Behav Res (2018) 50:313–322
DOI 10.3758/s13428-017-0872-z

A Web-based interface to calculate phonotactic probability for words and nonwords in Modern Standard Arabic

Faisal Aljasser[1] · **Michael S. Vitevitch**[2]

Published online: 24 March 2017

Abstract A number of databases (Storkel *Behavior Research Methods, 45*, 1159–1167, 2013) and online calculators (Vitevitch & Luce *Behavior Research Methods, Instruments, and Computers*, *36*, 481–487, 2004) have been developed to provide statistical information about various aspects of language, and these have proven to be invaluable assets to researchers, clinicians, and instructors in the language sciences. The number of such resources for English is quite large and continues to grow, whereas the number of such resources for other languages is much smaller. This article describes the development of a Web-based interface to calculate phonotactic probability in Modern Standard Arabic (MSA). A full description of how the calculator can be used is provided. It can be freely accessed at http://phonotactic.drupal.ku.edu/.

Keywords Phonotactic probability · Modern Standard Arabic · Online calculator

Phonotactic probability refers to the frequency with which phonological segments and sequences of phonological segments occur in the words of a given language (Vitevitch & Luce, 2005). Phonological segments or sequences of phonological segments that occur frequently in certain positions in a given language are said to be *high* in phonotactic probability, whereas less frequently occurring phonological segments and sequences of phonological segments are said to be *low* in phonotactic probability. A number of studies have shown that phonotactic probability has significant effects on how quickly and accurately spoken words are recognized, segmented from fluent speech, produced, and learned (for a brief review, see Vitevitch, Luce, Pisoni, & Auer, 1999, and Auer & Luce, 2005).

Sensitivity to the phonotactic patterns of one's native language develops as early as 9 months of age (Jusczyk & Luce, 1994) and may be used by infants to segment words from fluent speech (Mattys & Jusczyk, 2001). Evidence suggests that the learning of novel words in both children (Storkel, 2001) and adults (Storkel, Armbruster, & Hogan, 2006) is influenced by phonotactic probability.

In adulthood, knowledge of phonotactic patterns is observed in the subjective ratings of sequence typicality (Bailey & Hahn, 2001; Vitevitch, Luce, Charles-Luce, & Kemmerer, 1997). Adults may also use phonotactic information to help them segment words from fluent speech in their native language (Pitt & McQueen 1998) and in a second-learned language (Weber & Cutler, 2006). Furthermore, both spoken words and specially constructed nonwords that vary in phonotactic probability differ in how quickly and accurately they are recognized (Vitevitch & Luce, 1998, 1999, 2005; Vitevitch et al., 1997). Moreover, Goldrick and Larson (2008) observed that phonotactic patterns influence certain speech production processes, as well.

Researchers have continued to investigate how phonotactic probability influences language processing in various populations of speakers, including those who stutter (Anderson & Byrd, 2008), children with phonological delays (Storkel, 2004), and children with specific language impairment (Leonard, Davis, & Deevy, 2007). Some researchers have even begun to investigate how phonotactic probability might

✉ Michael S. Vitevitch
mvitevit@ku.edu

[1] Department of English Language and Translation, Qassim University, Buraydah, Saudi Arabia

[2] Spoken Language Laboratory, Department of Psychology, University of Kansas, 1415 Jayhawk Blvd., Lawrence, KS 66045, USA

influence other cognitive processes, such as decision making (e.g., Vitevitch & Donoso, 2012).

Note that all of the previously described research has considered how knowledge of the phonotactic patterns of *English* influenced processing in some way. Vitevitch, Chan, and Goldstein (2014) suggested that English could be viewed as a model language, just as certain animal species are viewed as model organisms in biological research. They cautioned, however, that

> there are several characteristics of English (spoken as a first language by around 335 million people worldwide) that may limit how broadly studies using this model language can generalize to the other 6000 or so languages found in a diverse range of language families (e.g., Indo-European, Sino-Tibetan, Afro-Asiatic, Austronesian) spoken by the other 6 billion or so people on the planet. (Vitevitch et al., 2014, p. 59)

Given the large number of other languages that are spoken on the planet, and the great variety found in those languages regarding phoneme inventories, morphological productivity, syntactic rules, and so forth, it is important to consider the extent to which certain well-documented effects in English might generalize to other languages (e.g., Vitevitch & Stamer, 2006). Regarding the influences of phonotactic probability on language processing, some work has begun in other languages, such as Dutch (e.g., Adriaans, 2006; Messer, Leseman, Boom, & Mayo, 2010; Rispens, Baker, & Duinmeijer, 2015; van der Kleij, Rispens, & Scheper, 2016). The primary factor that limits researchers' ability to examine the influences of phonotactic probability in other languages is the existence or availability of databases or online calculators to compute various linguistic measures of interest to language scientists (e.g., Marian, Bartolotti, Chabal, & Shook, 2012; Storkel, 2013). To help language scientists, clinicians, and educators explore other languages, we describe here a Web-based interface to compute phonotactic probability of the most widely used variety of Arabic, Modern Standard Arabic (MSA).

We decided to create this resource for MSA because it is the most widely spoken Semitic language. Consider that it is an official language in 27 countries (only English and French are spoken in more countries in the world), with over 290 million native speakers worldwide ("Arabic Language", Furman, Goldberg, & Lusin, 2007). Lewis, Simons, and Fennig (2016) list Arabic as the fourth global language in terms of the number of first-language speakers (after Chinese, Spanish, and English). Moreover, the number of Arabic-as-a-foreign-language classes has increased around the world at universities and language centers. Enrollment in Arabic classes in the United States, for example, dramatically increased by 126.5% between 2002 and 2006 (according to a Modern Language Association language report; Furman et al., 2007, cited in Palmer, 2008).

Another reason that Arabic is increasingly spoken around the world is that the Quran (the holy book of Muslims) is written in Arabic, so Arabic is widely learned as a foreign language in non-Arab Muslim countries and by other Muslims all over the world. However, whereas some Muslims will be learning Arabic as a foreign language to be able to speak it, others will learn only enough (i.e., the Arabic writing system) to enable them to read the Quran. Faizal, Khattab, and McKean (2015) recently developed a psycholinguistic database based on Quranic Arabic (i.e., from Quranic recitations). This freely available database offers phonotactic probability calculations (based on the method used in Vitevitch & Luce, 2004) of Quranic Arabic as one of its features. Although the database developed by Faizal et al. will be a valuable source for those interested in investigating Quranic Arabic, psycholinguistic research in Arabic may require a resource that is more representative of the language as it is naturally used by speakers. To this end, we have developed the present tool. Below, we provide a detailed description of the Arabic variety (i.e., MSA) that we used for the present database.

Modern Standard Arabic

Like *Chinese*, *Arabic* can be viewed as an umbrella term that encompasses several different language varieties. One variety of Arabic is Classical Arabic, whose use is now limited to some forms of literature or religious context. MSA, on the other hand, is the current standard form used for written and formal spoken communications. Although it follows much of, but not exactly, the same syntactic and phonological rules as Classical Arabic, the two varieties diverge lexically. MSA does not use many of the archaic words of Classical Arabic (e.g., *sajanjal* "mirror"). It continuously adds new words either through direct borrowing or loan translation, to keep up with the rapidly growing fields of technology and science (Ryding, 2005).

Another variety of Arabic is what is usually referred to as *colloquial* Arabic. This variety consists of the regional dialects that are used in certain geographic locations for informal spoken communications (e.g., the Egyptian and Moroccan dialects). Native speakers of Arabic will use both MSA and the local dialect in their proper contexts. In Saudi Arabia, for example, whereas students might use the local dialect to talk to each other in the university café, they will typically use MSA to give a presentation in class. This simultaneous use of MSA and the local dialect by Arabic speakers creates a situation of diglossia (see Ferguson, 1959, for a discussion). Explaining the phenomena of diglossia in Arabic, Ferguson posited two language varieties: the high variety (standard

Arabic) and the low variety (the regional dialect). However, this division should not suggest that the two varieties of Arabic are discrete. Empirical evidence suggests that spoken Arabic should be viewed as a continuum of varieties in which interaction and mixing between varieties takes place (e.g., Parkinson, 2003), but in which MSA plays a vital role.

Typically, children learning Arabic as their first language start their exposure to MSA at school (Holes, 1995), with some preschool exposure mainly through TV programs (e.g., animation). MSA is used in most Arab countries in public speaking and most other formal contexts. It is consistently used in the mass media in both written (e.g., newspapers) and spoken (radio, TV) forms.

However, we should emphasize here that the use of MSA and the local variety in different Arabic-speaking countries is not as homogeneous as it may sound. Arabic-speaking countries vary in the natures of the exact situations that call for using MSA instead of the local variety. Whereas in most Gulf countries a class presentation or a formal speech by a politician is typically given in MSA, the same may not be true in Morocco, Lebanon, or Egypt, where the local dialect could be used instead. For example, the late Egyptian president Jamal Abdul-Nasir was famous for delivering most of his public speeches using the Egyptian dialect. On the other hand, Saudi kings do typically deliver their speeches in MSA. This heterogeneous use of MSA, its learning context, and the varying levels of exposure to it in different Arabic-speaking countries make it difficult to assume that Arabic speakers from different Arabic-speaking countries have comparable levels of mastery of MSA. This is a point that researchers need to keep in mind when investigating MSA using participants from different dialectal backgrounds.

For example, conducting their study in Israel, Ibrahim and Aharon-Peretz (2005) compared the processing of Palestinian Arabic and literary Arabic (i.e., MSA). They concluded that Palestinian Arabic and MSA retain their statuses as first and second languages in the cognitive systems of adult native Arabic speakers. However, Boudelaa and Marslen-Wilson (2013) argued that the psycholinguistic parallel between the local dialect and MSA in a language community is reliant on the sociolinguistic context in which MSA is learned and used. They maintained that, in Israel, Hebrew typically plays the role that MSA does in other Arabic-speaking countries, and that is why, in this particular community, MSA is not expected to be processed as effectively as the dialectal Arabic.

Comparing the processing of roots and patterns in both MSA and Southern Tunisian Arabic (STA), Boudelaa and Marslen-Wilson (2013, p. 1471) provided empirical evidence that MSA and STA "are cognitively processed and represented in fundamentally similar ways," and they found no evidence of less-effective processing of MSA. Moreover, Boudelaa and Marslen-Wilson observed that the similarities between MSA and the STA dialect make it difficult to accept the argument that MSA is represented separately as a second language. One point they made is relevant to the present article: the very close match between the consonantal repertoires of MSA and other Arabic dialects. This close match, however, does not necessarily entail that these consonants are organized phonologically in parallel ways within MSA and the dialect. For instance, whereas the phonotactics of MSA do not allow consonant clusters in word-initial positions, some dialects do allow these. Najdi Arabic, for example, allows biconsonantal clusters in word-initial positions (Ingham, 1994).

For those learning Arabic as a foreign language, there is still debate about which variety of Arabic should be taught, but Badawi (2006) observed that MSA has been the variety taught most often in formal settings. Some have also called recently for teaching MSA as a lingua franca for nonnative speakers (Jaradat & Al-Khawaldeh, 2015). Thus, the online calculator that we developed will be of use not only to language scientists studying Arabic in laboratory settings, but also to instructors studying the acquisition of Arabic as a foreign language in the classroom.

Despite its important and growing status, MSA has been relatively understudied in psycholinguistics, and as we noted above, few resources exist for calculating Arabic sounds. The richness and complexity of sounds in the Arabic language, as well as its unique orthography, may have prevented more researchers from examining the language and developing these necessary resources.

MSA consists of 28 consonants. One distinctive feature of Arabic consonants is *emphasis*. MSA has four emphatic consonants:/d/,/t/,/ð/, and/s/. The emphatic form of these consonants is produced with the tongue retracted toward the pharyngeal wall (Amayreh & Dyson, 1998), or as Ryding (2005, p. 14) puts it "with a raised and a tensed tongue." The emphatic form of these consonants is indicated by the IPA symbol/ /in Table 1. In contrast, the vowel system in MSA is composed of only the three short vowels/i/,/a/, and/u/and their corresponding long ones,/i:/,/a:/, and/u:/. The difference between the short and long vowels in MSA is phonemic. For instance, *qad* "maybe" and *qaad* "he led" are two different words. The phonetic realization of the vowels is determined by the consonants surrounding them (Holes, 1995). The six vowels are shown in Table 2.

Finally, two diphthongs are found in the phonemic inventory of MSA. The vowel/a/combines with the glides/j/and/w/ to form the two diphthongs available in Arabic,/ai/and/au/, as in the English words *kite* and *cow*, respectively.

MSA morphology

As in other Semitic languages, Arabic morphology is based on root–pattern interaction (Holes 1995). The *root* consists of a number of discontinuous consonants (two to five consonants, with three-consonant roots being the most common). The

Table 1 Consonants of Modern Standard Arabic

	Bilabial	Labiodental	Dental	Alveodental	Palatal	Velar	Uvular	Pharyngeal	Glottal
Stops	b			d t dˤ tˤ		k	q		ʔ
Fricatives		f	ð θ ðˤ	z s sˤ	ʃ		ʁ χ	ʕ ħ	h
Affricates					ʤ				
Nasals	m			n					
Liquid				l					
Tap/trill				ɾ/r					
Glide	w				j				

Adapted from Ryding (2006) and Amayreh (2003). In each column, voiced segments are on the left, voiceless segments on the right

pattern, on the other hand, consists primarily of vowels that work as a template for the root. The root determines the semantic meaning, whereas the pattern determines the morphosyntactic properties of the word and creates its phonological structure (Wright, 1995). For words to be formed, mapping the root onto a pattern is necessary. One of the most commonly used examples of Arabic roots in the literature is {ktb}, which involves the notion of "writing." Examples of the mapping of this root into vowel patterns (sometimes also containing other consonants) are provided in Table 3.

Arabic writing system

Because we based the Arabic Phonotactic Probability Calculator on a database of written MSA, we describe in this section some important details regarding the orthography of Arabic. It is not unusual to use orthographic corpora for research in spoken-language processing. Indeed, the word frequency counts of Brysbaert and New (2009), based on movie subtitles, have been widely used by researchers in a variety of fields, with over 750 citations to date (as per Google Scholar). The work of Brysbaert and New was done in part to update and replace the word frequency counts developed by Kučera and Francis (1967; which were also based on written materials).

The orthography of Arabic can be described as *phonemic* or *shallow*, meaning that it has a close correspondence between sounds and letters. However, not all Arabic sounds are marked by letters. Whereas consonants and vowels are marked by letters, short vowels and gemination (consonant doubling or prolonging) are marked by diacritics. These are small marks used either under or above the letters representing the sounds they follow. They are not typically used in normal writing, rendering some words, especially those from the same root, homographic. Consider the following example: The word فهم without diacritics can mean *fahima* "he understood," *fuhima* "it was understood," *fahm* "understanding," or *fahhama* "he explained." If diacritics are used, the word will be written فَهِمَ, فُهِمَ, فَهْم, or فَهَّمَ, respectively.

What makes diacritics problematic is that their use is optional. They are only consistently used in the Quran, children's stories, and some language-teaching materials. Although the absence of diacritics will render many Arabic words homographic, skilled Arabic readers are assumed to be able to extract the meaning from the context in order to disambiguate these homographs.

Another interesting feature of Arabic orthography is that some letters representing certain sounds can be spelled differently, on the basis of fixed rules. *Hamza*, for example (the symbol for the glottal stop sound), can either sit on a "chair" that takes different letter shapes or stand on its own on the line, depending on its position in the word and the surrounding vowels (Holes, 1995; Ryding, 2005). The chair for *hamza* takes the shapes of the letters representing long vowels in Arabic (ي for/i:/, ا for/a:/, and و for/u:/). Therefore, *hamza* in words such as/ma:ʔ/"water,"/ʔum/"mother,"/bi:ʔah/ "environment," and/muʔallif/"author" will be spelled in

Table 2 Vowels of Modern Standard Arabic

	Front	Central	Back
Close	i/i:		u/u:
Open		a/a:	

Table 3 Mapping of the root {ktb} into different patterns

Pattern	Word	English Meaning
CaCaCa	Kataba (v)	He wrote
CaaCaCa	Kaataba (v)	He corresponded
CiCaaC	Kitaab (n)	book
CaaCiC	Kaatib (n)	writer
maCCaC	Maktab (n)	office
CuCuC	kutub (n)	books
maCCuuC	Maktuub (pp)	written

v, verb; n, noun; pp, past participle

Arabic as أم, بيئة, ماء and مؤلف, respectively, where the underlined parts spell out the *hamza*.

Hamza in Arabic comes in two forms: strong *hamza*, which is a consonant that is part of the sounds in the word, as in the examples above, and weak *hamza*, which is a helping sound that only appears at the beginning of the word to help with the pronunciation of its consonant clusters. Weak *hamza* always appears attached to a following short vowel and is only pronounced when the word is utterance-initial; otherwise, it is omitted (Ryding, 2005). The symbol (ا) is used to distinguish weak *hamza* from strong *hamza* (أ) in Arabic orthography.

Similar to the weak *hamza*, another letter in Arabic can be pronounced differently in pausal and juncture contexts. The *taa marbutah* (the "tied *taa*") letter is used as a feminine gender marker and only appears in word-final position in either nouns or adjectives. A word like طالب/tˁa:lib/"student" becomes طالبة/tˁa:libah/when in feminine form. However, some words are inherently feminine, such as مدرسة/madrasah/"school" and فكرة/fikrah/"idea." Generally, the *taa marbutah* is pronounced as/t/when the case-ending vowel attached to it is pronounced in juncture; otherwise, when the word is pronounced in pausal form, it is pronounced as/h/.

Another sound that has more than one spelling variant in Arabic is the long vowel/a:/. These variants depend not only on vowel position, but also on the derivational etymology of the word. Therefore, the vowel/a:/could be spelled differently even in the same position. This is particularly apparent when this vowel is in word-final position. For example, the homophonic words "stick or cane" and "disobeyed" in Arabic are both pronounced as/ʕasˁa:/, but they are spelled عصا and عصى, respectively.

Other letters in Arabic play a dual role. The letters waw (و) can represent either the consonant/w/or the long vowel/u:/. Similarly, the letter yaa (ي) can represent either the consonant/j/or the long vowel/i:/.

Method

The corpus underlying the Arabic Phonotactic Probability Calculator (APPC)

Because it represents the official language in the Arab world, MSA vocabulary is frequently changing, adding new words to accommodate new technical, political, and scientific terminology (see Ryding, 2005, for a discussion of borrowing in MSA). The first step in creating the calculator was to obtain a wordlist of MSA that also contained frequency-of-occurrence counts for the words. We further desired a corpus that was current and contained information regarding short vowels and gemination. Recall that short vowels and gemination in MSA are not marked by letters; rather, they are marked by diacritics, and these diacritics are not usually written. For a corpus to be selected, it had to include these diacritics so that words would not be ambiguous.

One corpus that fit these desiderata was Ar ten ten (Arts, Belinkov, Habash, Kilgarriff, & Suchomel, 2014), which is a written corpus of Arabic gathered in 2012, comprising 5.8 billion words. A large corpus was desired so that word frequency counts would be more reliable. Ar ten ten also includes diacritics, enabling us to disambiguate words, because a subset containing 115 million words was processed with MADA (Habash & Rambow, 2005; Habash, Rambow, & Roth, 2009). MADA is an Arabic language-processing tool that uses a morphological analyzer for MSA that works at disambiguating MSA words in context by reaching a preferred analysis for each undiacritized word (see Arts et al., 2014, section 4.3, for a full description of the processing with MADA). The 115-million-word subset, which was lemmatized and tagged for parts of speech by MADA, was loaded to a corpus manager called Sketch Engine (Kilgarriff, Rychly, Smrz, & Tugwell, 2004).

Other advantages of Ar ten ten include that only text in sentences was used, thereby better reflecting the contextual usage of those words (Arts et al., 2014). Also, the Web domains in the corpus are from different Arab countries (Palestine, Saudi Arabia, Syria, Egypt, United Arab Emirates, Jordan, Sudan, Morocco, and Lebanon), so that the use of MSA from different regions is also represented. Furthermore, Ar ten ten sampled text from social websites where users share their complaints and concerns (e.g., http://humum.net, used mainly in Egypt) or religious and social questions (e.g., http://m.islamweb.net, used mainly in the Gulf countries, Egypt, Morocco, and Algeria). This is particularly important because it indicates that in this corpus, MSA is used in various contexts rather than in fixed, scripted ones (e.g., news texts).

A frequency wordlist containing the 100,000 most frequent MSA words was purchased from Sketch Engine (https://www.sketchengine.co.uk/). For regular words, the lemma (i.e., the uninflected dictionary citation form of the word) was used for the database that underlies the APPC. For example, when a word form with the feminine marker *taa marbutah* (e.g., طالبة/ta:libah/) or the plural marker/u:n/(e.g., mu?minu:n مؤمنون) appeared in the corpus, the lemmatized uninflected forms/ta:lib/and/mu?min/were used in the calculator. However, for irregular forms, such as irregular (broken) plural forms, the internal structure (i.e., the pattern) of the word was changed (e.g., providing the lemma/tˁa:lib/طالب for the broken plural/tˁulla:b/طلاب.). To avoid loss of the phonological information in the Arabic vowel patterns contained in these irregular forms, we instead used the phonemic transcriptions of the irregular forms.

The lemmas in the wordlist were analyzed in fully vowelized Buckwalter (2002) transliteration. Buckwalter

transliteration represents Arabic script by using ASCII characters to romanize the orthographic forms of the words; it is not a phonological transcription system. In other words, it reflects the spelling variants in Arabic script of the same phonemes discussed above. For example, as in Arabic script, the Buckwalter transliteration uses several different characters for the glottal stop and the/a:/vowel.

Because we were interested in phonological rather than orthographic representations, we encoded all different variants (transliterations) of the same sound (i.e., glottal stop and/a:/) as one symbol, as is shown in Table 4. Moreover, to avoid Buckwalter's (2002) dual function of symbols, we used the uppercase symbols I, A, and U to transcribe the long vowels/i:/,/a:/, and/u:/, respectively, and we used the symbols Y and W to transcribe the two diphthongs/a /and/a /.

Also a choice was made of which pronunciations to keep for the weak *hamza* and *taa marbutah*, discussed above. Since weak *hamza* is only pronounced when the word is utterance-initial, we chose to omit it. Therefore, a word starting with a weak *hamza*, such as اسم, was transcribed as/ism/rather than/ʔism/. Also, the pronunciation/h/was chosen over/t/for *taa marbutah*, to preserve a more representative frequency of the sound/h/in word-final position in MSA (e.g.,/madrasah/"school").

There was a large number of homophones among the lemmas of the wordlist. These resulted from the Ar ten ten corpus providing the same lemmas for all inflected forms of the same word. There were also other homophonic words, such as the example/ as a:/given above. Multiple forms of the homophones were reduced to the most frequent form and included in the calculator.

All proper nouns and directly borrowed words (except those with no Arabic equivalents) were also removed from the corpus. The pronunciations and transcription of all the entries in the database (lemmatized wordlist) were manually checked and edited (where needed) by a trained native speaker of Arabic with the help of the Almaany online Arabic dictionary. The final database had 11,164 unique lemmas of MSA in phonemic transcription. Frequency counts of these lemmas were used to derive the positional segment and biphone frequencies of MSA, as we explain below.

As in Vitevitch and Luce (2004) and other works cited in this report, two measures are used to estimate phonotactic probability: (1) positional segment frequency (i.e., how often a particular segment occurs in a certain position in a word) and (2) biphone frequency (i.e., segment-to-segment co-occurrence probabilities of sounds within a word).

The positional segment frequency was calculated by searching the computer-readable transcriptions for all of the words in the dictionary (regardless of word length) that contained a given segment in a given position. The log (base 10) values of the frequencies with which those words occurred in the Ar ten ten corpus were summed together, and then divided by the total log (base 10) frequency of all the words in the corpus with a segment in that position, to provide an estimate of probability. Log values of the word frequency counts were used because log values better reflect the distribution of frequency of occurrence in the language (Zipf, 1935) and better correlate with performance in various psycholinguistic tasks than the raw frequency values (e.g., Balota,

Table 4 ASCII characters used in the calculator and International Phonetic Alphabet (IPA) transcriptions

	ASCII Used in APPC	IPA
Stops	t	t
	T	tˤ
	k	k
	q	q
	?	ʔ
	d	d
	D	dˤ
	b	b
Affricates	j	ʤ
Fricatives	f	f
	v	θ
	s	s
	S	sˤ
	$	ʃ
	x	x-χ
	H	ħ
	h	h
	*	ð
	Z	ðˤ
	z	z
	g	ɣ–ʁ
	E	ʕ
Nasals	m	m
	n	n
Trill	r	r
Glides	w	w
	y	j
Liquid	l	l
Gemination	~	
Vowels	i	i
	a	a
	u	u
	I	i
	A	a
	U	uː
	Y	aɪ
	W	aʊ

Most of the ASCII characters used in the calculator are based on Buckwalter's (2002) transliterations, with necessary amendments to accommodate certain characteristics of the MSA phonemic inventory

Pilotti, & Cortese, 2001). Thus, the estimates of position-specific segment frequencies are token- rather than type-based estimates of probability.

Consider the following example of word-initial/k/. All of the words in the dictionary that contained/k/in the initial position were selected, and the log values of the frequency counts for those words were summed and then divided by the summed log values of the frequency counts for all words in the corpus that have a segment in that position, to produce a token-based probability estimate that/k/will occur in the initial position of a word in MSA.

Now consider/k/in the second position of the word. Again, the log values of the frequency counts for those words with/k/in the second position would be summed and then divided by the summed log values of the Ar ten ten frequency counts for all words in the dictionary that have a segment in the second position, to produce a token-based probability estimate that/k/will occur in the second position of a word in MSA. In this instance, all words that are one phoneme in length will not be included in the denominator, because they do not contain a phoneme in the second position of the word. Similarly, when calculating the probability of a segment in the third position of a word, words that are only one or two phonemes long are not included in the denominator, because they do not contain a phoneme in the third position of the word.

The position-specific biphone frequency was calculated in a similar way. That is, all the instances that a sequence of two phonemes occurred together in specific adjacent positions in a word were counted. The log values of the frequencies of occurrence for the words in which the (position-specific) two-phoneme sequences were found were summed, and then divided by the summed log values of the Ar ten ten frequency counts for all words in the dictionary that contained phonemes in those two adjacent positions, to yield a token-based estimate of the position-specific biphone probability.

As in the English calculator described by Vitevitch and Luce (2004), the biphone probability that is calculated by the Arabic version of the Phonotactic Probability Calculator is based on the co-occurrence of segments within words, and not on the transitional probabilities of sounds between words as they might occur in fluent speech (cf. Gomez & Gerken, 2000).

How to use the calculator

The APPC can be freely accessed at http://phonotactic.drupal.ku.edu/to calculate the phonotactic probabilities of both real MSA words and pseudowords up to 15 phonemes in length. The gemination marker (~), which must be added after a geminated consonant, counts as a phoneme.

The landing page offers the user the option to calculate phonotactic probability for English (a revised version of the Vitevitch & Luce, 2004, site), Spanish, or Arabic. After selecting "Arabic" from the list of languages, a screen like the one depicted in Fig. 1 will appear. To calculate the phonotactic probabilities of your items (words or nonwords), enter their phonemic transcriptions using the ASCII characters provided in Table 4. One item should be entered per line. This can be done either by typing the item directly into the field and using <Enter> to move to the next line, or by copying your items from a text file and pasting them into the field (one item per line, using a hard return). There is no limit to the number of items you can enter (by either typing or copying and pasting) in the calculator's field. However, speed of calculation is affected by the number of the items that have been entered, as well as by the amount of traffic on the network.

In Fig. 1, the ASCII transcriptions of the two Arabic words/qa:l/("said") and/katab/("wrote") are entered. By clicking the Run button, the phonotactic probabilities of the items will be calculated and appear on a new page, as in Fig. 2. Each item will typically have three lines of results. On the first line of the

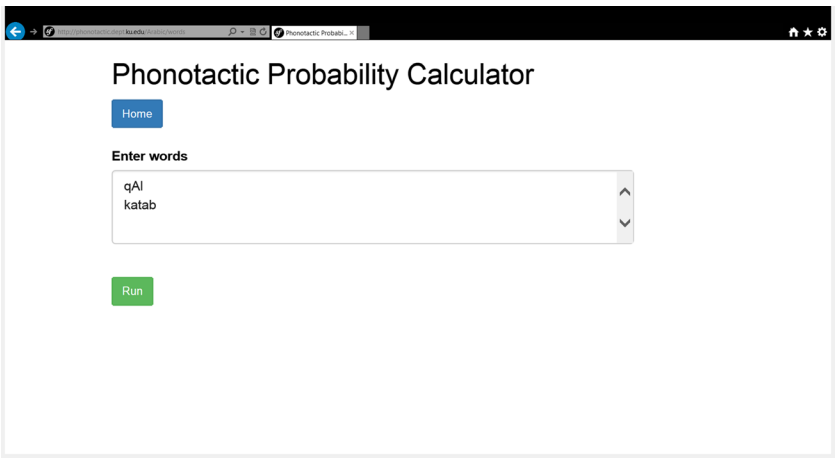

Fig. 1 Depiction of the MSA Phonotactic Probability Calculator's input field page. This is where the computer-readable phonemic transcription is entered

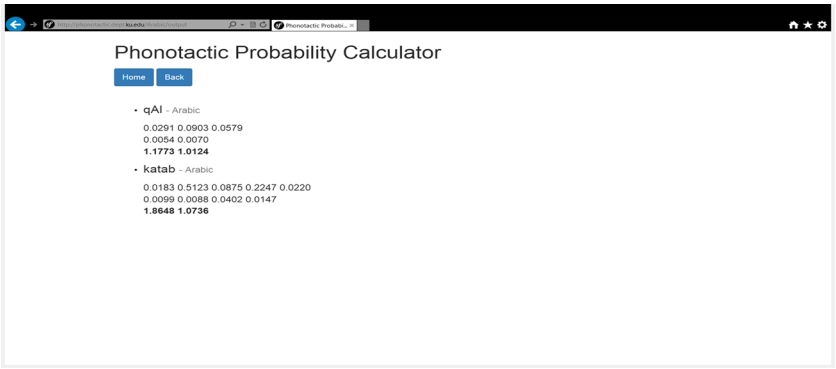

Fig. 2 Output page for the two Arabic words/qAl/and/katab/

output page, the word "Arabic" will appear next to the phonemic transcription of the item the user has entered, to show that "Arabic" (and not another language) has been selected. The second line contains the position-specific probability of each segment in the item that has been entered. In the three-phoneme word qAl/qa:l/, the phoneme/q/has an initial position probability of .0291;/a:/has a medial (second) position probability of .0903, and/l/has a final (third) position probability of .0579. If a zero position-specific probability appears in any position, it means that this sound does not appear in that specific position in MSA (i.e., that segment is illegal in that position).

The third line contains the position-specific biphone probability for each of the biphones of the item. The three phonemes of qAl/qa:l/will have two biphones,/qa:/and/a:l/, with biphone probabilities of .0054 and .0070, respectively. A zero biphone probability indicates that the sequence of phonemes is illegal in Arabic (e.g., consonant clusters in initial position).

The final line for each entry sums up all the phoneme probabilities and biphone probabilities for the entry. In the example of qAl, the sum of phoneme probabilities is 1.1773, and the sum of biphone probabilities is 1.0124. Following the convention established in Vitevitch and Luce (2004), the value of 1 has been added to these sums to draw attention to these commonly reported values. However, because probability cannot exceed the value of 1, the leading 1 should not be included when reporting results.

Conclusion

The APPC described in the present work was based very closely on the Phonotactic Probability Calculator for English described by Vitevitch and Luce (2004). Like the English version, the APPC provides a very simple measure of phonotactic probability that is relatively neutral regarding linguistic theory. Despite the simplicity of the metric, the Phonotactic Probability Calculator for English has proven useful to language researchers from a variety of areas for stimulating new research questions and examining a number of populations (e.g., older adults: Hunter, 2016; people with aphasia or apraxia: Cunningham, Haley, & Jacks, 2016; preschool children: Goodrich & Lonigan, 2015; children with specific language impairment: Richtsmeier & Goffman, 2015; and adult users of cochlear implants: Vitevitch, Pisoni, Kirk, Hay-McCutcheon, & Yount, 2002). The Phonotactic Probability Calculator for English has also proven useful to instructors in various language sciences. We hope that the APPC described here will similarly stimulate new research questions and prove useful in the classroom as well as in clinical research.

Acknowledgements We thank the Council for International Exchange of Scholars for funding F.A. through the Fulbright Scholars Program while he was at the University of Kansas, and the University of Kansas Information Technology department (especially Erica Boos, Chris Escalante, and Bob Lim) for their work on the interface.

References

Adriaans, F. (2006). *PhonotacTools (Test version) [Computer program]*. The Netherlands: Utrecht Institute of Linguistics OTS, Utrecht University.

Amayreh, M. M. (2003). Completion of the consonant inventory of Arabic. *Journal of Speech, Language, and Hearing Research, 46*, 517–529.

Amayreh, M. M., & Dyson, A. T. (1998). The acquisition of Arabic consonants. *Journal of Speech, Language, and Hearing Research, 41,* 642–653.

Anderson, J. D., & Byrd, C. T. (2008). Phonotactic probability effects in children who stutter. *Journal of Speech, Language, and Hearing Research, 51,* 851–866. doi:10.1044/1092-4388(2008/062)

Arts, T., Belinkov, Y., Habash, N., Kilgarriff, A., & Suchomel, V. (2014). arTenTen: Arabic corpus and word sketches. *Journal of King Saud University - Computer and Information Sciences, 26,* 357–371.

Auer, E. T., & Luce, P. A. (2005). Probabilistic phonotactics in spoken word recognition. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 610–630). Oxford, UK: Blackwell. doi:10.1002/9780470757024.ch25

Badawi, E. (2006). Arabic for nonnative speakers in the 21st century: A shopping list. In K. M. Wahba, Z. A. Taha, & L. England (Eds.), *Handbook for Arabic language teaching professionals* (pp. ix–xiv). Mahwah, NJ: Erlbaum.

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language, 44,* 568–591.

Balota, D. A., Pilotti, M., & Cortese, M. J. (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition, 29,* 639–647. doi:10.3758/BF03200465

Boudelaa, S., & Marslen-Wilson, W. D. (2013). Morphological structure in the Arabic mental lexicon: Parallels between standard and dialectal Arabic. *Language and Cognitive Processes, 28,* 1453–1473.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41,* 977–990. doi:10.3758/BRM.41.4.977

Buckwalter, T. (2002). Arabic transliteration. URL www.qamus.org/transliteration.htm

Cunningham, K. T., Haley, K. L., & Jacks, A. (2016). Speech sound distortions in aphasia and apraxia of speech: Reliability and diagnostic significance. *Aphasiology, 30,* 396–413.

Faizal, S. S. B., Khattab, G., & McKean, C. (2015). 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. In *Proceedings of the 18th International Congress of Phonetic Sciences*. Retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0968.pdf on April 26, 2016.

Ferguson, C. A. (1959). Diglossia. *Word, 15,* 325–340.

Furman, N., Goldberg, D., & Lusin, N. (2007). Enrollments in languages other than English in United States institutions of higher education, Fall 2006. *Modern Language Association of America*. Retrieved from www.mla.org/homepage on November 29, 2007.

Goldrick, M., & Larson, M. (2008). Phonotactic probability influences speech production. *Cognition, 107,* 1155–1164. doi:10.1016/j.cognition.2007.11.009

Gomez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences, 4,* 178–186.

Goodrich, J., & Lonigan, C. J. (2015). Lexical characteristics of words and phonological awareness skills of preschool children. *Applied Psycholinguistics, 36,* 1509–1531.

Habash, N., & Rambow, O. (2005). Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In K. Knight (Ed.), *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (pp. 573–580). Stroudsburg, PA: Association for Computational Linguistics.

Habash, N., Rambow, O., & Roth, R. (2009). MADA + TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR)* (pp. 102–109). Cairo, Egypt: MEDAR.

Holes, C. (1995). *Modern Arabic: Structure, functions and varieties*. London, UK: Longman.

Hunter, C. R. (2016). Is the time course of lexical activation and competition in spoken word recognition affected by adult aging? An event-related potential (ERP) study. *Neuropsychologia, 91,* 451–464. doi:10.1016/j.neuropsychologia.2016.09.007

Ibrahim, R., & Aharon-Peretz, J. (2005). Is literary Arabic a second language for native Arab speakers? *Journal of Psycholinguistic Research, 34,* 51–70.

Ingham, B. (1994). *Najdi Arabic: Central Arabian*. Amsterdam, The Netherlands: Benjamins.

Jaradat, A. A., & Al-Khawaldeh, N. N. A. (2015). Teaching Modern Standard Arabic for non-native speakers as a lingua franca. *Mediterranean Journal of Social Sciences, 6,* 490–499.

Jusczyk, P. W., & Luce, P. A. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language, 33,* 630–645.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The sketch engine. In *Proceedings of EURALEX, Lorient, France* (pp. 105–116). Available from www.euralex.org/elx_proceedings/Euralex2004/

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Leonard, L. B., Davis, J., & Deevy, P. (2007). Phonotactic probability and past tense use by children with specific language impairment and their typically developing peers. *Clinical Linguistics and Phonetics, 21,* 747–758.

Lewis, M. P., Simons, G. F., & Fennig, C. D. (Eds.). (2016). *Ethnologue: Languages of the world* (19th ed.). Dallas, Texas: SIL International. Online version at www.ethnologue.com

Marian, V., Bartolotti, J., Chabal, S., & Shook, A. (2012). CLEARPOND: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. *PLoS ONE, 7,* e43230. doi:10.1371/journal.pone.0043230

Mattys, S. L., & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition, 78,* 91–121.

Messer, M. H., Leseman, P. P., Boom, J., & Mayo, A. Y. (2010). Phonotactic probability effect in nonword recall and its relationship with vocabulary in monolingual and bilingual preschoolers. *Journal of Experimental Child Psychology, 105,* 306–323.

Palmer, J. (2008). Arabic diglossia: Student perception of spoken Arabic after living in the Arabic-speaking world. *Arizona Working Papers in Second Language Acquisition and Teaching, 15,* 81–95.

Parkinson, D. B. (2003). Verbal features in oral fusha in Cairo. *International Journal of the Sociology of Language, 163,* 27–41.

Pitt, M. A., & McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language, 39,* 347–370.

Richtsmeier, P. T., & Goffman, L. (2015). Learning trajectories for speech motor performance in children with specific language impairment. *Journal of Communication Disorders, 55,* 31–43.

Rispens, J., Baker, A., & Duinmeijer, I. (2015). Word recognition and nonword repetition in children with language disorders: The effects of neighborhood density, lexical frequency, and phonotactic probability. *Journal of Speech, Language, and Hearing Research, 85,* 78–92. doi:10.1044/2014_JSLHR-L-12-0393

Ryding, K. C. (2005). *A reference grammar of modern standard Arabic*. Cambridge, UK: Cambridge University Press.

Ryding, C. K. (2006). Teaching Arabic in the United States. In K. M. Wahba, Z. A. Taha, & L. England (Eds.), *Handbook for Arabic language teaching professionals in the 21st century* (pp.13–20). Mahwah, NJ: Lawrence Erlbaum Associates.

Storkel, H. L. (2001). Learning new words: Phonotactic probability in language development. *Journal of Speech, Language, and Hearing Research, 44,* 1321–1337.

Storkel, H. L. (2004). The Emerging lexicon of children with phonological delays: Phonotactic constraints and probability in acquisition. *Journal of Speech, Language, and Hearing Research, 47,* 1194–1212.

Storkel, H. L. (2013). A corpus of consonant–vowel–consonant real words and nonwords: Comparison of phonotactic probability, neighborhood density, and consonant age of acquisition. *Behavior Research Methods, 45,* 1159–1167. doi:10.3758/s13428-012-0309-7

Storkel, H. L., Armbruster, J., & Hogan, T. P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research, 49,* 1175–1192.

van der Kleij, S. W., Rispens, J. E., & Scheper, A. R. (2016). The effect of phonotactic probability and neighbourhood density on pseudoword learning in 6- and 7-year-old children. *First Language, 36,* 93–108.

Vitevitch, M. S., Chan, K. Y., & Goldstein, R. (2014). Using English as a "model language" to understand language processing. In A. Lowit & N. Miller (Eds.), *Motor speech disorders: A cross-language perspective* (pp. 58–73). Buffalo, NY: Multilingual Matters.

Vitevitch, M. S., & Donoso, A. J. (2012). Phonotactic probability of brand names: I'd buy that! *Psychological Research, 76,* 693–698.

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science, 9,* 325–329. doi:10.1111/1467-9280.00064

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language, 40,* 374–408.

Vitevitch, M. S., & Luce, P. A. (2004). A Web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, and Computers, 36,* 481–487. doi:10.3758/BF03195594

Vitevitch, M. S., & Luce, P. A. (2005). Increases in phonotactic probability facilitate spoken nonword repetition. *Journal of Memory and Language, 52,* 193–204. doi:10.1016/j.jml.2004.10.003

Vitevitch, M. S., Luce, P. A., Charles-Luce, J., & Kemmerer, D. (1997). Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language and Speech, 40,* 47–62.

Vitevitch, M. S., Luce, P. A., Pisoni, D. B., & Auer, E. T. (1999). Phonotactics, neighborhood activation and lexical access for spoken words. *Brain and Language, 68,* 306–311.

Vitevitch, M. S., Pisoni, D. B., Kirk, K. I., Hay-McCutcheon, M., & Yount, S. L. (2002). Effects of phonotactic probabilities on the processing of spoken words and nonwords by postlingually deafened adults with cochlear implants. *Volta Review, 102,* 283–302.

Vitevitch, M. S., & Stamer, M. K. (2006). The curious case of competition in Spanish speech production. *Language and Cognitive Processes, 21,* 760–770. doi:10.1080/01690960500287196

Weber, A., & Cutler, A. (2006). First-language phonotactics in second-language listening. *Journal of the Acoustical Society of America, 119,* 597–607.

Wright, W. (1995). *A grammar of the Arabic language*. Cambridge, UK: Cambridge University Press.

Zipf, G. K. (1935). *The psycho-biology of language: An introduction to dynamic philology*. Cambridge, MA: Houghton Mifflin.