled: Natural Language Processing Working Together with Arabic and Islamic Studies and runs for two years, from 2013 to 2015.

2. Taǧwīd theory and practice: types of qalqalah

The taǧwīd website quran1.net specifies three types or degrees of qalqalah: كبيرة [kubrā], very strong; صغيرة [saɡīrah], weak. Rules for applying qalqalah assume a prosodic boundary or pause immediately after the word carrying the qalqalah letter. Hence qalqalah letters are described as sākinah, silent: they are either marked with ُ or sukūn, or treated as such in pre-pausal position, where any trailing case endings (signified by short vowel diacritics) will not be pronounced. This especially applies to qalqalah letters which occur at the end of a verse. The website readwithitajweed.com identifies a strong qalqalah effect at the end of Qur’an 112.1 (cf. Fig. 1). Here, the final consonant in ُاَهَدُ (aḥad(in)) is ُdāl (one of the qalqalah set), and it carries a tānwi` diacritic, which categorises the word as an indefinite noun via the case ending –un. However, the word ِاَهَدُ is also verse terminal, and therefore, irrespective of its transliterated form ِاَهَدُ, it will be truncated and pronounced as ُءاَهَدُ, with a bouncing qalqalah effect on the letter ُdāl if the reader interprets verse endings as compulsory stops (cf. Brierley et al 2012). Readers are referred to Section 6 of this paper, plus our parallel paper (Sawalha et al 2014) for a summarised account of the automated IPA transcriptions for Arabic that appear in Fig. 1.2

| Arabic | Romanized | IPA | Meaning |
|---|---|---|---|
| قُلْ | qul | /qul/ | Say He is (is) Allah the One |
| هُوَ | huwa | /huwa/ | (is) Allah |
| ِلَا هَوَاء | l-lahu | /lalla:hu/ | ِءاَهَدُ |
| اَهَدُ | aḥadun | /aḥadun/ | /aḥad(in) |

Figure 1: Arabic words in Qur’an 112.1 transcribed in roman characters and IPA symbols, with a word-for-word translation

2.1 Contextual rules for qalqalah

Rules for applying different intensities كلبي [al-kubrā,
very strong) (al-kabīrah, strong) (al-saġīrah, weak) for qalqalah during recitation are well defined. If the qalqalah letter occurs as a geminate at the end of a word in pre-pausal and/or verse-terminal position, then that is qalqalah kabīrah. A similar rule identifies qalqalah saģīrah except the letter will not be geminate, and may not appear at the end of a verse. For qalqalah saģīrah the letter is word-internal but carries the sukūn diacritic and hence indicates a syllable boundary. Contextual examples of all three events are presented in Fig. 2. However, readers should note that application of any of these rules depends on whether or not the reciter chooses to realise a prosodic boundary or pause on/after the word in question. Without that boundary/pause, the qalqalah effect will be negligible, even though qalqalah is a permanent attribute of these letters: they are always majhūrah (unbreathed or voiced) and Ǧādīdah (intense or explosive), an ancient classification dating back over 1200 years and preserved in taqīd studies today (quran1.net; readwithtajweed.com).

| Intensity | Letter | Verse | ID |
|-----------|--------|-------|----|
| kabīrah   | بَطْشَ | نَبِيَّ ُبَشَرُ أَيُّهَا الْعَالَمُ | 111.1 |
| saģīrah  | طَبْشَ | ﴿بِتَرَبَّتْ بَيْنَ أَيْنَ أَلْحَابُ وَتَغَيَّرَ﴾ | 85.12 |

Figure 2: Orthographic signification of each qalqalah type: very strong; strong; weak

3. Qalqalah events algorithm

We have developed software for collecting all potential qalqalah events in the Qur’an. Input data is the entire text of the Qur’an rendered in fully vowelised Modern Standard Arabic from our Boundary Annotated Qur’an dataset for machine learning (Brierley et al. 2012). In a parallel paper for LREC (Sawalha et al. 2014), we present an updated version of this corpus with Arabic words mapped to their canonical pronunciation form, and example transcriptions are presented in Section 5 of the current paper. The qalqalah events algorithm first builds a Qur’an data structure of chapters, verses, and words, and then operates over this nested list to output two separate lists of verse-terminal and in-verse qalqalah sites for each letter in the qalqalah set {بَطْشَ} in the form of verse strings tagged with their chapter + verse ID. For an illustrative example, the commented Python and NLTK code in Table 1 specifies an RE for locating an instance of qalqalah saģīrah in a list of three Qur’anic verses, where each verse is in turn a list of word tokens, with each token appearing as a unicode string (e.g. بَطْشَ for bāṭša, the grip). The program actually returns this same word because it is the only match for qalqalah saģīrah in the input text of three Qur’anic verses from Fig. 2. The word بَطْشَ, bāṭša, the grip occurs in Qur’an 85.12. This verse contains another instance of qalqalah in the final word لَنَشْدَدَنَّ, (is) surely strong. However, this is not retrieved in Table 1 because the RE pattern applies to word-internal (not word-terminal) qalqalah. The RE in question operates over each verse string to determine whether each Arabic consonant (i.e. letter) belongs to the qalqalah set, is associated with sukūn, and is word-internal.

```
# -*- coding: utf-8 -*-
import codecs, nltk, re
from nltk.tokenize import *
tokenizer1=WhitespaceTokenizer()
data=codecs.open('arabicRE.txt','r','utf-8').readlines()
data = {tokenizer1.tokenize(index) for index in data}
data[0][0] = u'عو625\u650\u664\u656\u664\u644\u670' # get rid of unwanted \ufeef at beginning of string

#check if the word contains a qalqalah saģīrah (qalalah letter + sukun) in the middle of the word
p2= u''"\\u6261-\\u6552"*\\u6264,\\u657e,\\u6268,\\u626c,\\u6262,\\u626f\\u6552{\\u6261-\\u6552}+''

for verse in data:
  for word in verse:
    qalqalah_S=re.match(p2,word)
    if qalqalah_S:
      print word
```

Table 1: Regular expression search for one qalqalah type within a sample of Qur’anic verses

4. Frequencies for qalqalah types

Figures 3 to 5 in Section 4.1 show the top ten most frequent words for each qalqalah type: kabīrah, kabīrah, saģīrah. These have been obtained via searches over the entire Qur’an data structure of chapters, verses, and words for patterns matching an RE specification of each type such as the example given in Table 1. We have then created an instance of the FreqDist() Class from NLTK’s probability module for each qalqalah iteration, and obtained word counts via an fdist.items() method call which returns a list of words sorted in decreasing order of

http://fromkarachi.wordpress.com/2007/02/17/lesson-3-al-qalqalah-the-echo/
frequency (Table 2).

from nltk.probability import FreqDist
fdist = FreqDist(word for word in sağirah)
inspect = Fdist.items()
for index in inspect[:10]: print index[0],
    "\".join(index[1]),

Table 2: Generating raw frequencies for each qalqalah type in Python and NLT

4.1 Traditional Arabic parts-of-speech

Words in Figs. 3-5 are part-of-speech (POS) tagged very simply as nouns, verbs, or particles {N, V, P}.

| Count | Arabic word | POS | English meaning |
|-------|-------------|-----|----------------|
| 99    | رَبّ   | N   | lord            |
| 74    | يَبْنُ    | N   | in-truth        |
| 48    | أَمَّلاَمَقي | N   | the-truth       |
| 39    | يَبْنُ    | V   | love(s)         |
| 37    | أَمَّلاَمَقي | N   | the-truth       |
| 24    | أَمَّلاَمَقي | N   | the-truth       |
| 23    | يَبْنُ    | N   | lord            |
| 18    | أَمَّلاَمَقي | N   | stronger/mightier |
| 14    | أَمَّلاَمَقي | N   | right/due/truth |

Figure 3: Top ten most frequent kabrā word types

This sparse tripartite scheme follows traditional Arabic grammar (Wright, 1996; Ryding, 2005; Al-Ghalayini, 2005), and informs one of the syntactic annotations tiers in our source data: the Boundary Annotated Qur’an corpus (Brierley et al 2012).

| Count | Arabic word | POS | English meaning |
|-------|-------------|-----|----------------|
| 124   | وَلَدَّ     | P   | and-certainly  |
| 120   | تَمَّ     | P   | certainly/indeed |
| 98    | عَندَ     | N   | with/near/at    |
| 82    | بَعْدَ     | N   | after           |
| 78    | الكِتَابِ  | N   | the-book        |
| 77    | الكِتَابِ  | N   | the-book        |
| 77    | غَدَّ     | N   | a-punishment    |
| 58    | حَلَقَ     | V   | (has)-created   |
| 54    | بَعْدَ     | N   | after           |

Figure 4: Top ten most frequent kabīrah word types

As well as respecting traditional linguistic wisdom, this {N, V, P} scheme avoids the problem of mismatches between descriptive frameworks for Arabic and English (i.e. “Western”) grammar. For example, Arabic nouns subsume adjectives, adverbs, and some prepositions, while particles also subsume some prepositions, as well as conjunctions and negatives (Maamouri et al 2004). Hence the words after and before in Figs 4 and 5 are tagged as nouns because they are adverbs (of time).

| Count | Arabic word | POS | English meaning |
|-------|-------------|-----|----------------|
| 70    | نَبِيّ   | N   | before         |
| 48    | نَفَّر   | V   | flow(s)/sail(s) |
| 47    | إِبْرَاهِيمَ | N   | ibrahim        |
| 36    | بَعْدِهِ   | N   | before-them    |

Figure 5: Top ten most frequent sağirah word types

4.2 Arabic morphology: short vowel case endings

Readers will have noted the apparent repetitions in Figs 3 to 5 in the English translations of Arabic words which in turn have different frequencies but markedly similar orthography, differing only in their final short vowel diacritic. An example would be three word types for the-truth in Fig. 3. The final short vowel (damma; fathā; kasra) in each of these types denotes, respectively, the nominative, accusative and genitive case in Arabic: الدَمْمُ (الْكَبْرَا) (Fig.6).

| Count | Arabic word | POS | Case   | English meaning |
|-------|-------------|-----|--------|----------------|
| 48    | الدَمْمُ   | N   | nominative | the-truth |
| 37    | الدَمْمُ   | N   | genitive | the-truth |
| 24    | الدَمْمُ   | N   | accusative | the-truth |

Figure 6: Case endings

5. Exploring qalqalah prosody and keywords via the Semantic Pathways toolkit

One aspect of Qur’anic scholarship is stylistic comparison of Meccan versus Medinan chapters and verses to identify discriminatory features which can be used to determine the provenance of disputed chapters/verses (Sharaf 2011). This Mecca/Medina split lends itself to corpus comparison techniques from Corpus Linguistics. In a recent publication (Brierley et al 2013), we use the Semantic Pathways toolkit to visualize lexical differences in British versus American English, represented in the Lancaster-Oslo-Bergen (LOB) and Brown corpora respectively. Semantic Pathways implements keyword extraction and keyword-based document clustering for interactive information exploration and hypothesis-forming in the field of Visual Analytics. The initial corpus comparison appears as a collection-level gist comprising the ten most significant content words in the test set of documents with respect to (wrt) the reference set. Preliminary experiments in the Semantic Pathways command line interface on Meccan wrt Medinan Meccan chapters (and vice versa) uncover statistically significant qalqalah items. For example, the genitive form rabbi lord is one of the most frequent content words in the Qur’an. It is also positively key in Meccan versus Medinan sub-corpora in the Semantic Pathways collection-level gist. Similarly, another genitive form al-kitābi the-book, and غَدَّاَبَ a-punishment are statistically significant at document-level in the Medinan wrt Meccan comparison. Here, document-level
denotes the subset of documents, each represented by its top-ranking keyword as calculated by the log-likelihood statistic, in which the collection-level query term is also significant. Hence we might hypothesise that qalqlalah prosody is one way of highlighting salient items during recitation of the Qur’an.

6. Extracting a qalqlalah pronunciation guide from the Boundary-Annotated Qur’an Dataset for Machine Learning

Our parallel paper (Sawalha et al 2014) presents an updated version of our Boundary-Annotated Qur’an Dataset for Machine Learning (Brierley et al 2012), which includes two new prosodic and phonemic annotation tiers in the form of syllabified International Phonetic Alphabet (IPA) transcriptions for each Arabic word. These are based on our detailed mapping from Classical and Modern Standard Arabic to the IPA, which extends beyond one-to-one grapheme-phoneme correspondence as in SAMPA (Wells 2002), to capture and resolve compound orthographic events prior to automated transcription proper. A typical (though only moderately challenging) example would be the sequence of characters denoting the Arabic diphthong /aw/ as in مكتبة, namely: j-. The Arabic > IPA mapping and the mapping algorithm are both discussed in separate publications (Brierley et al under review; Sawalha et al forthcoming). The SALMA tagger used to capture frequencies of Arabic letters and diacritics at different orders of n-gram granularity, and thus verify the completeness of the mapping, is published in Sawalha (2011) and Sawalha and Atwell (2010).

6.1 Qalqlalah pronunciation guide

In Section 3 of this paper, we have presented software for gathering all qalqlalah sites in the Qur’an, first at verse level via the events algorithm, and then at word level via regular expressions. Results have been further subdivided into three qalqlalah types: kubra, kabira, saqiirah. Our source data is the Boundary-Annotated Qur’an (version 2.0), a user-friendly dataset for machine learning. Thus, we have been able to extract a qalqlalah pronunciation guide where qalqlalah words are mapped to their canonical IPA transcriptions and also tagged with their chapter-verse ID. This resource is open source and is suitable for both native and non-native Arabic speakers. Examples for each qalqlalah type are given in Fig.7.

![Figure 7: Example entries in the qalqlalah pronunciation guide](image).

| ID  | Qalqlalah type | Arabic word | IPA transcription |
|-----|----------------|-------------|------------------|
| 111.1 | kubra | مكتبة | /watabb/ |
| 112.1 | kabira | أحمد | /faHar/ |
| 85.12 | saqiira | بطن | /baJra/ |

7. Conclusions

We are interested in the prosodic effect of qalqlalah applied to a subset of Arabic consonants during correct Qur’anic recitation. Since this effect is rule-governed and signified orthographically, we have developed software for collecting all qalqlalah instances in the Qur’an, incorporating regular expression patterns which define each qalqlalah type. From this definitive list of events, we have generated a qalqlalah pronunciation guide with each item phonetically transcribed in IPA, utilising our state-of-the-art Arabic transcription technology. We have also found that high frequency qalqlalah content words are significant discriminators when determining Meccan/Medinan provenance of Qur’anic chapters, using state-of-the-art Visual Analytics technology. We therefore hypothesise that qalqlalah prosody is a salience marker, primarily in Qur’anic Arabic, and possibly in other varieties of Arabic, and will investigate this in future work. Qalqlalah is always latent in this subset of consonants, so their occurrence under certain constraints may subconsciously trigger connotations of significance for native Arabic speakers. This is research funded under the EPSRC “Working Together” theme. The events algorithm has been developed for our forthcoming, phonetics-based, inter-disciplinary study on the consonants of Modern South Arabian and Arabic (Watson and Al-Saqqaf 2014).

8. References

[Al-Ghalayyni. 2005. “جامع الدروس العربية "Jami’ Al-Duroos Al-Arabia" Saida - Lebanon: Al-Maktaba Al-Alsiyiah "المكتبة الالسيرة"].

Bird, S., Klein, E. and Loper, E. 2009. Natural Language Processing with Python. Sebastopol, CA. O’Reilly Media, Inc.

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

Brierley, C., Atwell, E., Rowland, C. and Anderson, J. 2013. Semantic Pathways: a novel visualization of varieties of English. In Journal of International Computer Archive of Modern and Medieval English (ICAME).

Brierley, C., Sawalha, M., Atwell, E. 2012. “Open-Source Boundary-Annotated Corpus for Arabic Speech and Language Processing.” In Proceedings of Language Resources and Evaluation Conference (LREC), Istanbul, Turkey.

Maamouri, M., Bies, A., Buckwalter, T. and Mekki, W. 2004. The Penn Arabic Treebank: Building a Large-Scale Annotated Corpus. Philadelphia. Linguistic Data Consortium.

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

Sawalha, M., Brierley, C. and Atwell, E. (Forthcoming). IPA transcription technology for Classical and Modern Standard Arabic.

Sawalha, M., Brierley, C. and Atwell, E. 2014. Automatically generated, phonemic Arabic-IPA pronunciation tiers for the Boundary Annotated Qur’an.
Dataset for Machine Learning (version 2.0). To appear in Proceedings of the 2nd. Workshop in Language Resources and Evaluation for Religious Texts, LREC 2014, Reykjavik.

Sawalha, M., Brierley, C., and Atwell, E. 2012. ‘Predicting Phrase Breaks in Classical and Modern Standard Arabic Text.’ In Proceedings of LREC 2012: Language Resources and Evaluation Conference, Istanbul, Turkey.

Sawalha, Majdi. 2011. Open-source Resources and Standards for Arabic Word Structure Analysis. Leeds: University of Leeds PhD.

Sawalha, M. and Atwell, E. 2010. ‘Fine-Grain Morphological Analyzer and Part-of-Speech Tagger for Arabic Text.’ In Proceedings of LREC’10: Language Resources and Evaluation Conference, Valetta, Malta.

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

Watson, J.C.E., and Al-Saqqaf, H. (2014 - under review). The Consonants of Modern South Arabian and Arabic. Project proposal submitted to the British Academy and under review.

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

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