ASCII Based Transcription Systems for Languages with the Arabic Script: The Case of Persian

Shadi Ganjavi[*], *Panayiotis G. Georgiou*[§], *Shrikanth Narayanan*[§,*]

[*]Department of Linguistics, [§]Department of Electrical Engineering
Speech Analysis and Interpretation Laboratory: sail.usc.edu
University of Southern California
ganjavi@usc.edu, [georgiou, shri]@sipi.usc.edu

Abstract

In this paper, we discuss transcription systems needed for automated spoken language processing applications in languages such as Persian that use the Arabic script for writing. The work is described in the context of a speech-to-speech translation system development for English and Persian. This system can easily be modified for Arabic, Dari, Urdu and any other language that uses the Arabic script. The proposed system has two components. One is a phonemic based transcription of sounds for acoustic modeling in Automatic Speech Recognizers and for Text to Speech synthesizer, using ASCII based symbols, rather than International Phonetic Alphabet symbols. The other is a hybrid system, that provides a minimally-ambiguous lexical representation that explicitly includes vocalic information; such a representation is needed for language modeling and machine translation.

1. Introduction

Arabic script is the second most widely used written script after Latin. Languages that use this script have posed a problem for automated language processing such as speech recognition and translation systems. For instance, the CSLU Labeling Guide [1] offers orthographic and phonetic transcription systems for a wide variety of languages, from the European languages with a Latin-based writing system like German and Spanish to languages like Mandarin and Cantonese, which use Chinese characters for writing. However, there is still no transcription system for languages like Arabic, Persian, Dari, Urdu and many others, which use the Arabic script [1, 2, 3].

The motivation for creating the proposed class of transcription schemes stemmed from the necessities of the Transonics project, a speech-to-speech translation system developed as a part of the DARPA Babylon program [4, 5] As is evident by the data presented in this table, there are two major sources of problem for any speech-to-speech machine translation. Consider the examples in Table 1.

Table 1 Examples of the transcription methods and their limitation.

	Arabic Script	USCPers	USCPron	USCPers+
'six'	شش	SS	SeS	SeS
'lung'	شش	SS	SoS	SoS
'100'	صد	$d	sad	$ad
'dam'	سد	sd	sad	sad

The first problem is that of unusual prevalence of "homographs" due to the convention of ignoring vowels in the written form as illustrated by the words "six" and "lung", where identical (transliterated Arabic) orthographic representations (Column 3) correspond to different pronunciations [SeS] and [SoS] respectively (Column 4). The second problem is that of "homophones" illustrated by the words "hundred" and "dam" in Table 1, which have similar pronunciation [sad] (Column 4), despite their different spellings (Column 3). The former is a sample of the cases in which there is a many to one mapping between orthography and pronunciation, a direct result of the basic characteristic of the Arabic script, viz., little to no representation of the vowels. Therefore, to employ a system with a direct 1-1 mapping between Arabic orthography and a Latin based transcription system (what we refer to as USCPers in our paper) would be highly ambiguous and would create many problems for our speech-to-speech translation system. The latter, on the

other hand, is a representative of the cases in which the same sequence of sounds would correspond to more than one orthographic representation. Therefore, using a pure phonetic transcription, e.g., USCPron, would provide the necessary data for *Automatic Speech Recognizer* (ASR), but not for *Dialog Manager* (DM) or the *Machine Translator* (MT). The goal of this paper is twofold (i) to provide an ASCII based phonemic transcription system similar to the one used in the International Phonetic Alphabet (IPA), in line of Worldbet [6] and (ii) to argue for an ASCII based hybrid transcription scheme, which provides an easy way to transcribe data in languages that use the Arabic script.

We will proceed in Section 2 to provide the USCPron ASCII based phonemic transcription system that is similar to the one used by the International Phonetic Alphabet (IPA), in line of Worldbet [6]. In Section 3, we will then present the USCPers orthographic scheme, which has a one-to-one mapping to the Arabic script. In Section 4 we will present and analyze USCPers+, a hybrid system that keeps the orthographic information, while providing the vowels. This system can be an easy way to transcribe data not just in Persian but also in languages that use the Arabic script.

2. Phonetic Labels (USCPron)

One of the requirements of an ASR system is a phonetic transcription scheme to represent the pronunciation patterns for the acoustic models. Persian has a total of 29 sounds in its inventory, six vowels (Section 2.1) and 23 consonants (Section 2.2). The system that we created to capture these sounds is a modified version of the International Phonetic Alphabet (IPA), called USCPronunciation (USCPron for short). In USCPron, just like the IPA, there is a one-to-one correspondence between the sounds and the symbols representing them. However, this system, unlike IPA does not require special fonts and makes use of ASCII characters.

2.1. Vowels

Table 2: Vowels

	Front	Back
High	i	u
Mid	e	o
Low	a	A

Persian has a six-vowel system, high to low and front and back. These vowels are: [i, e, a, u, o, A], as are exemplified by the italicized vowels in the following English examples: 'b*ea*t', 'b*e*t', 'b*a*t', 'p*u*ll', 'p*o*ll' and 'p*o*t'. The high and mid vowels are represented by the IPA symbols. The low front vowel is represented as [a], while the low back vowel is represented as [A]. There are no diphthongs in Persian, nor is there a tense/lax distinction among the vowels.

2.2. Consonants

In addition to the six vowels, there are 23 distinct consonantal sounds in Persian. Voicing is phonemic in Persian, giving rise to a quite symmetric system. These consonants are represented in Table 3 based on the place and manner of articulation and their voicing ([-v(oice)] and [+v(oice)].

Table 3: Consonants (Bilabial, Labio-Dental, Dental, Alveo-Palatal, Velar, Uvular, Glottal)

	Bilab	Lab-Dnt	Dnt	Alv-Plt	vlr	Uvlr	Glot
Stop [-v]	p		t		k		?
[+v]	b		d		g	q	
Fricative [-v]		f	s	S	x		h
[+v]		v	z	Z			
Affricate[-v]				C			
[+v]				J			
Liquids			l, r				
Nasals	m		n				
Glides				y			

Many of these sounds are similar to English sounds. For instance, the stops, [p, b, t, d, k, g] are similar to the italicized letters in the following English words: '*p*otato', '*b*all', '*t*ree', '*d*oll', '*k*ey' and 'd*g*' respectively. The glottal stop [?] can be found in some pronunciations of 'bu*tt*on', and the sound in between the two syllables of 'uh oh'. The uvular stop [q] does not have a correspondent in English. Nor does the velar fricative [x]. But the rest of the fricatives [f, v, s, z, S, Z, h] have a corresponding sound in English, as demonstrated by the following examples '*f*ine', '*v*alue', '*s*and', '*z*ero', '*sh*ore', 'plea*s*ure' and '*h*ello'. The affricates [C] and [J] are like their English counterparts in the following examples: '*ch*urch' and '*j*udge'. The same is true of the nasals [m, n] as in '*m*ake' and '*n*o'; liquids [r, l], as in '*r*ain' and '*l*ong' and the glide [y], as in '*y*esterday'. (The only distinction between Persian and English is that in Persian [t, d, s, z, l, r, n] are dental sounds, while in English they are alveolar.) As is evident, whenever possible, the symbols used are those of the International Phonetic Alphabet (IPA). However, as mentioned before because IPA requires special fonts, which are not readily available for a few of the sounds, we have used an ASCII symbol that resembled the relevant IPA symbol. The

only difference between our symbols and the ones used by IPA are in voiceless and voiced alveopalatal fricatives [S] and [Z], the voiceless and voiced affricates [C] and [J], and the palatal glide [y]. In the case of the latter, we did not want to use the lower case 'j', in order to decrease confusion.

3. Orthographic Labels (USCPers)

We proceed in this section to present an alternative orthographic system for the Arabic language, as a stepping stone in the creation of the USCPers+ system that will be presented later. The Persian writing system is a consonantal system with 32 letters in its alphabet [7]. All but four of these letters are direct borrowing from the Arabic writing system. It is important to note that this borrowing was not a total borrowing. In other words, many letters were borrowed without their corresponding sound. This has resulted in having many letters with the same sound (homophones). Five sounds are represented by 13 letters:

- [s] for 'س', 'ص' and 'ث';
- [t] for 'ت' and 'ط';
- [q] for 'ق' and 'غ';
- [h] for 'ه' and 'ح' and
- [z] for 'ز', 'ذ', 'ض', and 'ظ'.

In order to assign a symbol to each letter of the alphabet, the corresponding letter representing the sound of that letter was chosen. So, for instance for the letter 'پ', which is represented as [p] in USCPron, the letter 'p' was used in USCPersian, USCPers for short. These letters are:

Table 4: USCPers(ian) Symbols (Non-Homophonic Consonants)

Stop	پ p	ب b	د d	ک k	گ g	ع ?
Fricative	ف f	ش S	ژ Z	خ x		
Affricate	چ C	ج J				
Liquid	ر r	ل l				
Nasal	م m	ن n				

In the case of the homophonous letters, several strategies were used. If there were two letters with the same sound, the lower case and the upper case letters were used, as in table 5.

In all these cases, the lower case letter is assigned to the most wildly used letter and the upper case, for the other. In the case of the letters represented as [s] and [z] in USCPron, because the corresponding upper case letters were already assigned, other symbols were chosen. For the letters sounding [s], 's', '$' and '&' and for the letters sounding [z], 'z', '2', '7' and '#'.

Table 5 USCPers(ian) Symbols: Homophonic Consonants 1

[t]	ت t	ط T
[q]	ق q	غ Q
[h]	ه h	ح H

Table 6 USCPers(ian) Symbols: Homophonic Consonants 2

[s]	س s	ص $	ث &	
[z]	ز z	ض 2	ظ 7	ذ #

Among the consonants, the letters 'ی' and 'ر' can be used as a consonant as well as a vowel, [y] and [i] in the case of the former and [v], [o] and [u] in the case of the latter. However, in USCPers, the symbols 'y' and 'v' were assigned to them, leaving the pronunciation differences for USCPron to capture. For instance, the word for 'you' is written as 'tv' in USCPers, but pronounced as [to], but the word 'and' is written as 'v' and pronounced as [va].

As mentioned earlier the vowels in the Persian language are mostly not marked in orthography. The only letter in the alphabet that represents a vowel is the letter 'alef'. This letter has different appearances depending on where it appears in a word. In the word initial position, it appears as 'ا', elsewhere it is represented as 'ا'. Because the dominant sound that this letter represents is the sound [A], the letter 'A' was assigned to represent 'ا', which has a wider distribution; 'V' was assigned for the more restricted version 'ا'. In Persian, like in Arabic, diacritics mark the vowels, although they are not used in writing, unless to avoid ambiguities. Therefore, in our system, we ignored the diacritics.

Table 7 Non-Persian Letters

Borrowed Letters	USCPers Symbol	USCPron
ً	@	an
ِ	*	a
ۀ	Y	e
ء	^	no sound
ؤ	W	o

Finally in creating the one-to-one mapping between the Persian alphabet and USCPers, we need to deal with the issue of "pure Arabic" letters that appear in a handful of words. We see the same situation in the borrowed words in English, for instance the italicized letters in cañon or naïve, are not among the letters of the English alphabet, but they appear in some words used in English.

In order to ensure a one-to-one representation between the orthography and USCPers, these letters were each assigned a symbol, as presented on Table7.
USCPers, therefore, provides us with a way to capture each letter of the alphabet with one and only one ASCII symbol, creating a comparable system to USCPron for the orthography.

4. USCPers, USCPron: Two Way Ambiguity

As was noted in the previous section, vowels are not usually represented in orthography and there are many homophones letters. These two properties can give rise to two sources of ambiguity in Persian which can pose a problem for speech-to-speech machine translation: (i) in which two distinct words have the same pronunciation (homophones), like 'pair' and 'pear' in English and the Persian words like 'sd' and '$d', which are both pronounced as [sad] and (ii) in which one orthographic representation can have more than one pronunciation (homographs) similar to the distinction between the two English words convict (n) and convict (v), which are both spelled c-o-n-v-i-c-t, but different stress assignments create different pronunciations. It is important to note that English has a handful of such homographic pairs, while in Persian homographs are very common, contributing to much ambiguity. In this section, we will discuss the transcription system we have adopted in order to eliminate these ambiguities.

4.1. Homophones

The following examples illustrate the case in (i) (the letters with the same sounds are underlined):

Table 8: Same Pronunciation, Different Spellings

Gloss	USCPers	USCPron
'hundred'	$d	[sad]
'dam'	sd	[sad]
'life'	HyAt	[hayAt]
'backyard'	HyAT	[hayAt]
'Eve'	HvA	[havA]
'air'	hvA	[havA]

As is evident by the last column, in each case, the two words have similar pronunciation, but different spellings. The word for 'life' ends in 't', while the word for 'backyard' ends in 'T'. In the other examples, because there is no difference in the pronunciation of 'h'/'H' and 's'/'$', we get ambiguity between 'Eve'/'air' and 'hundred'/'dam'. Therefore, this type of ambiguity appears only in speech.

4.2. Homographs

The second case of ambiguity is illustrated by the following examples:

Table 9: Same Spelling, Different Pronunciations

Gloss	USCPers	USCPron
'lung'	SS	[SoS]
'six'	SS	[SeS]
'thick'	klft	[koloft]
'maid'	klft	[kolfat]
'Cut!'	bbr	[bebor]
'tiger'	bbr	[babr]

In these set of examples, we see that in the middle column two words that have the same orthographic representation correspond to different pronunciations (Column 3), marking different meanings, as is indicated by the gloss. This type of ambiguity arises only in writing and not speech.

4.3. Solution: USCPers+

Because of the ambiguity presented by the lack of vowels the data transcribed in USCPers cannot be used either by MT or for language modeling in ASRs, without significant loss of information. In order to circumvent this problem, we adopted a modified version of USCPers. In this new version, we have added the missing vowels, which would help to disambiguate. (Because this new version is USCPers+ vowels, it is called USCPers+.) In other words, USCPers+ provides both the orthographic information as well as some phonological information, giving rise to unique words. Let us reconsider the examples we saw above using this new transcription system. A modified version of Table 8 is presented in Table 10 and Table 11 is the modified version of Table 9.

Table 10: USPers+ Disambiguates Cases with Same Pronunciation & Different Spellings

'Gloss'	USCPers	USCPers+	USCPron
'hundred'	$d	$ad	[sad]
'dam'	sd	sad	[sad]
'life'	HyAt	HayAt	[hayAt]
'backyard'	HyAT	HayAT	[hayAt]
'Eve'	HvA	HavA	[havA]
'air'	hvA	havA	[havA]

Table 12: An example from the online newspaper, Hamshahri (September 16, 1996, Fourth Year, Number 1068), where the Persian (Arabic Script) transcription is converted into the forms of USCPers, USCPers+ and USCPron for the purposes of creating a Language Model, dictionary and lexicon. The Gloss is also provided here although not generated in our collection.

Persian	مدیر عامل سازمان پارکها گفت تاکنون در تهران ۱۸ هزار هکتار (معادل ۱۸۰ میلیون متر مربع) به فضای سبز شهر تهران در ٦ سال گذشته افزوده شده واین شهر در حال حاضر ٦٣٠ بوستان، گلستان و باغ شهری بزرگ و کوچك دارد
USCPers	mdyr?Aml sAzmAn pArkhA gft tAknvn dr thrAn hyJdh hzAr hktAr m?Adl $d v hStAd mylyvn mtr mrb? bh f2Ay sbz Shr thrAn dr SS sAl g#Sth Afzvdh Sdh v Ayn Shr dr HAl HA2r SS$d v sy bvstAn glstAn v bAQ Shry bzrg v kvCk dArd
USCPers+	modyr?Amel sAzmAn pArkhA goft tAkonvn dar tehrAn hyJdah hezAr hektAr mo?Adel $ad va haStAd mylyvn metr moraba? beh fa2Ay sabz Sahr tehrAn dar SeS sAl go#aSteh Afzvdeh Sodeh va Ayn Sahr dar HAl HA2er SeS$ad va sy bvstAn golestAn va bAQ Sahry bozorg va kvCak dArad
USCPron	modir?Amele sAzmAne pArkhA goft tAkonun dar tehrAn heJdah hezAr hektAr mo?Adele sad o haStAd milyun metre moraba? be fazAye sabze Sahre tehrAn dar SeS sAle gozaSte afzude Sode va in Sahr dar hAle hAzer SeSsad o si bustAn golestAn va bAqe Sahriye bozorg va kuCak dArad
Gloss	The managing director of Parks and Recreation Services said that so far since last year, 18,000 acre, which is equivalent to one hundred eighty million square foot, of green area has been added to Tehran in the past six years and this city currently has 630 parks, orchards and small and large garden.

Table 11: USCPers+ Disambiguates Cases with Same Spelling & Different Pronunciations

'Gloss'	USCPers	USCPers+	USCPron
'lung'	SS	SoS	[SoS]
'six'	SS	SeS	[SeS]
'thick'	klft	koloft	[koloft]
'maid'	klft	kolfat	[kolfat]
'Cut!'	bbr	bebor	[bebor]
'tiger'	bbr	babr	[babr]

Data in Column 4 and Column 2 of Tables 10 and 11, respectively, show that USCPron and USCPers can give rise to ambiguity, while no ambiguity exists in USCPers+, Column 3. To further illustrate this point, consider the following case (where the words 'thick' and 'maid' from Table 11 are used). Assume that ASR receives the audio input in (1) represented in USCPron:

(1) USCPron: [in koloft ast]
Gloss: this thick is
Translation: 'This is thick'

If ASR outputs USCPers, as in (2),

(2) USCPers: Ayn klft Ast

the MT output in the English language can choose either:

(3) a. This is thick
b. This is a maid

as a possible translation. However, using USCPers+ instead of USCPers would avoid this ambiguity:

(4) USCPers+: Ayn koloft Ast (cf. (2))

As evident, there is a significant benefit by using USCPers+.

The discussion of the conventions that have been adopted in the use of USCPers+ and USCPron, e.g., not including punctuations or spelling out numbers, is beyond the scope of this paper. However, it is important to note that by adopting a reasonable number of conventions in our transcription of USCPers+ and USCPron, we have been able to provide a complete transcription convention for acoustic model and language models for the ASRs, TTSs and MTs for our English to Persian translation system.

5. Examples

5.1. USCPers and Transonics

To demonstrate the use of the USCPers class of transcription schemes, the example of Table 12 shows the process of mining data from publicly available sources for the purpose of generating a language model for the Persian language.

5.2. Augmentation- Lebanese Arabic

As mentioned before, Persian is among a handful of languages, which use the Arabic script. Our group is also researching speech technologies in other languages, which employ the Arabic script. One of these is Lebanese Arabic, which is commonly spoken in Lebanon. The Tables 13, 14, and 15 illustrate how, the above-mentioned system is adopted for Lebanese Arabic.

6.Conclusions

It is suggested in this paper, that a better way to represent data at phonological/lexical level for language modeling and MT in languages that employ a consonantal writing system, e.g., Arabic script, is by using a hybrid system, which combines information provided by orthography and includes the vowels that are not represented in orthography. This would ensure a

uniqueness that otherwise is not available. It has also been suggested in this paper that a modification of IPA, which would allow the use of ASCII characters, is a more convenient way to capture data for acoustic modeling and TTS.

Persian data resources developed under the DARPA Babylon program will adopt the conventions described in this paper.

Table 13: Vowel

Arabic Script	USCPers	USCPers+	USCPron	USCPron -LA
ـَ		a	a	a
ـِ		e	e	i
ـُ		o	o	u
ًا	@	@	an	an
ا	A	A	A	A
آ	V	V		'A
ي	I	I	I	I
ي	E	E	E	E
و	v	v	U	U
و	v	v	O	O
ى	I	I	A	A
ة	F	F	a e	a[1] or e

Table 14: The Hamzeh

Arabic Script	USCPers	USCPers+	USCPron	USCPron-LA
أ	*	*	a	'a
ئـ ئ	Y	Y	e	'i
ء	^	^	no sound	no sound
ؤ	W	W	o	'u
ع			'	'

7.Acknowledgements

This work was supported by the DARPA Babylon program, contract N66001-02-C-6023. We would like to thank the following individuals for their comments and suggestion: Narineh Hacopian, Naveen Srinivasamurthy, and finally HS, MK and SS for working with the first versions of this system and making insightful suggestions.

Table 15: The Consonants

Arabic Script	USCPers	USCPers+	USCPron	USCPron -LA
ب	b	b	b	b
د	d	d	d	d
ك	k	k	k	k
ع	?	?	?	9
م	m	m	m	m
ن	n	n	n	n
ف	f	f	f	f
ش	S	S	S	$
خ	x	x	x	x
ر	r	r	r	r
ل	l	l	l	l
ج	J	J	J	j

8.References

[1]T. Lander, The CSLU Labeling Guide, OGI, http://cslu.cse.ogi.edu/corpora/corpPublications.html

[2]Alan S. Kaye. "Arabic," The World's Major Languages, ed. Bernard Comrie, Oxford University Press (1987).

[3]Yamuna Kachru, "Hindi-Urdu," The World's Major Languages, ed. Bernard Comrie, Oxford University Press (1987).

[4]"The DARPA Babylon program," http://darpa-babylon.mitre.org.

[5]Narayanan, et. al., Transonics: A speech to speech system for English-Persian interactions. (submitted, ASRU 2003)

[6]James L. Hieronymus, ASCII Phonetic Symbols for the World's Languages: Worldbet, AT&T Bell Labs, http://cslu.cse.ogi.edu/corpora/corpPublications.html

[7]Gernot, L/ Windfuhr, "Persian," The World's Major Languages, ed. Bernard Comire, Oxford University Press (1987).