Automatic Phonetization-based Statistical Linguistic Study of Standard Arabic

Fadi Sindran *fadi.sindran@faui51.informatik.uni-erlangen.de*
Faculty of Engineering/ Department of Computer Science
Pattern Recognition Lab
Friedrich-Alexander-Universität Erlangen-Nürnberg
Erlangen, 91058, Germany

Firas Mualla *firas.mualla@cs.fau.de*
Faculty of Engineering/ Department of Computer Science
Pattern Recognition Lab
Friedrich-Alexander-Universität Erlangen-Nürnberg
Erlangen, 91058, Germany

Tino Haderlein *Tino.Haderlein@cs.fau.de*
Faculty of Engineering/ Department of Computer Science
Pattern Recognition Lab
Friedrich-Alexander-Universität Erlangen-Nürnberg
Erlangen, 91058, Germany

Khaled Daqrouq *haleddaq@yahoo.com*
Department of Electrical and Computer Engineering
King Abdulaziz University,
Jeddah, 22254, Saudi Arabia

Elmar Nöth *noeth@cs.fau.de*
Faculty of Engineering/ Department of Computer Science
Pattern Recognition Lab
Friedrich-Alexander-Universität Erlangen-Nürnberg
Erlangen, 91058, Germany

Abstract

Statistical studies based on automatic phonetic transcription of Standard Arabic texts are rare, and even though studies have been performed, they have been done only on one level – phoneme or syllable – and the results cannot be generalized on the language as a whole. In this paper we automatically derived accurate statistical information about phonemes, allophones, syllables, and allosyllables in Standard Arabic. A corpus of more than 5 million words, including words and sentences from both Classical Arabic and Modern Standard Arabic, has been prepared and preprocessed. We developed a software package to accomplish a rule-based automatic transcription from written Standard Arabic text to the corresponding linguistic units at four levels: phoneme, allophone, syllable, and allosyllable. After testing the software on four corpora including more than 57000 vocabulary words, and achieving a very high accuracy (> 99 %) on the four levels, we used this software as a reliable tool for the automatic transcription of the corpus used in this paper and evaluated the following: 1) the vocabulary phonemes, allophones, syllables, and allosyllables with their specific percentages in Standard Arabic. 2) the best curve equation from the distribution of phonemes, allophones, syllables, and allosyllables normalized frequencies. 3) important statistical information, such as percentage of consonants and vowels, percentage of the consonants classified by the place and way of articulation, the transition probability matrix between phonemes, and percentages of syllables according to the type of syllable, etc.

Keywords: Statistical Studies, Standard Arabic, Phonetic Transcription, Phonetization, Ranked Frequency Distribution, Phonemes, Allophones, Syllables, Allosyllables, Fit of Equation.

1. INTRODUCTION

Arabic is a Semitic language and belongs to the family of Afro-Asian languages. It is an official language in more than 25 countries with more than 400 million speakers, and one of the six official languages of the United Nations, as well as Arabic is a liturgical language for more than 1500 million Muslims. In our work we will deal with Standard Arabic, which includes Classical Arabic (CA) and Modern Standard Arabic (MSA). Classical Arabic is the language of the Holy Qur'an and books of Arabic heritage. MSA, on the other hand, is the formal language in all Arab countries [1], taught in schools and universities, used in radio and television along with local dialects, and it is the predominant language in which books and newspapers are written. In this paper, we will present an automatically obtained comprehensive statistical study of Standard Arabic at the level of phonemes, allophones, syllables, and allosyllables. The fit of many equations to the distribution of frequencies at these four levels, which means testing the suitability of these equations to represent the curves resulting from the distribution of frequencies, will be tested. For this purpose we prepared a very large corpus with more than 5 million vocalized words containing both words and sentences from CA and MSA. The importance of this work lies in the following:

- To the best of our knowledge, it is the first fully automated comprehensive statistical study of linguistic units in Standard Arabic done using a fully vocalized corpus.
- It offers a deeper understanding of the structure of Standard Arabic in terms of phones and how they are related to each other. It also gives accurate statistical information about phonemes, allophones, syllables, and allosyllables.
- The equations that best fit the ranked frequency distribution of phonemes, allophones, syllables, and allosyllables were accurately identified and tested depending on the Goodness of Fit (GoF) parameters.
- The corpus used in the study is very large and carefully prepared. It is fully vocalized, so it can be very helpful in machine learning-based automatic vocalization of Standard Arabic texts. Vocalization is one of the major challenges for the Arabic natural language processing [2].
- The output of the automatic phonetization at phonemic and phonetic level are very important components, which can be used in Arabic Text-to-Speech (TTS), Automatic Speech Recognition (ASR), and Computer-Aided Pronunciation Learning (CAPL).

The main work in this paper can be divided into three steps as following:

1) Data preparation by building and pre-processing a very large fully vocalized corpus of words and sentences
2) Automatic phonetization on four levels: phonemes, allophones, syllables, and allosyllables. This was done based on our work [3] which we developed to achieve a letter-to-sound transcription with very high accuracy (> 99 %) at both phonemic and phonetic level.
3) Processing of the output of the previous step, obtaining statistics, and extracting important information from statistics. We will examine the fit of different equations to phoneme, allophone, syllable, and allosyllable frequencies in Standard Arabic. These equations are: Zipf (1), Yule (2), Sigured (3), Borodovsky and Gusein-Zade (4), and exponential (5) and (6). Zipf-Mandelbrot's law [4] is one of the most famous models proposed to account for frequency distributions of words and other linguistic units [5].

$$F_r = \frac{a}{r^b} \qquad (1)$$

$$F_r = \frac{a}{r^b} c^r \quad (2)$$

$$F_r = \frac{1-a}{1-a^n} a^{r-1} \quad (3)$$

$$F_r = \frac{1}{n} log \frac{n+1}{r} \quad (4)$$

$$F_r = ae^{br} \quad (5)$$

$$F_r = ae^{br} + ce^{dr} \quad (6)$$

$\boldsymbol{F_r}$ is the frequency of the symbol (in our case phoneme, allophone, syllable, or allosyllable) with rank $\boldsymbol{r}$, $\boldsymbol{r}$ is the rank of the symbol when frequencies are in descending order (the maximum frequency is corresponding with $\boldsymbol{r = 1}$ while $\boldsymbol{r = n}$ is corresponding with the minimum frequency), $\boldsymbol{n}$ is the number of symbols, and $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$, and $\boldsymbol{d}$ are the parameters to be estimated from the data.

The importance of finding the best fitting equations is as follows:

- Depending only on the best curve equations and the frequencies of the most frequent linguistic unit (phoneme, allophone, syllable, and allosyllable), we can approximately derive the frequencies of all rest linguistic units in a Standard Arabic text.
- Suitability of texts, in terms of phonological richness and balance, for training an Arabic ASR or TTS, can be automatically tested. Usually the linguists prepare these texts manually, and it requires a lot of time and effort. The best text for the training must fulfill the following characters:
    - ✓ It must contain all phonemes and allophones in Standard Arabic.
    - ✓ It must be phonetically balanced, that means it must maintain the same order of phoneme and allophone ranks as in Standard Arabic language and the same best fit equation of the ranked frequency distribution of phonemes and allophones with the least possible deviations. We calculate the root mean squared error (RMSE) values between the curve equations of the ranked frequency distribution of phonemes and allophones in Standard Arabic and the text we test. The closer the values are to zero, the more appropriate is the text for training.

2. RELATED WORKS

Moussa, in a lecture on the computerization of the Arab heritage [6], explained the relation between the consonants and vowels in Classical Arabic from a statistical point of view. His study was on two samples from the Qur'an: Surat al-A'raf (The Heights) /suratulʔaʕraːf/ (Arabic: " سُورَةُ الأَعرَاف"), which is the seventh chapter (sura) of the Qur'an, some short chapters, and Surat al-Baqara (The Cow) /suratulbaqarah/ (Arabic: "سُورَةُ البَقَرَة"), which is the longest sura of the Qur'an containing 6150 words. He presented the percentage of the vowel frequencies as well as the percentage of the frequencies of the short vowel in combination with the previous consonant. AbuSalim in [7] considered only the syllables within isolated words. He explained the syllabic structure of the vocabulary in the lexicon of contemporary Arabic (Arabic: "مُعجَمُ اللُّغَةِ العَرَبِيَّةِ المُعَاصِرَة"). Ibrahim in his master thesis [8] studied the syllables within sentences in Surat al-Baqara. The only study that addressed the phoneme distribution is [9] performed by Tambovtsev and Martindale. They studied the phoneme distribution in 95 languages to find the equation that best fits the distribution of the frequency of the occurrence of phonemes in these languages. For Arabic they used a sample containing 23727 phonemes. Previous studies did not address allophones and allosyllables, and the corpora used had no more than 50000 words, so the results of statistics cannot be generalized on Standard Arabic as a whole.

3. DATA PREPARATION

It is well known in statistics that the larger the sample is in size and variation, the more accurate is the study, and the more reliable the results can be generalized. The corpus we used in this work contains more than 5 million words and has been taken from both CA and MSA. It was also fully vocalized and without errors in writing. Almost all MSA texts are not vocalized [10], and the best automatic vocalization systems of Arabic text do not give the required accuracy we need in our study. For the aforementioned reasons, all non-vocalized or partly-vocalized texts have been vocalized, and the correct writing of all texts in the corpus has been checked manually. The large corpus we prepared consists of ten subcorpora:

1) Corpus1: the Holy Qur'an [11] (the main book of Islam)
2) Corpus2: the Van Dyck Version of the Holy Bible [12] (the main book of Christianity)
3) Corpus3: Nahj al-Balagha (Way of Eloquence) (in Arabic: "نَهجُ البَلَاغَة") [13], a book that contains sermons, letters, and sayings of Imam Ali ibn Abi Talib
4) Corpus4: Sahih al-Bukhari (in Arabic: "صَحِيحُ البُخَارِي") [14], the book that contains the most correct prophetic traditions
5) Corpus5: a book entitled Nihayet al Rutba fi Talab al-Hisba (in Arabic: " نِهَايَةُ الرُّتبَةِ فِي طَلَبِ الحِسبَة") [15] about Hisbah (business accountability) in Islam
6) Corpus6: a book, written by Ibn Hajar al-Haytami, entitled Tuhfatu'l Muhtaj fi Sharh Al-Minhaj, (in Arabic: "تُحفَةُ المُحتَاج فِي شَرحِ المِنهَاج") [16].
7) Corpus7: developed by King Abdul-Aziz City for Science and Technology (KACST), Saudi-Arabia, and contains 367 fully diacritized sentences with rich phonetical information content. Each sentence consists of 2 to 9 words. More information about this corpus is given in [17].
8) Corpus8: written by Katharina Bobzin; it contains 306 sentences selected from her book [Arabic Basic Course] (in German: "Arabisch Grundkurs") [18], which is a beginners' book for German speakers learning Arabic.
9) Corpus9: the Umm Alqura list of most frequently used Arabic words [19], prepared by Institute of Arabic Studies in the University of Umm Alqura. This list contains 5446 words.
10) Corpus10: contains texts taken from Arab news sites on the internet as well as stories and poems written in Standard Arabic. This corpus has 21618 words.

Additional information about the large corpus is given in Table 1.

all words	5,348,714
vocabulary	278,377
syllables	14,623,679
phonemes	33,250,270

TABLE 1: Information and statistics about the large corpus.

4. AUTOMATIC PHONETIZATION AND PRE-PROCESSING

Phonetization is the transcription from written text to sound symbols. A rule-based approach has been adopted, where expert linguists defined letter-to-sound language-dependent rules at both phonemic and phonetic level, and a lexicon of exceptions. Standard Arabic is known to have a clear correspondence between orthography and the phonetic system. Therefore, in this work, we adopted a rule-based approach for the transcription. At the phonemic level, we adopted the Arpabet coding [20] developed by the Advanced Research Projects Agency (ARPA). Table 2 shows the phonemic transcription of Arabic consonants, vowels, and diphthongs using both Arpabet and International Phonetic Alphabet (IPA) encodings.

cons.	Arpa.	IPA	cons.	Arpa.	IPA	cons.	Arpa.	IPA	cons.	Arpa.	IPA
ئ ,ؤ ,إ ,أ ,ء	E	ʔ	د	D	d	ض	DD	dˤ	ك	K	k
ب	B	b	ذ	DH	ð	ط	TT	tˤ	ل	L	l
ة ,ت	T	t	ر	R	r	ظ	DH2	ðˤ	م	M	m

ث	TH	θ	ز	Z	z	ع	AI	ʕ	ن	N	n
ج	JH	g	س	S	s	غ	GH	ɣ	ه	H	h
ح	HH	ħ	ش	SH	ʃ	ف	F	f	و	W	w
خ	KH	x	ص	SS	sˤ	ق	Q	q	ي	Y	y
short vowel	**Arpa.**		**IPA**	**long vowel**		**Arpa.**		**IPA**	**diphthong**	**Arpa.**	**IPA**
َ	AE		ɑ	َا, َى		AE:		ɑː	َوْ	AW	aw
ُ	UH		u	ُوْ		UW		uː	َيْ	AY	ay
ِ	IH		i	ِيْ		IY		iː			

TABLE 2: Arpabet and IPA encoding of the Arabic phonemes (cons.: consonant, Arpa.: Arpabet).

The encoding of the allophones, presented by our group in [3], has been adopted at the allophonic level: we extended the encoding developed by Alghamdi and others [21] in order to cover all phonetic variations of Arabic phonemes. Our encoding consists of two characters and three digits. The first two characters represent the phoneme, the first digit represents if the allophone is pharyngealized or not, the second digit symbolizes sound duration, and the last digit represents all phonetic variations except the features related to pharyngealization. Tables 3 and 4 explain this encoding.

position	**value interpretation**		
1, 2 (phoneme)	two letters represent a phoneme (cf. Table 4)		
3 (pharyngealization)	0 (not pharyngealized)	1 (pharyngealized)	
4 (duration)	0 (short)	1 (shorter)	2 (long)
5 (phonetic variations)	0 (the sound is a phoneme)	1 (allophone for "ن")	
	2 (released with a schwa)	3 (allophone for "ض" followed by "ط")	
	4 (allophone for "ت" followed by "د")	5 (allophone for "د" followed by "ت")	
	6 (allophone for "ل" followed by "ر")	7 (allophone for "ق" followed by "ك")	
	8 (allophone for "ط" followed by "ت")	9 (allophone for /ُ/ or /ِ/)	

TABLE 3: The modified encoding of the allophones at the phonetic level.

Arpabet phoneme	E	B	T	TH	JH	HH	KH	D	DH	R	Z
representing letters	hz	bs	ts	vs	jb	hb	xs	ds	vb	rs	zs
Arpabet phoneme	S	SH	SS	DD	TT	DH2	AI	GH	F	Q	K
representing letters	ss	js	sb	db	tb	zb	cs	gs	fs	qs	ks
Arpabet phoneme	L	M	N	H	W	Y	AE	UH	IH	AW	AY
representing letters	ls	ms	ns	hs	ws	ys	as	us	is	aw	ay

TABLE 4: Positions 1 and 2 in the encoding at the phonetic level.

Table 5 shows some examples of the encoding at the phonetic level.

text	سَامِحني /saːmiħniː/ (forgive me)
encoding	ss000 as020 ms000 is010 hb000 ns000 is020
text	بَدَأتُ أتَعَلَّمُ العَرَبِيَّةَ /badaʔtuʔataʕallamulʕarabiyyah/ (I began to learn Arabic)
encoding	bs000 as000 ds000 as010 hz000 ts000 us010 hz000 as000 ts000 as000 cs000 as010 ls000 ls000 as000 ms000 us010 ls000 cs000 as100 rs100 as100 bs000 is010 ys000 ys000 as000 hs000
text	شُكرًا لَا أدَخِّن /ʃukranlaːʔudaxxin/ (thank you I do not smoke)
encoding	js000 us010 ks000 rs100 as110 ls001 ls000 as020 hz000 us000 ds000 as110 xs000 xs000 is009 ns000
text	هُم مِنَ الجَزَائِر /humminalgazaːʔir/ (they are from Algeria)
encoding	hs000 us019 ms000 ms000 is000 ns000 as010 ls000 jb000 as000 zs000 as020 hz000 is009 rs000

TABLE 5: Examples of the encoding at the phonetic level.

At the syllable level, we categorized the syllables into eight types as follows: CV, CD2, CL, CVC, CD2C, CLC, CVCC, and CLCC, where C refers to a consonant, V to a short vowel, L to a long vowel, and D2 is a diphthong.

We implemented the automatic phonetization and syllabication [3] in Matlab. A pre-processing has been done to get the frequency of occurrence of the single phonemes, allophones, syllables, and allosyllables in descending order. Then we determined the equation that best fit the distribution at the four levels. From the phonemic transcription, we obtained the transition matrix between phonemes, as well as we counted the number of syllables of each type. The most important results will be presented and discussed in the next section.

5. EVALUATION

5.1. At Phoneme Level

Table 6 shows the frequency and percentage of the phonemes (vowels, consonants, and diphthongs). The percentage of the consonants, classified by the place and way of articulation, is presented in Tables 7, 8 respectively. Frequencies and percentages are in descending order. The frequency of linguistic units has a major role in corpus linguistics [5]. Table 9 explains the fit of the equations (1) to (6), described by the Goodness of Fitting (GoF) parameters with other relevant information, to the ranked frequency distribution of phonemes. The transition probability matrix between phonemes is given in Table 10.

rank	phoneme	frequency	perc. (%)	rank	phoneme	frequency	perc. (%)
1	AE	5,726,982	17.22	19	IY	488,655	1.47
2	IH	3,436,070	10.33	20	K	469,113	1.41
3	L	2,590,992	7.79	21	HH	408,852	1.23
4	UH	2,182,079	6.56	22	S	381,182	1.15
5	N	1,879,210	5.65	23	AW	359,329	1.08
6	AE:	1,863,610	5.60	24	AY	330,687	0.99
7	M	1,421,216	4.27	25	JH	280,159	0.84
8	H	1,213,770	3.65	26	SS	251,751	0.76
9	E	1,212,534	3.65	27	UW	244,910	0.74
10	T	1,024,767	3.08	28	DH	231,325	0.70
11	R	1,014,625	3.05	29	KH	226,339	0.68
12	B	846,222	2.55	30	SH	206,646	0.62
13	W	773,620	2.33	31	TH	172,156	0.52
14	AI	752,534	2.26	32	DD	159,119	0.48
15	Q	697,179	2.10	33	TT	151,267	0.45
16	Y	688,572	2.07	34	Z	127,841	0.38
17	F	669,909	2.01	35	GH	121,356	0.37
18	D	579,637	1.74	36	DH2	66,055	0.20

TABLE 6: Frequencies and percentages of the phonemes in descending order (perc.: percentage)

Based on Table 6, we conclude the following:

- Regarding the short vowels, “AE” is most frequent (≈ 17.22 %), followed by “IH” (≈ 10.33 %) and “UH” (≈ 6.56 %), respectively.
- Regarding the long vowels, “AE:” is most frequent (≈ 5.60 %), followed by “IY” (≈ 1.47 %) and “UW” (≈ 0.74 %), respectively.
- The difference between the percentages of the two diphthongs is very small (≈ 0.09 %), with (≈ 1.08 %) for “AW” and (≈ 0.99 %) for “AY”.
- Regarding the consonants, “L” is most frequent (≈ 7.79 %), followed by “N” (≈ 5.65 %) and “M” (≈ 4.27 %), respectively.
- “DD”, “TT”, “Z”, “GH”, and “DH2” occur with a percentage of below 0.5 %, and “DH2” is the least frequent phoneme in Standard Arabic (≈ 0.2 %).

- Although Standard Arabic contains only three short vowels, three long vowels, and two diphthongs, they occur with a percentage of (≈ 44 %). The remaining (≈ 56 %) are shared by the 28 consonants altogether.

place of articulation	consonants	percentage (with respect to consonants only) (%)
alveodental	T, D, R, Z, S, SS, DD, TT, L, N	43.83
glottal	E, H	13.03
bilabial	B, M	12.18
pharyngeal	HH, AI	6.24
uvular	KH, GH, Q	5.61
labiovelar	W	4.16
palatal	Y	3.70
labiodental	F	3.60
alveopalatal	JH, SH	2.61
interdental	TH, DH, DH2	2.52
velar	K	2.52

TABLE 7: Percentages of the consonant classes according to the place of articulation.

way of articulation	consonants	percentage (with respect to consonants only) (%)
stop	E, B, T, D, Q, K	25.94
fricative	TH, DH, HH, KH,Z, S, SH,AI, GH,F,H	24.23
nasal	M, N	17.73
lateral	L	13.92
glide	W, Y	7.85
trill	R	5.45
emphatic fricative	SS, DH2	1.71
emphatic stop	DD, TT	1.67
affricate	JH	1.50

TABLE 8: Percentages of the consonant classes according to the way of articulation.

fit equation	SSE	R-square	Adj. R-sq	RMSE	Variables values	number of variables
1	0.0319	0.9765	0.9758	0.0307	a = 1.0360, b = 0.7961	2
2	0.0061	0.9955	0.9952	0.0136	a = 1.0230, b = 0.5802, c = 0.9598	3
3	1.0993	0.1901	0.1901	0.1772	a=0.3550	1
4	No fitting					1
5	0.1165	0.9141	0.9116	0.0585	a = 0.9277, b = -0.1666	2
6	0.0038	0.9972	0.9970	0.0109	a =1.5020, b = -1.0190, c = 0.5000, d = -0.0934	4

TABLE 9: Fit of Equations (1) to (6) to the ranked frequency distribution of phonemes.

The Goodness of Fit parameters we used are:

- The sum of squared errors (SSE). A value closer to 0 points out that the fit will be more useful for prediction because of the smaller random error component in the model.
- R-square $\in$ [0, 1], with a value closer to 1 indicating that a greater portion of variance is accounted by the model.
- Adjusted R-square (Adj. R-sq) ≤ 1, with a value closer to 1 indicating a better fit.
- Root mean squared error (RMSE). It is an assessment of the standard deviation of the random part in the data. A value closer to 0 demonstrates a fit that is optimal for prediction.

More details about GoF parameters can be found in [22]. We tested different types of coordinates in terms of suitability for fitting and found that the best case is when the horizontal axis represents the rank of the phonemes and the vertical axis represents the normalized frequencies of the phonemes with regard to the maximum frequency or the sum of all frequencies. We received the same result with allophones, syllables, and allosyllables. The normalized frequencies with respect to the maximum frequency was adopted in this paper at all levels.

From Table 9, we found the following:

- Exponential equation (6) best fits the ranked frequency distribution of phonemes in Standard Arabic. This is contrary to the result reached by Tambovtsev and Martindale in [9] that Borodovsky and Gusein-Zade equation (4) provides the best fitting at this level.
- Yule's equation (2) gives slightly better fit than Zipf's equation (1), but Zipf's equation has the advantage that there are only two variables to be estimated versus three for Yule's equation.

Figure 1 shows the best fitting equation which is the exponential equation (6). The corresponding phonemes to the ranks are given in Table 6.

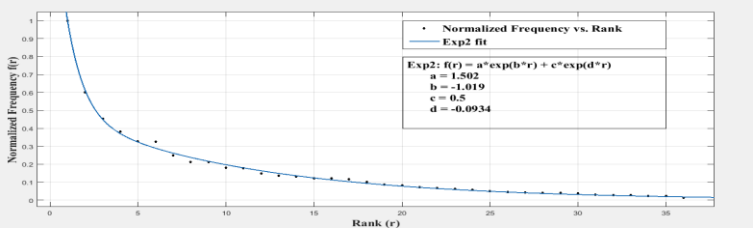

FIGURE 1: Fit of the exponential equation (6) to the ranked frequency distribution of phonemes.

	AE	IH	L	UH	N	AE:	M	H	E	T	R	B	W	AI	Q	Y	F	D
AE	0	0	0.1491	0	0.0967	0	0.0751	0.0576	0.0495	0.0913	0.0593	0.0459	0.0188	0.0485	0.0422	0.0148	0.0296	0.0353
IH	0	0	0.1607	0	0.1396	0	0.0582	0.0897	0.0625	0.017	0.0464	0.0415	0.0572	0.0324	0.0296	0.0531	0.0368	0.0211
L	0.1821	0.16	0.0823	0.097	0.0011	0.1609	0.0707	0.0031	0.0478	0.0044	0.0001	0.0145	0.0127	0.0169	0.016	0.008	0.0097	0.0004
UH	0	0	0.1179	0	0.1004	0	0.0889	0.1153	0.0671	0.0217	0.0382	0.0524	0.0743	0.05	0.0322	0.0112	0.0441	0.0146
N	0.2387	0.1069	0.0356	0.0356	0.1201	0.0534	0.0274	0.0242	0.0481	0.0183	0.0038	0.0232	0.0488	0.0261	0.0229	0.0204	0.0318	0.0092
AE:	0	0	0.1236	0	0.0644	0	0.0658	0.0577	0.1276	0.0412	0.0514	0.051	0.0327	0.0292	0.0359	0.0613	0.0498	0.0302
M	0.2788	0.1759	0.0114	0.1674	0.0047	0.1663	0.0605	0.0018	0.0084	0.0102	0.009	0.0034	0.0068	0.0051	0.0019	0.0292	0.0047	0.0014
H	0.0479	0.3106	0.006	0.4564	0.0016	0.1466	0.0017	0.0023	0.001	0.0006	0.0066	0.0007	0.0015	0.0001	0.0001	0.0008	0.0005	0.0019
E	0.39	0.3231	0.0003	0.0398	0.0022	0.0712	0.003	0.0001	0.0001	0.0105	0.0005	0.0004	0.0005	0.0001	0.0001	0.0014	0.0001	0.0002
T	0.3704	0.2792	0.0056	0.1553	0.0153	0.0319	0.0034	0.0053	0.0043	0.0649	0.0024	0.0048	0.0022	0.0028	0.0035	0.0011	0.0028	0.0003
R	0.3062	0.2327	0.0008	0.1125	0.0017	0.0743	0.0045	0.0029	0.0041	0.007	0.0958	0.0091	0.002	0.0064	0.0061	0.0011	0.0054	0.0065
B	0.2301	0.3798	0.0173	0.0787	0.0411	0.0762	0.0002	0.0028	0.0005	0.0058	0.0086	0.029	0.001	0.0061	0.0029	0.0012	0.0002	0.0146
W	0.8443	0.0333	0.0001	0.0291	0	0.0756	0	0	0.0001	0	0.0001	0	0.0049	0.0002	0.0001	0	0.0001	0
AI	0.5062	0.1366	0.0248	0.0895	0.0115	0.0487	0.0096	0.0016	0.0006	0.0333	0.009	0.0082	0.004	0.0014	0.0025	0.003	0.0027	0.028
Q	0.2484	0.0951	0.0056	0.0761	0.0009	0.1459	0.0005	0.0008	0.0003	0.0238	0.0116	0.0079	0.0024	0.001	0.0272	0.0032	0.0037	0.0148
Y	0.4742	0.0727	0	0.2062	0	0.0645	0	0	0	0	0	0	0	0	0	0.1461	0	0
F	0.3952	0.2067	0.0034	0.0476	0.0016	0.0467	0.0002	0.005	0.0004	0.0075	0.0087	0.0002	0.0021	0.0127	0.0013	0.0038	0.0115	0.0002
D	0.3407	0.1833	0.0048	0.1291	0.0026	0.049	0.0031	0.0011	0.0043	0.0048	0.0145	0.0048	0.0027	0.0052	0.0028	0.0172	0.0032	0.1447
IY	0	0	0.0755	0	0.1056	0	0.0951	0.1103	0.0688	0.0152	0.0709	0.0432	0.0404	0.0497	0.0626	0.0139	0.0392	0.0494
K	0.444	0.1085	0.0043	0.1343	0.0032	0.1206	0.0165	0.0009	0.0024	0.0078	0.0226	0.0043	0.0016	0.0044	0.0002	0.0003	0.0083	0.0002
HH	0.3665	0.1509	0.0063	0.1066	0.0053	0.0815	0.0178	0.0002	0.0002	0.0346	0.0245	0.0031	0.0326	0.0001	0.0031	0.0073	0.0034	0.0089
S	0.2981	0.1193	0.0346	0.0828	0.0108	0.076	0.0254	0.0011	0.0111	0.0956	0.0166	0.0141	0.0042	0.0035	0.0096	0.0011	0.0016	0.0002
AW	0	0	0.5271	0	0.0341	0	0.0548	0.0037	0.0312	0.0255	0.0134	0.0303	0.0669	0.0165	0.0276	0.008	0.0166	0.0072
AY	0	0	0.0349	0	0.1342	0	0.0309	0.1692	0.0432	0.0368	0.123	0.032	0.0256	0.0297	0.0071	0.1234	0.0382	0.0192
JH	0.3155	0.1766	0.0164	0.0979	0.0126	0.1013	0.0179	0.0311	0.0007	0.0142	0.0247	0.0051	0.0017	0.011	0	0.0003	0.0006	0.0057
SS	0.3389	0.1345	0.0474	0.0525	0.0048	0.0795	0.0032	0.0002	0.0001	0.0003	0.0279	0.0091	0.0038	0.0027	0.0002	0.0009	0.0168	0.0183
UW	0	0	0.1565	0	0.1848	0	0.0547	0.0321	0.0312	0.0214	0.0955	0.0731	0.0093	0.0639	0.015	0.0077	0.0323	0.0625
DH	0.1709	0.1173	0.0138	0.0524	0.019	0.4105	0.0022	0.0121	0.0014	0.0018	0.027	0.005	0.0005	0.0008	0.0018	0.0016	0.0062	0.0001
KH	0.2138	0.3847	0.0124	0.0848	0.0004	0.0635	0.0014	0	0.0005	0.045	0.0362	0.0267	0.0044	0	0	0.0016	0.0099	0.0031
SH	0.303	0.0813	0.0001	0.0546	0.0003	0.1065	0.0077	0.0155	0.0003	0.0601	0.0315	0.0068	0.0015	0.0049	0.0011	0.0041	0.0027	0.003
TH	0.3393	0.0744	0.0541	0.213	0.0272	0.095	0.0102	0.0004	0.0005	0.0012	0.0053	0.016	0.0007	0.0001	0.0019	0.0002	0.0003	0
DD	0.3119	0.2961	0.0126	0.0885	0.001	0.0991	0.0156	0.001	0.0004	0.0019	0.0092	0.0023	0.0038	0.0131	0.0001	0.0013	0.0003	0.0003
TT	0.2718	0.1487	0.0991	0.0947	0.0067	0.0924	0.0032	0.0037	0.0183	0.0013	0.0154	0.0103	0.0083	0.0314	0.0033	0.0013	0.0295	0.0003
Z	0.2726	0.1826	0.0061	0.0943	0.0084	0.0995	0.0133	0.0017	0.011	0.001	0.0051	0.001	0.0103	0.0049	0.0019	0.0052	0.0005	0.0031
GH	0.171	0.0428	0.0145	0.047	0.2287	0.0725	0.0046	0.0006	0.0002	0.0129	0.0166	0.0029	0.003	0.0001	0.0001	0.0026	0.0071	0.0014
DH2	0.2828	0.0775	0.0058	0.1127	0.0012	0.2394	0.0062	0.0948	0.0002	0.0016	0.0009	0.0002	0.0004	0.0007	0.0001	0.0001	0.0017	0

	IY	K	HH	S	AW	AY	JH	SS	UW	DH	KH	SH	TH	DD	TT	Z	GH	DH2
AE	0	0.0236	0.0236	0.0225	0	0	0.0183	0.0163	0	0.014	0.01	0.0116	0.0098	0.0094	0.009	0.0063	0.0058	0.0059
IH	0	0.0308	0.017	0.0193	0	0	0.0074	0.0124	0	0.0173	0.0073	0.0107	0.0107	0.0043	0.0058	0.005	0.0041	0.0019
L	0.0088	0.0099	0.0171	0.0003	0.0149	0.0294	0.01	0.0001	0.0068	0.0001	0.0071	0.0001	0.0001	0	0.0004	0.0025	0.0047	0
UH	0	0.0263	0.0175	0.0225	0	0	0.0181	0.0141	0	0.0066	0.01	0.0126	0.0059	0.0052	0.0089	0.0053	0.0171	0.0014
N	0.0289	0.0176	0.0081	0.0061	0.0011	0.0014	0.0044	0.0042	0.006	0.0029	0.0028	0.0038	0.0053	0.001	0.0022	0.0033	0.0035	0.0029
AE:	0	0.022	0.0189	0.0146	0	0	0.0133	0.0079	0	0.0242	0.0384	0.0064	0.0064	0.007	0.0062	0.0082	0.0033	0.0016
M	0.0141	0.0056	0.001	0.0045	0.0065	0.0041	0.0004	0.0004	0.0133	0.0002	0.0004	0.0007	0.0008	0.0003	0.0002	0.0005	0.0002	0.0001
H	0.003	0	0	0	0.0001	0.0016	0.0005	0	0.0073	0.0002	0	0	0	0	0.0001	0.0002	0	0
E	0.008	0.0016	0	0.0031	0.0637	0.0668	0.0015	0.0001	0.0053	0.0013	0.0046	0	0.0005	0	0	0	0	0
T	0.0232	0.0011	0.0026	0.0004	0.0043	0.0068	0.0003	0.0006	0.0033	0.0002	0.0003	0.0004	0.0003	0.0001	0.0002	0.0002	0.0002	0
R	0.0324	0.0058	0.0182	0.0031	0.0091	0.0043	0.0051	0.0005	0.0208	0.0001	0.0002	0.0024	0.0009	0.0115	0.0055	0.0008	0.0003	0
B	0.0281	0.0022	0.0029	0.0018	0.0007	0.0417	0	0.0007	0.0161	0.0002	0.0002	0.0005	0.0001	0.0029	0.0045	0.0008	0.0007	0
W	0.0077	0	0.0001	0.0001	0.0001	0.0017	0	0	0.0022	0	0	0	0	0	0	0	0	0
AI	0.0166	0.0003	0.0001	0.0013	0.0044	0.0176	0.0018	0.0016	0.011	0.0009	0	0.0003	0.0005	0.0134	0.0064	0.0012	0	0.0012
Q	0.0304	0.0001	0.0001	0.0015	0.2505	0.0065	0	0.0079	0.0287	0.0003	0	0.0002	0	0.002	0.0026	0	0	0
Y	0.0096	0	0	0	0.0111	0.0015	0	0	0.0138	0	0	0	0	0	0	0	0	0
F	0.2026	0.0003	0.0004	0.015	0.0045	0.0024	0	0.0036	0.0056	0	0.0002	0.0001	0	0.0037	0.001	0.0001	0	0.0057
D	0.0345	0.0009	0.0009	0.0013	0.001	0.0101	0.0009	0.001	0.0229	0.0004	0.0065	0.0004	0.0002	0.0002	0.0002	0.0002	0.0005	0.0001
IY	0	0.0196	0.0363	0.0125	0	0	0.0113	0.0088	0	0.0115	0.0047	0.0181	0.009	0.0056	0.0052	0.0074	0.0088	0.0014
K	0.013	0.0228	0.0002	0.004	0.0138	0.0081	0	0.0001	0.043	0.0006	0	0.0005	0.0092	0	0.0001	0.0001	0	0
HH	0.0336	0.0057	0.0427	0.0059	0.0044	0.0274	0.0027	0.0076	0.0062	0.0018	0	0.0014	0.0027	0.0037	0.0001	0.0006	0	0.0008
S	0.015	0.0056	0.0047	0.0931	0.0024	0.0232	0.0087	0	0.0326	0	0.0054	0	0	0	0.0036	0	0	0
AW	0	0.0184	0.008	0.0071	0	0	0.0346	0.0078	0	0.0015	0.0039	0.0058	0.0028	0.0307	0.0039	0.0042	0.0071	0.0014
AY	0	0.0196	0.0056	0.036	0	0	0.0022	0.0016	0	0.0009	0.0197	0.004	0.0259	0.0311	0.0028	0.0006	0.0017	0.0007
JH	0.0191	0	0.0005	0.0007	0.0051	0.0103	0.0431	0	0.0761	0.0002	0	0.0001	0.0001	0	0	0.0115	0	0
SS	0.0347	0	0.0157	0	0.0195	0.0068	0	0.1324	0.0455	0	0	0	0	0	0.0028	0	0.0014	0
UW	0	0.011	0.0139	0.0309	0	0	0.0224	0.0123	0	0.0091	0.0037	0.005	0.0043	0.0052	0.0143	0.023	0.0046	0.0005
DH	0.0689	0.0294	0.0004	0.0002	0.0006	0.0036	0.0004	0.0002	0.0116	0.0387	0.0005	0.0001	0.0001	0	0.0001	0	0.0004	0.0001
KH	0.0178	0	0	0.0005	0.0055	0.0204	0.0003	0.01	0.0172	0.0181	0.0142	0.0022	0	0.0007	0.0037	0.0008	0	0
SH	0.033	0.0155	0.0003	0	0.0049	0.0739	0.0007	0	0.0089	0	0.0005	0.176	0	0	0.0005	0	0.0007	0
TH	0.0269	0.0004	0	0	0.0121	0.0021	0	0	0.0054	0	0.0002	0.0001	0.1127	0	0	0	0.0003	0
DD	0.0444	0.0001	0.0048	0	0.0016	0.0052	0.0007	0	0.0232	0	0.0003	0	0	0.0529	0.0074	0	0.0008	0
TT	0.0262	0.0013	0.002	0.0002	0.0055	0.0086	0.0001	0.0001	0.0256	0.0001	0.0002	0.0008	0.0002	0.0001	0.0891	0	0.0002	0
Z	0.0376	0.0011	0.0004	0.0001	0.0628	0.0202	0.0005	0	0.0213	0.0001	0	0.0001	0	0	0	0.133	0.0001	0
GH	0.0561	0	0	0.0041	0.002	0.2929	0	0.0061	0.0046	0.0002	0	0.0024	0.0001	0.0013	0.0005	0.0005	0.0005	0
DH2	0.059	0.0001	0.0001	0	0	0.0006	0	0	0.0057	0.0001	0	0	0	0	0	0	0.0001	0.1078

TABLE 10: The transition probability matrix between phonemes.

The transition probability matrix is an important component in Hidden Markov Model (HMM)-based state-of-the-art phoneme recognition systems.

5.2. At Allophone Level

Table 11 shows the phonemes and the corresponding allophones in Standard Arabic, Table 12 represents the frequency and percentage of the allophones, and Table 13 explains the fit of the equations (1) to (6) to the ranked frequency distribution of allophones.

phoneme	corresponding allophones	phoneme	corresponding allophones
E	hz000, hz002	GH	gs000
B	bs000, bs002	F	fs000, fs001
T	ts000, ts001, ts002, ts005, ts008, ts100	Q	qs000, qs001, qs002
TH	vs000, vs001	K	ks000, ks001, ks002, ks007
JH	jb000, jb001, jb002	L	ls000, ls001, ls100
HH	hb000	M	ms000, ms001
KH	xs000	N	ns000
D	ds000, ds001, ds002, ds004, ds100	H	hs000
DH	vb000, vb001, vb100	W	ws000, ws001
R	rs000, rs001, rs006, rs100	Y	ys000, ys001
Z	zs000, zs001	AE	as000, as010, as100, as110
S	ss000, ss001, ss100	UH	us000, us009, us010, us019, us100, us109, us110,us119
SH	js000, js001	IH	is000, is009, is010, is019
SS	sb000, sb001	AE:	as020, as120
DD	db000, db001	UW	us020, us120
TT	tb000, tb001, tb002, tb003	IY	is020
DH2	zb000, zb001	AW	aw000, aw100
AI	cs000	AY	ay000, ay100

TABLE 11: The phonemes and their corresponding allophones in Standard Arabic.

Based on Table 11, we conclude the following:

- There are 93 allophones in Standard Arabic, 21 for the vowels, 4 for the diphthongs, and 68 for the consonants.
- 16 allophones occur for the short vowels, i.e. 4 allophones for “AE” (“as000”, “as010”, “as100”, “as110”), 4 allophones for “IH” (“is000”, “is010”, “is009”, “is019”), and 8

allophones for “UH” (“us000”, “us010”, “us100”, “us110”, “us009”, “us019”, “us109”, “us119”).

- There are 5 allophones for the long vowels, i.e. 2 allophones for “AE:” (“as020”, “as120”), 2 allophones for “UW” (“us020”, “us120”), and 1 allophone for “IY” (“is020”).
- 4 allophones occur for the diphthongs, i.e. 2 allophones for “AW” (“aw000”, “aw100”), and 2 allophone for “AY” (“ay000”, “ay100”).
- There are 7 phonemes in Standard Arabic that do not have phonetic variations, they are: “HH”, “KH”, “AI”, “GH”, “N”, “H”, and “IY”.

alloph.	freq.	perc. (%)	alloph.	freq.	Perc. (%)	alloph.	freq.	perc. (%)
ls000	2,511,018	7.55	ay000	259,128	0.78	ds001	17,142	0.05
as000	2,164,903	6.51	sb000	251,751	0.76	is009	15,825	0.05
as010	2,008,683	6.04	vb000	230,570	0.69	ss001	11,495	0.03
is010	1,850,217	5.56	xs000	226,339	0.68	vs001	9,818	0.03
ms000	1,421,216	4.27	js000	206,646	0.62	jb001	8,243	0.02
as020	1,379,567	4.15	aw100	200,085	0.60	sb001	7,842	0.02
is000	1,378,893	4.15	is019	191,135	0.57	ss100	7,593	0.02
ns000	1,331,816	4.01	vs000	172,156	0.52	js001	7,153	0.02
hs000	1,213,770	3.65	us100	170,005	0.51	rs001	7,121	0.02
hz000	1,210,487	3.64	us020	159,490	0.48	zs001	6,117	0.02
us010	1,031,091	3.10	aw000	159,201	0.48	zb001	5,396	0.02
ts000	967,527	2.91	db000	157,943	0.48	vb001	5,344	0.02
bs000	843,195	2.54	tb000	150,533	0.45	tb001	4,094	0.01
as100	812,447	2.44	us019	148,377	0.45	us009	3,671	0.01
ws000	773,663	2.33	zs000	127,841	0.38	ds002	3,484	0.01
cs000	752,534	2.26	gs000	121,356	0.37	bs002	3,027	0.01
as110	741,005	2.23	ms001	94,664	0.28	ts005	2,679	0.01
rs100	718,262	2.16	ws001	88,376	0.27	ks002	2,543	0.01
qs000	695,590	2.09	us120	85,420	0.26	hz002	2,047	0.01
ys000	688,588	2.07	ls100	79,718	0.24	db001	1,892	0.01
fs000	669,909	2.01	ay100	71,546	0.22	qs002	1,573	0.00
ds000	551,471	1.66	ls001	66,615	0.20	ts002	1,443	0.00
us000	501,455	1.51	zb000	66,055	0.20	tb003	1,176	0.00
is020	488,655	1.47	fs001	59,506	0.18	us109	1,052	0.00
as120	483,869	1.46	ts100	55,502	0.17	jb002	824	0.00
ks000	466,570	1.40	us119	44,152	0.13	vb100	755	0.00
hb000	408,852	1.23	qs001	42,764	0.13	tb002	587	0.00
ss000	373,589	1.12	ys001	36,690	0.11	ds004	295	0.00
rs000	296,363	0.89	ts001	34,212	0.10	rs006	256	0.00
us110	282,276	0.85	ks001	32,910	0.10	ts008	147	0.00
jb000	279,335	0.84	ds100	22,003	0.07	ks007	16	0.00

TABLE 12: Frequencies and percentages of the allophones.

Based on Table 12, we found the following:

- We realize the phonetic variations in Standard Arabic and how often they occur in the language. For example, from the sum of the percentages of allophones whose first digit of the encoding is equal to 1 (≈ 11.36 %), we conclude that the phenomenon of pharyngealization is very common in Standard Arabic.
- Regarding the allophones for the short vowels, “as000” is most frequent (≈ 6.51 %), followed by “as010” (≈ 6.04 %) and “is010” (≈ 5.56 %), respectively.
- Regarding the allophones for the long vowels, “as020” is most frequent (≈ 4.15 %), followed by “is020” (≈ 1.47 %) and “as120” (≈ 1.46 %), respectively.
- The allophone “ay000” for the diphthong “AY” is most frequent (≈ 0.78 %) followed by “aw100” (≈ 0.60 %), “aw000” (≈ 0.48 %), and “ay100” (≈ 0.22 %), respectively.
- Regarding the allophones for the consonants, “ls000” is most frequent (≈ 7.55 %), followed by “ms000” (≈ 4.27 %) and “ns000” (≈ 4.01 %), respectively.

- The allophone “ks007” for the phoneme “K” is the least frequent allophone in Standard Arabic. It occurs only 16 times in the large corpus which has 33,250,270 phonemes.

fit equation	SSE	R-square	Adj. R-sq	RMSE	Variables values	number of variables
1	0.7347	0.8254	0.8235	0.0899	a = 1.3160, b = 0.6483	2
2	0.0238	0.9944	0.9942	0.0163	a = 1.0530, b = 0.1314, c = 0.9461	3
3	4.6281	-0.1000	-0.1000	0.2243	a = 0.5588	1
4	No fitting					1
5	0.0496	0.9882	0.9881	0.0233	a = 0.9536, b = -0.0701	2
6	0.0207	0.9951	0.9949	0.0152	a = 0.8241, b = -0.0619, c = 0.3791, d = -0.4976	4

TABLE 13: Fit of Equations (1) to (6) to the ranked frequency distribution of allophones.

Similarly to the fit of the ranked frequency distribution of phonemes from Table 13, we found the following:

- Like the result at the phonemic level, Exponential equation (6) best fits the ranked frequency distribution of allophones in Standard Arabic.
- With a very negligible difference from the exponential equation (6), the exponential equation (5) and Yule’s equation (2) offer an excellent fitting.

Figure 2 shows the best fitting equation which is the exponential equation (6).

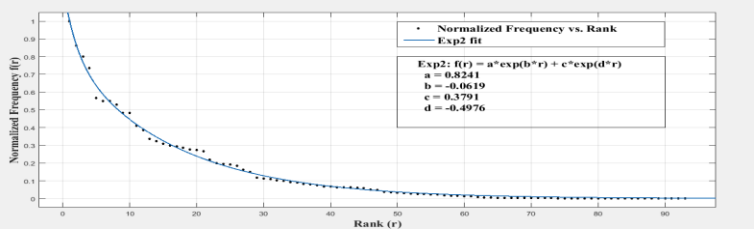

FIGURE 2: Fit of the exponential equation (6) to the ranked frequency distribution of allophones.

5.3. At Syllable Level

The importance of syllabication lies in the followings:

- Syllabication of a text to the corresponding syllables is very helpful for the correct pronunciation of this text.
- Stress in Standard Arabic depends on the syllable types in the word [23].
- The meters of Arabic poetry, as well as the phenomenon of pharyngealization, are syllable-based.

The syllabication has been done according to Algorithm 1. In this algorithm, “***P***” denotes the phonemic transcription of a given text. Moreover, “***S***” denotes the transcription we obtain after replacing consonants, short vowels, long vowels, and diphthongs in “***P***” with the symbols C, V, L, and D2, respectively.

1: find the number of syllables $\mathbf{N_s}$ which is the total number of occurrences of the symbols **V**, **L**, and **D2** in ***S***
2: **if** $\mathbf{N_s} > 1$ **then**
3: **while** ***l***(***S***) > 3 **do** # ***l***(***S***) (the length of ***S***)
4: **if** the fourth symbol in ***S*** is **C then**
5: take the first three phonemes in ***P*** as one syllable
6: take the first three symbols in ***S*** as the type of this syllable
7: remove the first three phonemes or symbols from ***P*** and ***S***
8: **else**
9: take the first two phonemes in ***P*** as one syllable
10: take the first two symbols in ***S*** as the type of this syllable
11: remove the first two phonemes or symbols from ***P*** and ***S***
12: **end if**
13: **end while**
14: take the phonemes in ***P*** as one syllable
15: take the symbols in ***S*** as the type of this syllable (1)
16: **else**
17: **do** (1)
18: **end if**

ALGORITHM 1: Syllabication Algorithm.

An example for the syllabification of the word "الشَّارِع" /ʔaʃʃaːriʕ/ (the street), according to Algorithm 1, is given in Table 14.

P	ء	ـَ	ش	ش	اَ	ر	ـِ	ع
***P* (Arpabet)**	E	AE	SH	SH	AE:	R	IH	AI
S	C	V	C	C	L	C	V	C
1st iteration	C	V	C	**C**				
1st syllable	E	AE	SH					
2nd iteration				C	L	C	**V**	
2nd syllable				SH	AE:			
3rd iteration						C	V	C
3rd syllable						R	IH	AI

TABLE 14: Syllabication of the word "الشَّارِع" using Algorithm 1.

After doing the automatic syllabication we found the following:

- The total number of unique syllables is 4921.
- The most frequent syllable in Standard Arabic is "وَ" "WAE" (≈ 3.60 %), followed by "هَ" "HAE" (≈ 3.25 %) and "لَا" "LAE:" (≈ 2.83 %), respectively.
- About 99.51 % of the syllables occur with a percentage of < 1 %, ≈ 96.57 % with < 0.1 %, ≈ 86.47 % with < 0.01 %, ≈ 70.84 % with < 0.001 %, ≈ 46.92 % with < 0.0001 %, and ≈ 16.03 % of the syllables in Standard Arabic occur with a percentage of < 0.00001 %.

The types of the syllables with their frequencies and percentages are given in Table 15, while the fit of the equations (1) to (6) to the ranked frequency distribution of syllables is presented in Table 16.

type of syllable	CV	CVC	CL	CD2	CLC	CVCC	CD2C	CLCC
frequency	7,415,763	3,904,533	2,563,124	685,515	32,269	18,384	3,912	99
percentage (%)	50.71	26.70	17.53	4.69	0.22	0.13	0.03	0.00

TABLE 15: Frequencies and percentages of the eight syllables types.

Regarding Table 15, we find the followings:

- The most frequent syllable type in Standard Arabic is “CV” (≈ 50.71 %), followed by “CVC” (≈ 26.70 %) and “CL” (≈ 17.53 %), respectively.
- “CLCC” is the least frequent syllable type (≈ 0.00 %). This is reasonable because the type occurs only when stopping at the end of a sentence.
- The first four types “CV”, “CVC”, “CL”, and “CD2” cover almost all syllables in Standard Arabic (> 99 %), while the remaining four ones include less than 0.5 %.

fit equation	SSE	R-square	Adj. R-sq	RMSE	variables values	number of variables
1	1.1250	0.8484	0.8483	0.0151	a = 1.4570, b = 0.6985	2
2	No fitting					3
3	6.1120	0.1760	0.1760	0.0352	a = 0.5374	1
4	No fitting					1
5	0.3448	0.9535	0.9535	0.0084	a = 0.7659, b = -0.0390	2
6	0.1003	0.9865	0.9865	0.0045	a = 0.1528, b = -0.0097, c = 0.7498, d = -0.0724	4

TABLE 16: Fit of Equations (1) to (6) to the ranked frequency distribution of syllables.

From Table 16, we found the following:

- With a very negligible increase from the exponential equation (5), Exponential equation (6) best fit the ranked frequency distribution of syllables in Standard Arabic.
- To some extent (R-square ≈ 0.85) Zipf’s equation (1) fits the ranked frequency distribution of syllables, while Yule’s equation (2) and Gusein-Zade’s equation (4) do not provide any fitting.

Figure 3 shows the best fitting equation which is the exponential equation (6).

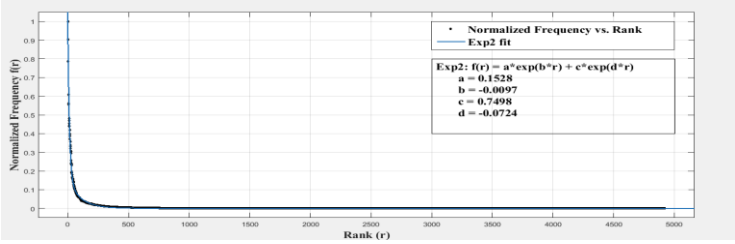

FIGURE 3: Fit of the exponential equation (6) to the ranked frequency distribution of syllables.

5.4. At Allosyllable Level

At this level we found the following:

- There are 10628 unique allosyllables in Standard Arabic.
- About 99.84 % of the allosyllables occur with a percentage of < 1 %, ≈ 98.13 % with < 0.1 %, ≈ 91.72 % with < 0.01 %, ≈ 77.71 % with < 0.001 %, ≈ 52.34 % with < 0.0001 %, and ≈ 18.62 % of the allosyllables in Standard Arabic occur with a percentage of < 0.00001 %.

The fit of the equations (1) to (6) to the ranked frequency distribution of allosyllables is given in Table 17.

fit equation	SSE	R-square	Adj. R-sq	RMSE	Variables values	number of variables
1	0.9659	0.8727	0.8727	0.0095	a = 1.4660, b = 0.6924	2
2	No fitting					3
3	6.2150	0.1808	0.1808	0.0242	a = 0.5483	1
4	No fitting					1
5	0.6321	0.9167	0.9167	0.0077	a = 0.7237, b = -0.0358	2
6	0.0912	0.9880	0.9880	0.0029	a = 0.7952, b = -0.0958, c = 0.1827, d = -0.0084	4

TABLE 17: Fit of Equations (1) to (6) to the ranked frequency distribution of allosyllables.

From Table 17, we found the following:

- Exponential equation (6) best fits the ranked frequency distribution of allosyllables in Standard Arabic.
- Similarly to the results of fitting at syllable level, the exponential equation (5) and Zipf's equation (1) fit the ranked frequency distribution of allosyllables rather well, while Yule's equation (2) does not provide any fitting.

Figure 4 shows the best fitting equation which is the exponential equation (6).

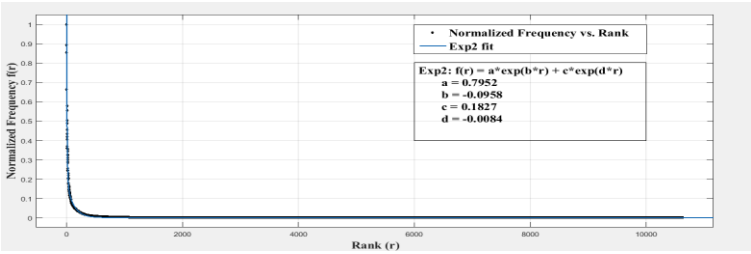

FIGURE 4: Fit of the exponential equation (6) to the ranked frequency distribution of allosyllables.

Determination of all the possible syllables in Standard Arabic at the phonemic and phonetic level is very important in any state-of-the-art Arabic CAPL system. We recorded the allosyllables we obtained in our work here and utilized them in a TTS system we developed for Standard Arabic pronunciation and spelling teaching. Due to the large number of unique syllables (4921) and allosyllables (10628), it is not able to present them with their frequencies here in this article, but all the transcriptions at phoneme, allophone, syllable, and allosyllable level, the frequency and percentage of unique syllables and allosyllables, and other outputs we obtained after doing the automatic transcription and preprocessing on the large corpus described in Section 3, will be available for research purposes upon request.

6. CONCLUSION

We accomplished a comprehensive automatic statistical study of Standard Arabic on four levels: phoneme, allophone, syllable, and allosyllable. The transcriptions we obtained in this work at the four previous levels can be utilized for machine learning-based automatic transcription of Standard Arabic texts on these four levels. The equations that best fit the ranked frequency distribution at these levels have been verified. We found that the exponential equation (6) provides the best fitting at all levels. A large corpus with more than 5 million words was used, and the phonetic transcription was fully automatic using a software package that we developed and achieved a very high accuracy (> 99 %) at both phonemic and phonetic level. The results of this work can be generalized to the Arabic language as a whole and can be used in Standard Arabic ASR, TTS, and CAPL systems.

7. REFERENCES

[1] A. Masmoudi, M. Ellouze Khemakhem, Y. Estève, L. Hadrich Belguith, N. Habash, “A corpus and phonetic dictionary for Tunisian Arabic speech recognition,” in: LREC, 2014, pp. 306–310.

[2] S. Harrat, M. Abbas, K. Meftouh, K. Smaili, “Diacritics restoration for Arabic dialects,” in: 14th Annual Conference of the International Speech Communication Association (Interspeech), 2013, pp. 1429–1433.

[3] F. Sindran, F. Mualla, T. Haderlein, K. Daqrouq, E. Nöth.G. “Rule-Based Standard Arabic Phonetization at Phoneme, Allophone, and Syllable Level.” International Journal of Computational Linguistics (IJCL), vol. 7, pp. 23-37, Dec. 2016.

[4] D.M.W. Powers, “Applications and explanations of Zipf's law”. Association for Computational Linguistics, 1998, pp. 151-160.

[5] A. Lüdeling, M. Kytö, Eds.: “Corpus linguistics: an international handbook”. Berlin, Mouton de Gruyter, 2008. Vol. 2, pp. 803-821.

[6] A. H. Moussa, [Computerization of the Arab heritage] (in Arabic: حَوسَبَةُ التُّرَاثِ العَرَبِي). Internet: http://majma.org.jo/res/seasons/19/19-1.pdf, [October 15, 2016].

[7] I. AbuSalim, [The syllabic structure in Arabic language] (in Arabic: البُنيَةُ المَقطَعِيَّةُ فِي اللُّغَةِ العَرَبِيَّة), Magazine of the Jordan Academy of Arabic 33 (1987), pp. 45–63.

[8] A. A.-R. A. Ibrahim, [The syllable system in Surat al-Baqara] (in Arabic), Master’s thesis, Arabic Department, Faculty of Arts, Islamic University Gaza, Palestine (2006).

[9] Y. Tambovtsev, C. Martindale, “Phoneme frequencies follow a yule distribution,” SKASE Journal of Theoretical Linguistics 4 (2007), pp. 1–11.

[10] M. Elshafei, H. Al-Muhtaseb, M. Alghamdi, “Statistical methods for automatic diacritization of Arabic text,” in: The Saudi 18th National Computer Conference. Riyadh, 2006.

[11] [Holy Qur’an] (in Arabic: “القُرأَنُ الكَرِيم”). [On-line]. Available: http://www.holyquran.net/quran/index.html [October 13, 2016].

[12] [Holy Bible] (in Arabic: “الكِتَابُ المُقَدَّس”). [On-line]. Available: http://ar.arabicbible.com/arabic-bible/word.html [October 13, 2016].

[13] S. Razi. [Nahj al-Balagha] (in Arabic: “نَهجُ البَلَاغَة”). [On-line]. Available: http://ia600306.us.archive.org/7/items/98472389432/nhj-blagh-ali.pdf [October 13, 2016].

[14] M. al-Bukhari. [Sahih al-Bukhari] (in Arabic: “صَحِيحُ البُخَارِي”). [On-line]. Available: http://shamela.ws/browse.php/book-1681 [October 13, 2016].

[15] A. al-Shaizari. [Nihayet al Rutba fi Talab al-Hisba] (in Arabic: “نِهَايَةُ الرُّتبَةِ فِي طَلَبِ الحِسبَة”). [On-line]. Available: http://shamela.ws/browse.php/book-21584 [October 13, 2016].

[16] I. al-Haytami. [Tuhfatu’l Muhtaj fi Sharh Al-Minhaj] (in Arabic: “تُحفَةُ المُحتَاجِ فِي شَرحِ المِنهَاج”). [On-line]. Available: http://shamela.ws/browse.php/book-9059 [October 13, 2016].

[17] M. Alghamdi, A. H. Alhamid, M. M. Aldasuqi, “Database of Arabic Sounds: Sentences,” Technical Report, King Abdulaziz City of Science and Technology, Saudi Arabia, 2003. (In Arabic).

[18] K. Bobzin. [Arabic Basic Course] (in German: “Arabisch Grundkurs”). Wiesbaden, Germany: Harrassowitz Verlag, 2009.

[19] [The Mecca list of common vocabulary] (in Arabic: “قَائِمَةُ مَكَّةَ لِلمُفرَدَاتِ الشَّائِعَة”). [On-line]. Available: http://daleel-ar.com/2016/09/08/قائمة-مكة-للمفردات-الشائعة/ [October 13, 2016].

[20] Arpabet, Internet: https://en.wikipedia.org/wiki/Arpabet [October 23, 2016].

[21] M. Alghamdi, Y. O. M. El Hadj, M. Alkanhal, “A manual system to segment and transcribe Arabic speech,” in: IEEE International Conference on Signal Processing and Communications (ICSPC), 2007, pp. 233–236.

[22] Evaluating Goodness of Fit. Internet: https://de.mathworks.com/help/curvefit/evaluating-goodness-of-fit.html?requestedDomain=www.mathworks.com [October 26, 2016].

[23] M. Zeki, O.O. Khalifa, A.W. Naji, “Development of an arabic text-to-speech system,” in: International Conference on Computer and Communication Engineering (ICCCE), 2010.