The Challenges and Pitfalls of Arabic Romanization and Arabization

Jack Halpern

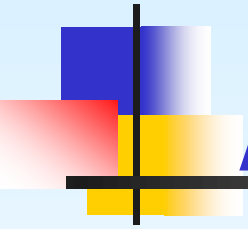
CEO

漢

The CJK Dictionary Institute, Inc.

المؤسسة المعجمية للغات الشرقية

日中韓辭典研究所



Automatic Romanizer of Arabic Names

ARAN

الناقل اللغوي الآلي للأسماء العربي

Non-Arabic Name Arabizer

NANA نعنع

نقل عربي نقل عربي

Transcription and Transliteration

☺ *Never the Twain Shall Meet* ☺

Transliteration: representing the source script letters (graphemes *not* phonemes) with the characters of another script. محمد > /mHmd/

Transcription: representing the source script of a language in the target script in a manner that reflects pronunciation. This includes:

1. Phonetic transcription represents the actual speechsounds.
محمد > [muħɛ̈mmɛ̈d] (IPA)
2. Phonemic transcription represents the phonemes of the source.
محمد > /muHammad/
3. Popular transcription roughly represents pronunciation.
محمد > Mohammed, Muhammad, Moohammad, Moohamad…+200

NANA Conversion Modes

- Latin Clinton → كلينتون
- Japanese 埼玉(Saitama) → سايتاما.
- Chinese 杨海洋 (Yang Haiyang)→هاييانغ يانغ
- Korean 부산 (Busan) → بوسان.

Diphthong Ambiguity for 福井 /fu-ku-i/

No.	Arabic	Google hits	Buckwalter
1	فوكوئي	468	fwkw}y
2	فوكوئ	9	fwkw}
3	فوكوي	1950	Fwkwy
4	فوكويي	335	Fwkwyy

Long and Short Vowels

No.	Kanji	Kana	Phonemic	Arab1	Arab2	Arab3
1	太田	おおた	oota	أوتا		
2	風馬	ふうま	fuuma	فوما		
3	敬子	けいこ	keiko	كيكو	كيكو	
4	空野	くうの	kuuno	كونو		
5	久野	くの	kuno	كونو		
6	日枝	ひえだ	hieda	هيدا	هيئدا	فيئيدا
7	芳江	よしえ	yoshie	يوشي	يوشيئي	ئيوشيئي

Orthographic Ambiguity

- Short vowel omission (كاتب / kAtb /)
- Short vowel Representation (جامعة /jaami`a/)
- Multiple long /aa/

'alif Tawiila (ا) سوريا

'alif maduuda (آ) آسيا

'alif maqSuura (ى) ا آسيالوسطى.

- Long vowel omission (هدا /haadha/)
- Long /aa/ ambiguity (شكرا, انا) (/an/, /a/)

Orthographic Ambiguity

- Otiose alif is silent كتبوا > /katabuu/
- Omission of *shadda* (مُحَمَّد > محمد) /Muhammad/
- Omission of *tanwiin* شكرا ¥$ukrAF¥ (شُكْرَاً)
- Complex hamza rules (فوكوؤوكا vs فوكوأوكا)
- Hamza omission (سائيتاما >سايتاما)
- Phonological alternation (لرجل الطويل)
  /'alrajulu alTawiilu/> /'arrajulu-Ttawiilu >
- Shortening long vowels (القاهرة في)
  /fii-lqaahira/ > /fi-lqaahira/

Output from ARAN modules

Unvocalized	Vocalized	Phonemic	Graphemic	Phonetic	Popular
(input)	(ADAN)	(ATAN)	(AXAN)	(APAN)	(AVAN)*
محمد	مُحَمَّد	muHammad	mHmd	muħɛ̈mmɛ̈d	Muhammad
قابوس	قَابُوس	qaabuus	qAbws	qɑːbuːs	Qaboos
جمال	جَمَال	jamaal	jmAl	dʒɛ̈mɛ̈ːl	Jamal
مكة	مَكَّة	makka	mkp	mɛ̈kkɛ	Mecca

*Only one popular variant is shown, but in reality there could be dozens. For example, for قابوس AVAN generates Qabuus, Qabus, Qabous, Qabooss, … and many more.

ARAN Processing of قابوس /qAbws/

Conversion process	ARAN module	Input	Output	Remarks
Phonemic Transcription	ATAN	ق ا بِ وس	/qaabuus/	linguistic representation of phonemes
English Transcription	ATAN	ق ا بِ وس	Qaboos	Standard English spelling
Popular Transcriptions	ATAN	ق ا بِ وس	Qabuus, Qabus, Qabous, Qabooss, Qaaboos, Kaboos, Kabuus, Gabous…	some of the many popular variants

ARAN Processing of قابوس /qAbws/

Conversion process	ARAN module	Input	Output	Remarks
Phonetic Transcription	APAN	ق ابـ وس	[qɑːbuːs]ْ	scientific transcription in IPA
Unvocalized Transliteration	AXAN	ق ابـ وس	/qAbws/	Buckwalter transliteration of unvocalized Arabic
Vocalized Transliteration	AXAN	قَ ابُ وس	/qaAbuws/	Buckwalter transliteration of vocalized Arabic
Diacrticization	ADAN	ق ابـ وس	قَ ابُ وس	adding vowels (vocalization) and diacrtics to unvocalized Arabic
Arabization	NANA	Qabuus, Qabus, etc.	ق ابـ وس	converting non-Arabic to Arabic script

Major Arabic Romanization Systems

Example: شولوخ

System	Example	Description
ALC-LC	shwlwkh	Romanization standard of the American Library Association-Library of Congress.
DIN	šūlūḫ	This refers to DIN 31635, the DIN standard for Arabic transliteration.
IPA	ʃuːluːx	International Phonetic Alphabet, a scientific system of uniquely and accurately representing speech sounds.
English	Shoulokh	One of many (at least 10) possible popular transcriptions.
Buckwalter	$wlwx	A strict transliteration system widely used in information processing.

Sample output from ADAN module

Unvocalized	Vocalized	Transcription	English
محمد	مُحَمَّد	muHammad	Muhammad
إبـ راهـ يم	إِبْرَاهِـ ـيم	'ibrahiim	Abraham
إسحاق	إِسْحَاق	isHaaq	Isaac
الـ ريـ اض	الرِّيَـ اض	arriyaaD	Riyadh
مـ كة	مَـكَّة	makkah	Mecca
الـ قاهرة	الْـ قَـاهِرَة	alqaahirah	Cairo

Variation in Arabic Names

Standard	Buckwalter	English	Variant	Error	Remarks
أبو ظبي	>bw Zby	Abu Dhabi	ظبي بو	ظبأبو	V: omit hamza E: alif maqsura replaces yaa'
الإسكندرية	Al<skndryp	Alexandria	اسكندريه	الاسكندريا	V: omit hamza E: haa' replaces taa' marbuuTa
جدة	jdp	Jeddah	جدّة	جده	V: explicit shadda E: haa' replaces taa' marbuuTa
الأردن	Al>rdn	Jordan	الاردن		V: omit hamza
بالو ألتو	bAlw >ltw	Palo Alto	الـ تو الـ و آلـ تو الـ و		V1: omit hamza V2: madda replaces hamza
الرياض	AlryAD	Riyadh	الرّياض		V: explicit shadda
طوكيو	Twkyw	Tokyo		توكيو	E: taa' replaces Taa'

Popular Transcriptions

Arabic	Buckwalter Transliteration	Popular Transcription
معـمر	mEmr	Moammar
معـمر	mEmr	Muammar
معـمر	mEmr	Mu'ammar
معـمر	mEmr	Mu`ammar
معـمر	mEmr	Mo'ammar
معـمر	mEmr	Moammar
معـمر	mEmr	Moamer
معـمر	mEmr	Moamar
معـمر	mEmr	Mohamar

MSA Flavors

Arabic	English	Phonemic	Phonetic Gulf	Phonetic Egyptian	Phonetic Levantine
قابوس	Qaaboos	/qaabuus/	[qɑːbuːs]	[ʔɑːbuːs]	[qɑːbuːs]
جمال	Jamal	/jamaal/	[dʒɛ̈mɛ̈ːl]	[gɛ̈mɛ̈ːl]	[ʒɛmɛ̈ːl]

Variation in Arabic Names

Standard	Buckwalter	English	Variant	Error	Remarks
ظ بي أب و	>bw Zby	Abu Dhabi	ظ بي اب و	ظ بى أب و ظ بى اب و	V: omit hamza E: *'alif maqsura* replaces *yaa'*
سـكـندريـة	Al<skndryp	Alexandria	سـكـندريـة	سـكـندريـا	V: omit *hamza* E: *haa'* replaces *taa' marbuuTa*
ألـ تـو بـا لـو	bAlw >ltw	Palo Alto	الـ تـو بـا لـو آلـ تـو بـا لـو		V1: omit hamza V2: *madda* replaces hamza
طوكـيو	Twkyw	Tokyo		تـوكـيو	E: *taa'* replaces *Taa'*

English-to-Arabic Errors

NUM	English	Error	Google_Hits	Google_Arabic	CJKI_Arabic
1	Muhammad	NO	24,700,000	محمد	محمد
2	Moohammad	YES	2,720	موحمد	محمد
3	Moohamad	YES	975	موهماد	محمد
4	Mohammad	NO	15,100,000	محمد	محمد
5	Mohamad	NO	4,030,000	محمد	محمد
6	Muhamad	NO	795,000	محمد	محمد
7	Mohamed	NO	23,700,000	محمد	محمد
8	Mohammed	NO	26,700,000	محمد	محمد
9	Mohemmed	YES	940	موه يم يد	محمد
10	Muhemmed	YES	12,800	موه يم يد	محمد
11	Muhamed	NO	742,000	محمد	محمد
12	Muhammed	NO	7,800,000	محمد	محمد
13	Moohammed	YES	1,840	موهم يد	محمد
14	Mouhammed	NO	102,000	محمد	محمد

Variants and Errors

Rank	Type	ARABIC	Buckwalter	Googlits	Remarks
1	N	الاسكندريه	AlAskndryp	2,930,000	Normalized, no hamza
2	S	الإسكندريه	Al Askndryp	690,000	Standard form, with hamza
3	E	الاسكندريه	AlAskndryh	89,200	No hamza, taa marbuta replaced by haa
4	V	الاسكندرِيّه	AlAskndry~p	954	Explicit shadda
5	E	الإسكندريه	Al Askndryh	897	taa marbuta replaced by haa
6	V	الإسكندرِيّه	Al Askndry~p	245	no hamza, shadda explicit
7	E	الاسكندريا	AlAskndryA	80	hamza omitted, taa marbuuta replaced by alif
8	V	الإسْكَنْدَرِيَّه	Al AskanodaryyAp	24	fully vocalized
9	E	الاسكندرِيّه	AlAskndry~h	12	no hamza, shadda explicit, taa marbuta replaced by haa
10	E	الإسكندريا	Al AskndryA	7	taa marbuta replaced by alif tawiila
11	E	الاسكندرِيّه	AlAskndry~h	5	taa marbuuta replaced by haa, shadda explicit