The Qur'an Lexicon Project: A database of lexical statistics and phonotactic probabilities for 19,286 contextually and phonetically transcribed types in Qur'anic Arabic

Siti Syuhada BINTE FAIZAL, Ghada KHATTAB, Cristina MCKEAN

School of Education, Communication, and Language Sciences, Newcastle University, NE1 7RU, UK
E-mail: s.s.binte-faizal@ncl.ac.uk, ghada.khattab@ncl.ac.uk, cristina.mckean@ncl.ac.uk

ABSTRACT

Reciting and memorizing the Qur'an forms a major part of religious practice for 1.6 billion Muslims around the world; in non-Arabic-speaking Muslim communities, it also provides Muslim speakers of other languages with their first exposure to the Arabic script and language. However, little research has been completed regarding the psycholinguistic processing of Qur'anic Arabic. In this paper, we present the first psycholinguistic database for Qur'anic Arabic, which comprises lexical variables (length: character, syllable, phone; frequency: item, syllable, biphone, phone; lexical uniqueness point, orthographic and phonological neighbourhood sizes, and orthographic and phonological Levenshtein distances) as well as phonotactic probabilities (positional segment and biphone) for 19,286 types that we contextually and phonetically transcribed based on Qur'anic recitation. This open-source resource will be useful for researchers studying Qur'anic Arabic lexical and phonological processing as well as for making systematic cross-linguistic comparisons that allow better delineation of language-specific and language-general processes in language processing.

Keywords: Arabic, lexicon, corpus, lexical statistics, phonotactic probability.

1. INTRODUCTION

The Qur'an, written solely in Arabic, is the religious text of around 1.6 billion Muslims all over the world, of which a large proportion are non-Arabic speakers [5]. In many Muslim communities, especially in the Indo-Pak and South-east Asian regions, Qur'anic recitation and memorization constitutes a major component in the religious education of children, to the extent that parents send their children to schools and classes for the sole purpose of learning to read, recite, and/or memorize the Qur'an. It is thus unsurprising that for many Muslims, the first (and often only) exposure to the Arabic script and language is through the Qur'an. Despite this and the Qur'an's large user base, there have been only two published experiments on the effects of Qur'anic memorization on serial memory skills [8] and on the statistical learning of grammar [16], and none on the psycholinguistic processing of Qur'anic Arabic. A major impediment to the development of such research has been the lack of data regarding the psycholinguistic attributes of Qur'anic Arabic (e.g. word frequency, neighbourhood density, length) that are needed to support the design of empirical psycholinguistic studies.

In order to overcome the above limitation and develop a better understanding of the statistical patterns in the language one is exposed to via Qur'anic recitation and/or memorization, we compiled a database of lexical variables (character length, syllable length, phone length, item frequency, syllable frequency, biphone frequency, phone frequency, lexical uniqueness point, orthographic and phonological neighbourhood sizes, and orthographic and phonological Levenshtein distances) as well as phonotactic probabilities (positional segment and biphone) for 19,286 types in the Qur'an corpus that we contextually and phonetically transcribed based on Qur'anic recitation. This is the first psycholinguistic database for Qur'anic Arabic, which is a significant step forward from past Qur'anic projects such as the Tanzil project [14] and the Qur'anic Arabic Corpus [3], which served to provide a verified Qur'an text and an annotated Qur'an resource respectively.

2. METHOD

2.1. Development of the Qur'an Lexicon

We used the Qur'anic Arabic Corpus [3] that was built on the verified Arabic text of the Qur'an distributed by the Tanzil project [14]. In this corpus, 77,430 orthographic tokens had already been segmented following the whitespaces between them in the text. The corpus also had the position of each token in the text annotated by its *surah* (chapter) number, sentence number, and word position in the sentence. Each token also had its own Buckwalter transliteration that uses ASCII characters to represent Arabic orthography.

For the Qur'an lexicon, we scripted special rules to convert each token's Buckwalter transliteration into a contextual broad phonetic transcription that

takes into account co-articulatory effects in continuous Qur'anic recitation that are marked orthographically in the script. Pauses in the Qur'anic recitation are reflected in sentence endings and compulsory pause markers, which the transcription also takes into account. It is important to note that this corpus is unique in that all the words appear in a certain order and are recited in that order. Due to strict rules of recitation, or *tajweed*, the pronunciation of a word depends on the position of the word in a sentence as well as the word that precedes or follows it; thus context plays a huge role in the pronunciation of a word. This makes the Qur'an lexicon different from other lexicons that were created from corpora with words in isolation.

What this means is that the phonetic transcription in this corpus is not necessarily how one would read the word in isolation, but is based on how one would recite the word, taking into account the *tajweed* rules of recitation. For example, at the end of words, a long vowel ending is shortened when it is assimilated with a *sukun* (ْ) in the next word: e.g. فَلَا (Buckwalter transliteration: falaA; phonemic transcription: fa.laa; contextual phonetic transcription: fa.la) that is followed by ٱقْتَحَمَ (Buckwalter transliteration: {qotaHama; phonemic transcription: ʔɪq.ta.ħa.ma; contextual phonetic transcription: q.ta.ħa.ma.). Such contextual transcription ensures that the Qur'an corpus accurately reflects the characteristics of items as they are recited or heard by memorizers of the Qur'an.

Each token's contextual phonetic transcription was manually cross-checked with a professional *qari* (Qur'an reciter) recitation and verified by a proficient Qur'anic Arabic reader. Approximately 10% of the corpus was also manually checked and verified by a *hafidz* (someone who has memorized the entire Qur'an). The final corpus had 77,430 tokens, with 18,994 unique orthographic representations and 19,286 unique phonetic representations. It was these representations that were used to calculate all the lexical and phonotactic probability variables, rather than more traditional phonological variables adopted in the literature. This is because we did not seek to make any assumptions about the reciters' phonological representations, but rather plan to investigate their nature in future work.

2.2. Variables calculated to date for the Qur'an lexicon

2.2.1. Length

For length measures, number of characters, syllables, and phones are provided for each item. Diphthongs and geminates were treated as singular phones for the purpose of phone counts.

2.2.2. Frequency

An N-gram extraction tool [15] was used to compute the following frequencies in the Qur'an corpus: *item, syllable, biphone,* and *phone*. For *item frequency*, both raw and log-transformed counts were provided. For *syllable, biphone,* and *phone* frequencies, both overall and position-specific counts were provided.

Figure 1: Type and token counts for number of phones in the Qur'an lexicon

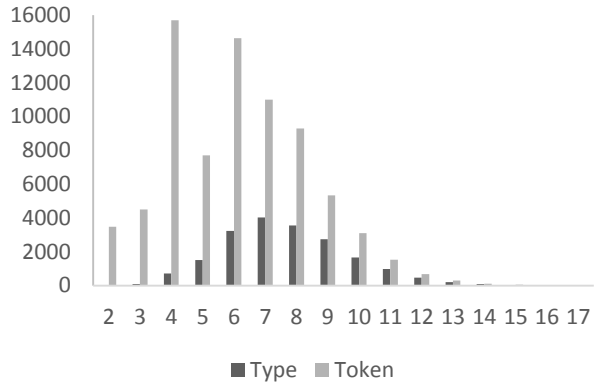
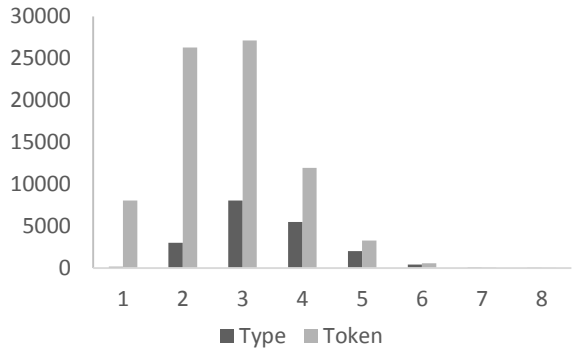


Figure 2: Type and token counts for number of syllables in the Qur'an lexicon



2.2.3. Lexical uniqueness point

This is defined as the point at which a set of phonemes or graphemes is no longer a subset of some other set of phonemes or graphemes [4]. The script for the lexical uniqueness point calculator for Hebrew [4] was modified to suit the Arabic script and the special characters used in our phonetic transcription. The lexical uniqueness point was then calculated for each item in the phonetic and orthographic Qur'an lexicons.

2.2.4. Neighbourhood size

Neighbourhood size measures were computed using LINGUA [9]. Orthographic neighbourhood density

(ON) is a measure of orthographic similarity referring to the number of words that can be obtained by changing a single letter in the target word, while holding the identity and positions of the other letters constant [1] [2].

Phonological neighbourhood density (PN) is the phonological analogue of orthographic neighbourhood density and reflects the number of words that can be obtained by changing a single phoneme in the target word while holding the other phonemes constant and preserving the identity and positions of the other phonemes [12] [13]. PN was computed using Qur'an Arabic contextual phonetic transcription.

Figure 3: Mean phonological Levenshtein distance (PLD20) and phonological N (PN) as a function of length.

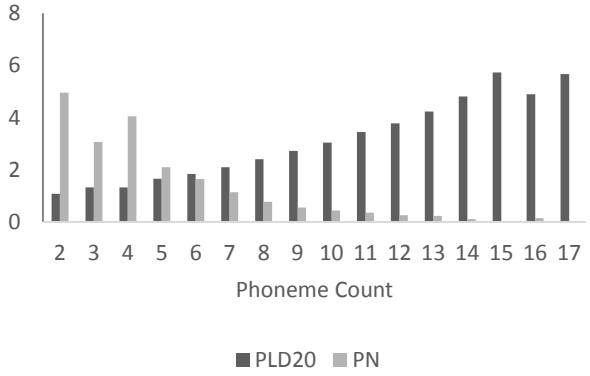
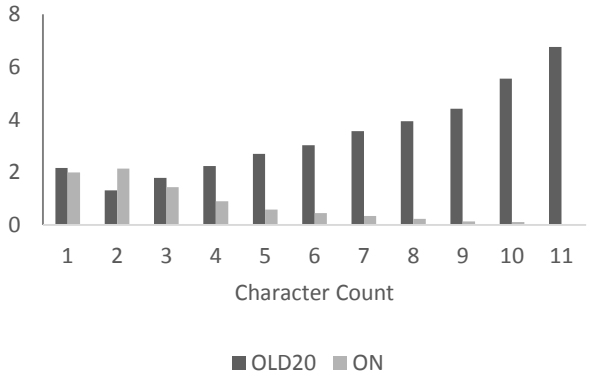


Figure 4: Mean orthographic Levenshtein distance (OLD20) and orthographic N (ON) as a function of length.



2.2.5. Levenshtein distance

Orthographic Levenshtein distance (OLD20) and phonological Levenshtein distance 20 (PLD20) were developed from a standard computer science metric of string similarity defined as the number of insertions, deletions, and substitutions needed to generate a string of elements, such as letters or phonemes, from another [11]. In order to create usable metrics of orthographic and phonological similarity, orthographic and phonological Levenshtein distances were first calculated between every word and every other word in the Qur'an Lexicon. OLD20 and PLD20 represent the mean orthographic and phonological Levenshtein distances, respectively, from a word to its 20 closest neighbors. Like phonological N, PLD20 was computed using Qur'an Arabic contextual phonetic transcription.

The Levenshtein measures have been shown by Yarkoni et al. [11] to circumvent many limitations that are linked to traditional neighborhood measures such as orthographic N, to the extent of being more powerful predictors of word recognition performance in English (see [10] and [11]). For instance, the utility of OLD20 and PLD20 as a measure of similarity or distinctiveness extends to words of all lengths and especially to long words, wherein the utility of orthographic N and phonological N is limited, as most long words (e.g. television, intermission) have few or no orthographic and phonological neighbours. This is especially significant in Arabic, which is an agglutinative language and thus, has naturally longer words (see Figures 3 and 4). We would therefore recommend researchers to consider using O/PLD20 as neighbourhood measures instead of O/PN, especially when constructing Arabic stimuli, or to at least consider using both measures together.

2.2.6. Phonotactic probability

Following the work of Vitevich and colleagues [6] [7], two token-based measures of position-specific phonotactic probability were computed: positional segment and biphone. *Positional segment probability* was calculated by dividing the sum of log (10) frequencies of all the items in the lexicon that contain a given segment in a given position by the total log (10) frequency of all the items in the lexicon that have a segment in that position [6] [7]. Log-values of the frequency counts were used as they better reflect the distribution of frequency of occurrence and better correlate with performance than with raw frequency counts [7]. For each item in the Qur'an lexicon, we then computed the *positional segment sum* (adding the positional segment probability for each sound in the target item) and *positional segment average* (dividing the positional segment sum by the number of sounds in the target item).

The biphone probability was computed in a similar manner, except that pairs of adjacent sounds were used in the calculations. *Biphone probability* was calculated by dividing the sum of log (10) frequencies of all the items in the lexicon that contain

a given pair of sounds in a given position by the total log (10) frequency of all the items in the lexicon that have a pair of sounds in that position [6] [7]. For each item in the Qur'an lexicon, we then computed the *biphone sum* (adding the positional segment probability for each sound in the target item) and *biphone average* (dividing the positional segment sum by the number of sounds in the target item).

Table 1: Descriptive lexical statistics and phonotactic probabilities in the Qur'an Lexicon

	M	*SD*	Min	Max
Item Freq	4.02	23.66	1.00	1264.00
Log (Item Freq)	0.45	0.29	0.30	3.10
Syllable Count	3.41	1.00	1.00	8.00
Phoneme Count	7.71	2.04	2.00	17.00
Character Count	5.27	1.45	1.00	11.00
OLD20	2.77	1.11	1.00	9.40
PLD20	2.37	0.91	1.00	10.00
ON	0.66	1.03	0.00	8.00
PN	1.13	1.59	0.00	18.00
Uniqueness Point	6.28	1.84	2.00	15.00
PosSegAv	0.12	0.04	0.00	0.35
PosSegSum	0.90	0.41	0.01	3.18
BiPhonAv	0.02	0.01	0.00	0.09
BiPhonSum	0.12	0.10	0.00	1.45

3. CONCLUSION

To summarize, we have generated and provided measures of frequency, length, orthographic and phonological similarity, and phonotactic probabilities for a set of 19,286 'phonetic' types that are based on an overt contextual phonetic transcription which is unique to Qur'anic recitation. To our knowledge, this represents the first such psycholinguistic database for Qur'anic Arabic, a language used by over a billion people. This resource, which will be made freely available, should be useful for researchers studying Qur'anic Arabic lexical and phonological processing. More generally, it will also be useful to researchers who are interested in making systematic cross-linguistic comparisons that allow better delineation of language-specific and language-general processes in language processing.

4. REFERENCES

[1] Coltheart, M., Davelaar, E., Jonasson, J., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535-555). Hillsdale, NJ: Erlbaum.

[2] Davis, C. J. (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic characteristics. *Beh. Res. Met., 37*, 65-70.

[3] Dukes, K. (2009). *Quranic Arabic Corpus.* Retrieved 31 January 2015 from http://corpus.quran.com

[4] Francom, J., Woudstra, D., & Ussishkin, A. (2009). *Creating a Web-based Lexical Corpus and Information-extraction Tools for the Semitic Language Maltese.* Soc. Esp. para el Proc. del Lenguaje Natural, San Sebastian, Spain.

[5] Pew Research Center. (2011). *The Future of the Global Muslim Population.* Retrieved 31 January 2015 from http://www.pewforum.org/2011/01/27/the-future-of-the-global-muslim-population/

[6] Storkel, H. L., & Hoover, J. R. (2010). An online calculator to compute phonotactic probability and neighbourhood density on the basis of child corpora of spoken American English. *Beh. Res. Met., 42*(2), 497-506.

[7] Vitevitch, M. S., & Luce, P. A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Beh. Res. Met, Instruments, & Computers, 36*(3), 481-487.

[8] Wagner, D. A., & Spratt, J. E. (1987). Cognitive consequences of contrasting pedagogies: The effects of Quranic preschooling in Morocco. *Child Dev., 58*, 1207-1219.

[9] Westbury, C., Hollis, G. & Shaoul, C. (2007). LINGUA: The Language-Independent Neighbourhood Generator of the University of Alberta. *The Mental Lexicon*, *2*(2), 273-286.

[10] Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *J. of Mem. & Lang., 60*, 502-529.

[11] Yarkoni, T., Balota, D. A., & Yap, M. J. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psych. Bulletin & Review, 15*, 971-979.

[12] Yates, M. (2005). Phonological neighbors speed visual word processing: Evidence from multiple tasks. *JEP: Learning, Memory, & Cognition, 31*, 1385-1397.

[13] Yates, M., Locker, L., & Simpson, G. (2004). The influence of phonological neighborhood on visual word perception. *Psych. Bulletin & Review, 11*, 452-457.

[14] Zarabbi-Zadeh, H. (2008). *Tanzil Project.* Retrieved 31 January 2015 from http://tanzil.net/wiki/

[15] Zhang, L. *N-Gram Extraction Tools.* Retrieved 31 January 2015 from http://homepages.inf.ed.ac.uk/lzhang10/ngram.html

[16] Zuhurudeen, F. M., & Huang, Y. T. (2013). *Effects of Statistical Learning on the Acquisition of Grammatical Categories through Qur'anic Memorization: A Natural Experiment.* BUCLD.