



Romanized Arabic

- Romanized Arabic - spelling Arabic words in Roman letters.
- Why use Romanized Arabic?
 - No access to an Arabic keyboard.
 - Can't type in Arabic.
 - It is "cool"
 - As an "in-group" language for fun, privacy...
- Who uses Romanized Arabic?
 - Arabic native speakers of all ages and backgrounds
 - But especially the younger generation.





Where is Romanized Arabic used?

- Chatrooms
- Internet Relay Chat (IRC)
- Instant messaging
- Forums
- Blogs
- E-mails
- Cellphone text messaging
- Schools such as: Arab Academy
- Signs on the streets in Egypt





Why Study Romanized Arabic?

- If one would like to study the usage of Arabic on the internet, then one has to research Romanized Arabic.



Why study Romanized Arabic? - Linguistically

- Tells us how people really “feel” and “think” about the phonology, morphology, syntax, etc. of their language (Hentschel).
- Study linguistic phenomena:
 - Sound structure of language.
 - Pattern of switching between languages or dialects.
 - Syntax of Romanized Arabic (RA) vs. Modern Standard Arabic (MSA), or any of the spoken dialects?



Romanized Arabic studies

- Past studies
 - Palfreyman and al Khalil (2003)
 - Warschauer, El-Said, and Zohry (2002)
 - *Limited in number and scope of dialects studied.*

- Basis Technology chat study
 - Broad range of Arabic dialects: Levantine, Iraqi, Gulf, Maghrebi, and Egyptian.
 - Results are based on what we've seen in the corpus.
 - Created the Reverse Transliterator software to convert Arabic chat to Arabic script.



Basis RA Corpus

	Phase I	Phase II	Total	Percentages
	Unique Tokens	Unique Tokens	Unique Tokens	per Tokens
Algeria	47	180	227	0%
Bahrain	5	58	63	0%
Egypt	2326	1781	4107	9%
Iraq	0	4523	4523	10%
Israel	0	0	0	0%
Jordan	155	635	790	2%
Kuwait	1192	5760	6952	15%
Lebanon	2460	4004	6464	14%
Libya	0	0	0	0%
Morroco	194	2908	3102	7%
Oman	231	51	282	1%
Palestine	0	361	361	1%
Qatar	0	355	355	1%
Saudi Arabia	82	2498	2580	5%
Sudan	0	343	343	1%
Syria	51	227	278	1%
Tunisia	147	152	299	1%
UAE	1737	6636	8373	18%
Yemen	0	0	0	0%
Multiple	7710	180	7890	17%
				0
Total	16337	30652	46989	100%



Representation of the Phonemic Consonants

Arabic	Romanized Arabic	Alternative spellings	Example
ث	'T	'C, TH	thamin ثمين
ج	G	J	Jamal جمال
ح	7	H	nroo7 نروح
خ	'7	KH, 5, *7, 7*, 7'	sa5ef سخيف
ذ	'D	'Z, TH	t3theeb تعذيب

Red = unused standard mappings



Representation of the Phonemic Consonants

Arabic	Romanized Arabic	Alternative spellings	Example
ش	'S	SH, ch	She شيء
ص	9	s	9ar صار
ض	'9	D, 9*, *9, 9'	'9arab
ط	6	T	al5wa6eer الخواطر
ظ	'6	Th, 6*, *6, 9*	'6ahir ظاهر



Representation of the Allophones

Arabic Letter	Allophonic Variations	Examples
ث	t, s	Talateh Thalatheh Salaseh ثلاثة
ج	J, G, dj, y, sh	Rujjal/ rijjal Riggal Rudjal Rayyal رجال Bwish بوج
خ	K	Kalid خالد



Representation of the Allophones

Arabic Letter	Allophonic Variations	Examples
ظ	D, z	Dal ظل Buza بوظة
ق	g, ʒ, k, A, 2, ‘	Qasim Gasim ʒasim Kasim Asim 2asim ‘asim قاسم
ك	tch, tsh, ch,	3indich 3inditsh 3indich عندك



Challenges of Reading Romanized Arabic

- Phonemic spelling
 - ani 6lbt mn al group owner → طلبت من
- Romanized Arabic spelling could be ambiguous, ex: “mn”
 - من from
 - مَنْ who/whom
 - مَنّْ grace/favor
- Misspellings:
 - wassalto wassalam 3la sayyed al khalq ajma3een
 - الصلاة والسلام على سيد الخلق أجمعين
 - وصلته



Multiple uses for same character

- Single quotes used as consonants or just single quotes!
 - oh ba3dain 'thara' oh 'electrons'
 - lak Zeus anaa akbar 3alim tharra.
- Geminate consonants and long vowels may not be represented correctly.
- Numbers used for Arabic characters or just as numbers!
 - Fe el 90'Z
- Case variations for fun or proper nouns:
 - KiLLeH Mn O5OoYuH e93'eeR



“Fuzzy standard” for Romanized Arabic

- Inconsistent use of the “standard” for mapping Arabic letters into RA .
 - Example: *9 used for ظ and ض .
- Some hypothesized representations are not used:
 - Example: 'T, 'C, 'Z



Spelling affected by English?

- Some people's spelling could be affected by English, such as:
 - Using [x] for the sound [ks]
 - *Thanx*
 - *bel 3ax hum a9lan maynlamoon*
بالعكس هم أصلاً ما ينلامون



Representation of the English Word

- Lot of switching with English and other languages.
 - English words spelled in Romanized Arabic/ Arabicized?
 - *laken al kibord e7'tarab*
 - *wa almasengar mo radi yefta7*
 - English chat language:
 - *why don't u think of ppl ill 6ale3oohum min bladhum*
 - *tmaam w u*
 - English misspellings:
 - *cuz i heard that if they don't inturpreate it right.. it will come true..*



Morpheme Boundaries - Prefixes

- Prefixes could appear detached from the word.
 - min nahiyat el 3alakat bi chabab , fa ana arfod aya 3alaka ma3a chab kharej netak ederassa
 - kell el nhar betdalik 7ebseta bel beyt w 3al tari2 kamen
 - im not saying it 7alal bi ma3na feeh 2aya wela 7adeeth
 - bas 7abit sa7e7ak enu il nis bit oul
 - Ana ib america.
- short prefixes such as single letter conjunctions and single letter prepositions are “detachable” from the word.



Morpheme Boundaries - Suffixes

- Suffix detachment happens less often:
 - '6arabt **hom** '6arb
 - fe wa7dah ba7reneyah o amsawyah fe nafs **ha** chethee

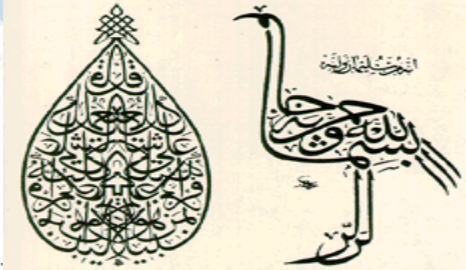


Morpheme Boundaries

- Prefixes and suffixes may be separated from the word by a dash or an underscore:
 - el-7ath
 - el-3azizeh

 - ako we7da ilexam **malat-ha** yom ilsabet
 - gadeesh darajet **7araret-ha**
 - haih malaak **theb7at_ha** el'3airaah
- Words may be separated/connected by a dash!
 - **Intal-3iraqi** am ana? ← /anta al3irqai/
 - Khasi2t **yabnal-zinna** ← /ya ibna alzinna/

Quranic language



- Use of dashes for harakat in transliterated MSA or Quranic language for rhyming and poetic sound:
 - 3alayhimu-s-salaam
 - Assalamu 3alaika yabna amir-ul-mu2mineen wabna sayyid-il-wasiyyina
Assalamu 3alaika yabna Fatima sayyidati nisa2-il-3alameen Assalamu
3alaika ya thaarallahi wabna thaarihi wa-l-watr-al-mawtuur
 - allahumma-l3an-il-3isaabata-l-lati jaahadati-l-7usayn
 - wa 3ala 3aliyyi-bni-l-7usayn





Word boundaries

- Joined words
 - Example: Function words attached to next word:
 - **Fillah** في الله
 - **Yarait** ياريت
 - **Matsadeg** ما تصدق
 - Content words joined, especially after common phrases:
 - **Sub7anallah** سبحان الله



Arabish words are creatively coined

- People try to be smart and creative..
 - shaklhaaaa wallah shrat eryayel..mafeha n3omah.. :unsure: al7en chay & shesh0o beygonon haih malaak theb7at_ha el'3airaah..agra afkarekom :P ma a7es enneh malaame7ha **male-eyyeh** at all :huh: 7lwa! heeeh e9ara7a wl 7aqeqa toqal she is pretttyy y3ybny hal jamal





Representation of sentence boundaries

- lack of punctuation
 - ana kenet be7dar tv 3a 6ol **ow** ana z3'eer **ow** haida sababli el badanei **ow** serst kteer nase7 belnesbi la 3omri **ow** haida sbabli no3 men el ete2ab **ow** ba3den seret ma ro7 3al madraseh bas men ba3ed ma fetet 3al jam3a seret nafseti a7san **ow** ssar wazni 3adi metel yali be 3omri **ow** hala2 ana enssan 6abe3i bas mesh lazem ne7rem awladna men el tv bas lazem ne5alehom ya3mlo sport **ow** yesta3mlo el internet **ow** el mo6ala3a men shan ma ykon el tv el she be 7ayaton **ow** yt3la2o feh we yeser kel 7ayaton **ow** 3ala fekra el tv beyb3o el awla 3an el le3eb **ow** haida el she mohemla elon men shan 7aleton el nefseeh



Sentence boundary detection

- Using commas as periods:
 - Alla y2awikoun wa ya3tikoun alf 3efiye , welli byejtème3 bi Zahle mesh metl li byejtème3 bi Jounieh , le anno Zahle b3ide 3an el Shar2ieh wa marat ktire el mo3ta2al deggre byekhdou 3a 3anjar 3end el souriyin , baynama li byo3ta2al bi Jounieh 3am ye3te2lou mokhabarat el jaysh el lebnene , ba3da shway ar7am.
- Sometimes the only punctuation we see is a smiley face:
 - y5tee 7abeetk mn '3eer ma a3rfk mn jed entee 5a6eerah o klamk ybrd el glb o yakther elle ybeelhm akwaa3 mo koo3 wa7d bs :)
ya5te ma dre men wain t6al3een halkalamm loool a5af dam ensana 3araftaha loool estmeree o allah yr3aki loool ya haila hop yr7am omek la treden 3ala hal 3ainat l2nek kbeera o 6ool 3omr el kbeer k beer



Language soup

- Arabic chat switches between
 - Languages -- Arabic, English, French, etc.
 - Arabic dialects.
 - MSA and one's own dialect.
 - Quranic language, MSA, other languages & Arabic dialects.





Mixing between Arabic script and RA

<fpyuigjrh> sho ?
<euu> بدل ما بتقعد ساعه وانتى تكتب بكلمة انجليزي
<fpyuigjrh> esma3 wala
<gpyof> eh!
<fpyuigjrh> otrok elly b2edak o ta3al hon
<fpyuigjrh> wala ya tabel
<fpyuigjrh> b7ki ma3ak
<euu> الله يخزيك على هالوراق اللي طابعهن
<fpyuigjrh> beddal 10 snen la te3mal zayhom
<fpyuigjrh> beddak 10 snen la te3mal zayhom
<fpyuigjrh> Abooooooya
<fpyuigjrh> HaMeL baba
<oszjb> الله يغضب عليك
<oszjb> من ولد
<iwnurqg> lol
<fpyuigjrh> yel3an aboy be 60 kondara
<iwnurqg> loooooooooooooool
<oszjb> أنا الظاهر معرفتش اربي
<oszjb> يا خسارة تعبي
<fpyuigjrh> rabby 7alak yaba
<fpyuigjrh> yel3anak sho hamel
<iwnurqg> lol
<oszjb> يا خسارة وجع ال9 شهور





Mixing in Arabic and English acronyms & abbreviations

- Mix of English and Arabic chat acronyms and abbreviations are used.
- Arabic chat abbreviations:
 - mbrooooooooook q8 3laaa ilfoooz
 - ya alaaaaaaaah shoofaw telvisyoon leq8
 - ilq8yeeen yn3doon 3la ela9abee3
 - wow masha2allah JAK roonny rabena yekrmeeek



Non-linguistic markers

- Mimic vocal expressions using:
 - Repetition of the same letter
 - Using upper case vs. lower case
- People duplicate both consonants and vowels, unlike real speech.
 - Anger
 - Shouting
 - Singing
 - Emphasis
 - Surprise, etc.
 - *ZEUS SHINO A3'AAAAAAAAAAR MIN MINO???????????*
 - *my msg contains many imgs oo kolaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
yrooooooooooooooooooooo7*
 - *yallah allah ygawwezna kulna ya rab aminnnn fatah muaddaba wa
gamela ist whats imporanttttttttttttttttttttttttttttttt*



Summary

- Arabic chat challenges for natural language processing
 - Inconsistent Romanization
 - Mixing languages, scripts, dialects
 - Loose boundaries: morphemes, words, sentences
 - Inconsistent use of punctuation
 - And more...



sHKrAn jazeeeeeeelAn!!!!!!!

Lo000o0o0o0L ☺