CENTER FOR ADVANCED
STUDY OF LANGUAGE

Dynamic Materials for Teaching and Learning Social Media Arabic

Julie Yelle, C. Anton Rytting, Paul Rodrigues, Michael Maxwell, Jennifer Boutz, Hannah Benninger

{jyelle, crytting, prr, mmaxwell, jboutz, hbenning}@umd.edu

(Presented by Julie Yelle)

23 May 2018

Motivation

- Media Arabic: “the language of newspapers, magazines, and internet news sites” (Elgibali & Korica, 2014)
- Social Media Arabic: the language of Internet forums, Twitter, Facebook, etc.
- Language for specific purposes (LSP): language courses specific to a content area
- Goals: production vs. reception

Social Media vs. Written Dialect

Social Media Arabic	Other Written Dialect
Lack of context obscures dialect identification	Context facilitates dialect identification
Spelling conventions are inconsistent	Spelling conventions are consistent
Typographical errors are frequent	Typographical errors are limited
Whitespace errors and inconsistencies are common	Whitespace usage is consistent
Punctuation is routinely omitted	Punctuation is routinely included
Written in Arabic and/or Roman (Arabizi) script	Written in Arabic script
Meaning is conveyed by text and emojis	Meaning is conveyed by text alone

Dialect Identification

ابي قلبك يعاني و يتألم

Dialect Identification

ابي قلبك يعاني و يتألم

Dad, your heart is suffering and feeling in pain

Dialect Identification

ابي قلبك يعاني و يتألم

I want your heart to suffer and feel in pain

Loanwords

- Internet jargon
  - Snapchat: السناب [*al-snāb*]
  - Whatsapp: واتس [*wāts*]
  - Followers: فولورز [*fūluwarz*]
- Transliterated named entities
  - FC Barcelona: البرسا [*al-barsā*]




Loanwords (cont.)

- Foreign words with Arabic morphological affixes

- Other borrowings
  - Group: قروب [*qrūb*]
  - Action: أكشن [*akshin*]

Neologisms

- Celebrity-focused word coinages
  - o سهلاوي [*sahlāwī*]: pertaining to or a fan of Moroccan singer Souhila ben Lacchab
  - o إيهابي [*īhābī*]: pertaining to or a fan of Algerian singer Ihab Amir
  - o سوهابي [*sūhābī*]: pertaining to or a fan of Souhila and Ihab

Dysphemisms

Hezbollah	حزب الله	*ḥizb Allāh*	“party of God”
	حزباله	*ḥizbālah*	“garbage-party”
	حزب اللات	*ḥizb al-Lāt*	“party of al-Lat”

Nasrallah	نصر الله	*naṣr Allāh*	“victory of God”
	نصر اللات	*naṣr al-Lāt*	“victory of al-Lat”

Interjections

- Many varieties of laughter
- Transliterated acronyms
  - LOL→ لول
- Elongation for emotional effect
  - لووووووول (looooooool)
  - مبروووووووك (congratulaaaaaations)
- Ululation
  - لولولولي
  - يويويوووووو

Stylistic Use of Characters

- Taṭwīl

  جـــــداً → جداً

- Use of non-standard characters for decorative purposes

من عظّم ربّہ وحسُن توگلہ علیہ وامتلأ قلبہ بحبّہ غناہ عمّن سواہ و گفاہ بہ و اگرمہ وأجزل لہ ألم
یقل سُبحانہ: (أليسَ الله بكافٍ عبده)

من عظّم ربّه وحسُن توكله عليه وامتلأ قلبه بحبّه غناه عمّن سواه و كفاه به و اكرمه وأجزل له ألم
يقل سُبحانه: (أليسَ الله بكافٍ عبده)

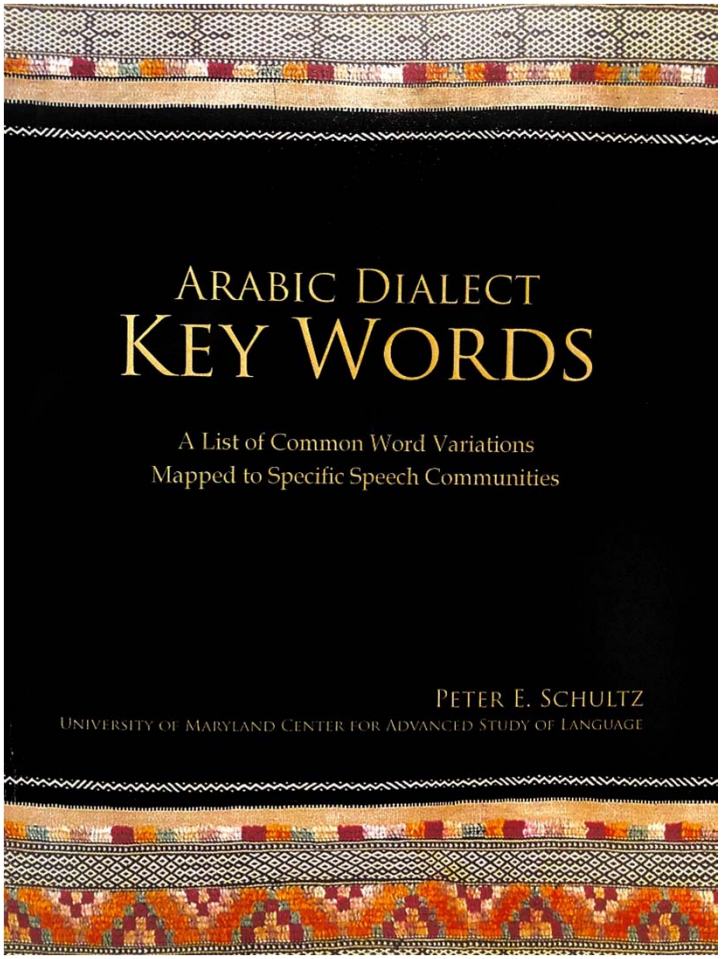
Lexicons

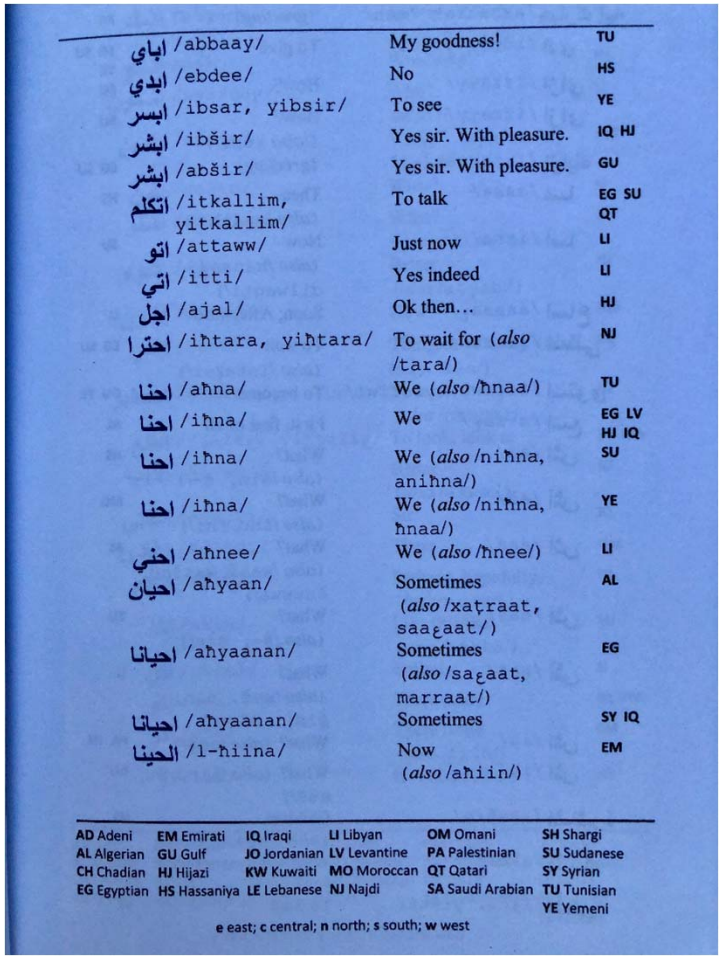
- Twiktionary
- Dialect lexicons with web examples
- Word frequency lists
- *Arabic Dialect Key Words: A List of Common Word Variations Mapped to Specific Speech Communities* (Schultz, 2011)

Arabic Dialect Key Words

- Indexes by region of distinctive key words and phrases
- Includes maps of geographic range of use
- Searchable by:
  - o Arabic word
  - o English word
  - o Major dialect

Arabic Dialect Key Words





اباي	/abbaay/	My goodness!	TU
ابدي	/ebdee/	No	HS
ابسر	/ibsar, yibsir/	To see	YE
ابشر	/ibšir/	Yes sir. With pleasure.	IQ HJ
ابشر	/abšir/	Yes sir. With pleasure.	GU
اتكلم	/itkallim, yitkallim/	To talk	EG SU QT
اتو	/attaww/	Just now	LI
اتي	/itti/	Yes indeed	LI
اجل	/ajal/	Ok then…	HJ
احترا	/iḥtara, yiḥtara/	To wait for (*also* /tara/)	NJ
احنا	/aḥna/	We (*also* /ħnaa/)	TU
احنا	/iḥna/	We	EG LV HJ IQ
احنا	/iḥna/	We (*also* /niḥna, aniḥna/)	SU
احنا	/iḥna/	We (*also* /niḥna, ħnaa/)	YE
احني	/aḥnee/	We (*also* /ħnee/)	LI
احيان	/aḥyaan/	Sometimes (*also* /xaṭraat, saaɛaat/)	AL
احيانا	/aḥyaanan/	Sometimes (*also* /saɛaat, marraat/)	EG
احيانا	/aḥyaanan/	Sometimes	SY IQ
الحينا	/l-ħiina/	Now (*also* /aḥiin/)	EM

AD Adeni	EM Emirati	IQ Iraqi	LI Libyan	OM Omani	SH Shargi
AL Algerian	GU Gulf	JO Jordanian	LV Levantine	PA Palestinian	SU Sudanese
CH Chadian	HJ Hijazi	KW Kuwaiti	MO Moroccan	QT Qatari	SY Syrian
EG Egyptian	HS Hassaniya	LE Lebanese	NJ Najdi	SA Saudi Arabian	TU Tunisian
					YE Yemeni

e east; c central; n north; s south; w west

Arabic Data Science Toolkit

- Framework for Arabic language feature extraction
- Designed for data scientists not necessarily familiar with Arabic or NLP
- Arabic-specific linguistic and stylistic features
  - o Spelling and normalization
  - o Social media features
  - o Stylistic features
  - o Word category features

Arabic Data Science Toolkit (cont.)

Attached Waw Frequency

{"attached_waws": 1, "detached_waws": 0}

Bag Of Words Frequency - no Diacritics

{"bag_of_words": {"1 :"يوم", 1, "وَجِيه": 1, "ناس": 1, "قُول": 1, "قَتَل": 1, "صامِد": 1, "حِين": 1, "حَنَا": 1, "جِي": 1, "جَدِيد": 1, "أُخْت": 1}}

Bag Of Words Frequency

{"bag_of_words_without_diacritics": {":"يوم", 1, "وجيه": 1, "ناس": 1, "قَول": 1, "قَتَل": 1, "صامد": 1, "حين": 1, "حنا": 1, "جي": 1, "جديد": 1, أخت": 1
1}}

Repeated Letters

{"repeated_letters": {}}

MisspelledWords

{"misspelled_words": {"اختة": {"hypotheses": {"آختة": {"edits": [["أ" ,"ا"]], "cost": "1.000000"}, "أختة": {"edits": [["أ" ,"ا"]], "cost":
"1.000000"}, "آختاه": {"edits": [["ا" ,""] ,["أ" ,"ا"]], "cost": "3.000000"}}}}}

Questions