

The Conundrum of New Online Languages : Translating Arabic Chat

Joel Ross

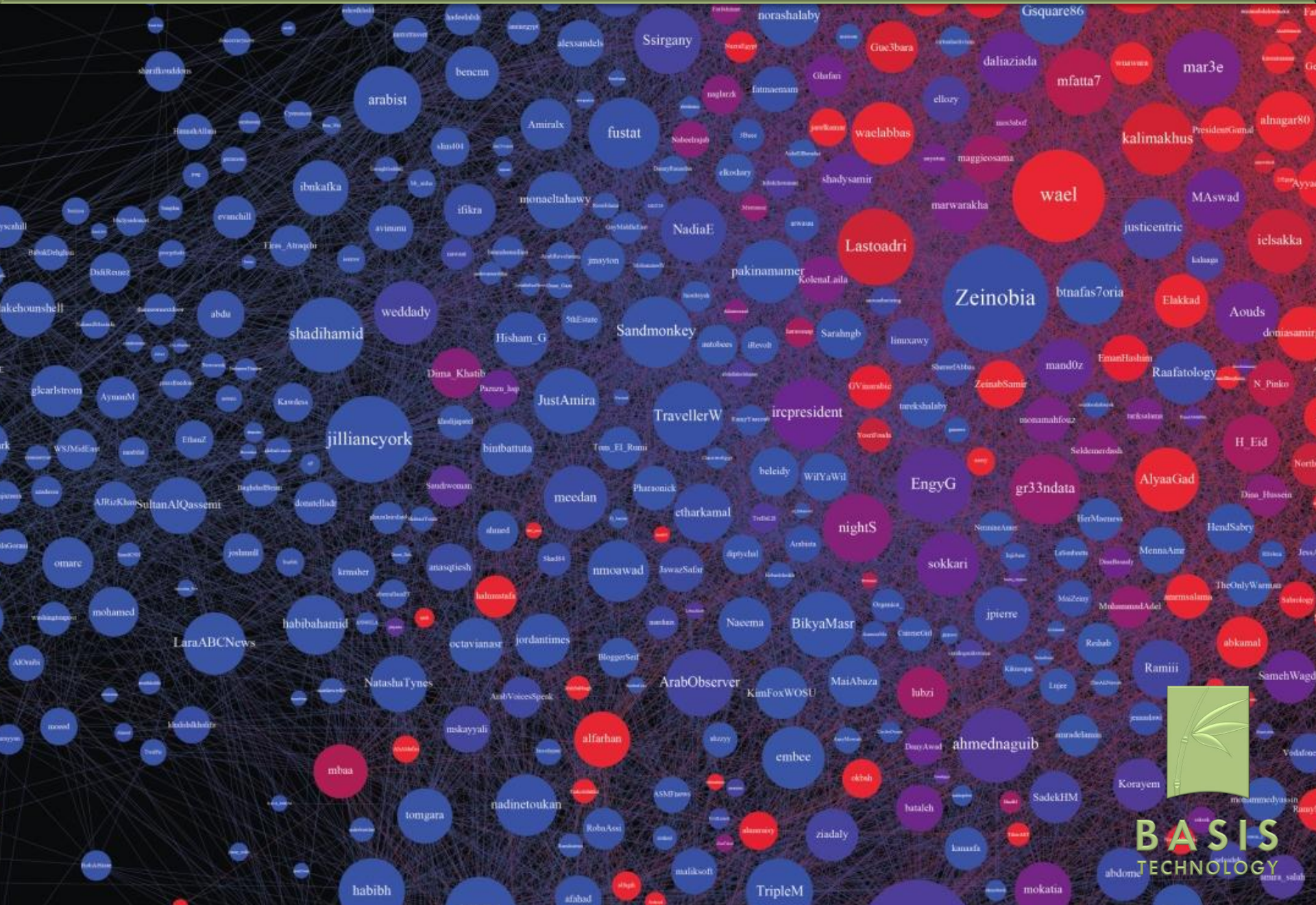
Basis Technology

VP, Defense & Intelligence Services





BASIS
TECHNOLOGY

Egyptian Revolution in Twitter



What is “Chat”?

Informal text language used primarily in online communications:

- Discussion boards
- Chat rooms (AIM, Second Life, IRC, JDate, etc.)
- SMS
- Facebook 
- Emails
- Twitter 
- MySpace (20th century) 



Arabic Chat Sample

Messages: Simplify | Expand (Group by Topic)

Author

1 **Hi**

Firas

rita24smile



Welcome to this group...

2 **I want your opinion**

Guitta

guittafaccoul



Hi Ana Guitta, Hebe 2agrif chou ra2ykon bi hal group. W kamen hebe kil wahad, iza baddo, y3ote ra2yo bi balado libnan....

6 **Re: I want your opinion**

Jean_Lannoury1



Libnen ya out3it sama, chou 3imlo fik ya balade, 7abbe min trabak biknouzidini, b7ibbak ya libnen ya watane. W bazed..... Ana bwefi2 3ala fikrit inno lmessages...

8 **Re: I want your opinion**

gabrylgahleh



ana bshaj3ak ya jean w bou2af ma3ak, leizim tota7 "a poll" zinweino inta ma3 aw dod il ra2abi 3al mail heik minshouf ra2y il aktariyi w min koun 7afazna 3a...



Significance of Arabic Chat

Friday, 03 June 2011 | Rajab 2, 1432 | Last updated at 15:49

LOG IN | REGISTER

arab news.com

The Middle East's Leading English Language Daily

HOME SAUDI ARABIA MIDDLE-EAST WORLD ECONOMY SPORTS LIFE & STYLE OPINION

RSS

Law & You Italy National Day

HOME / SAUDI ARABIA / 'ARABIZI IS DESTROYING THE ARABIC LANGUAGE'

SEARCH

'Arabizi is destroying the Arabic language'

By RENAD GHANEM | ARAB NEWS

Published: Apr 19, 2011 22:25 Updated: Apr 19, 2011 22:25

JEDDAH: Arabizi, a term that describes a system of writing Arabic in English, is now more popular than ever, especially online.

Parents and teachers are becoming more concerned over the popularity of this new trend. Some see it as a threat to the Arabic language.

A non-English speaker does not need to speak the language to communicate with others in Arabizi. Numbers are also mixed in Arabizi to represent some letters in Arabic, such as 2, 5, 6, 7 and 9.

Most Arab Internet users find this way easier than typing in Arabic. Teachers fear that this will weaken their Arabic language ability or even replace the language in the future. Arabic professional professors from the Arab world consider it a war against the Arabic language to make it disappear in the long run.

SECTION LIST

- Sharing husbands' burden irks wives
- With fecup-reading, women seek to get to 'bottom' of problems
- Drunk driving: 243 licenses seized this year
- Prince Sultan honors Mobily for DCA support
- Making Haj, Umrah more comfortable
- Yemeni crisis: Obama's aide mobilizes GCC
- Spreading sweetness at fest
- 90% absence of students denied

Problems Understanding Chat



BASIS
TECHNOLOGY

English Chat Examples

LOL! AFAIK he was the last **1 2** C his gf. She is a QT IMHO but a real PITA. L**8**R. ;-)

Laughing out loud! As far as I know, he was the last **one to** see his girlfriend. She is a cutie, in my humble opinion, but a real pain in the abdomen. L**ate**. [wink]

Quite humourous! I believe he was the last person to observe his girl. I feel she is pretty, but a bother. Regards.
[nod, nod; wink, wink; say no more]



Numerals as Letters

0=O

1=I

3=E

5=S

6=b

8=B

9=g

*ABS**5**OLUTELY*

ab50lutely

*a**6**so**1**ute**1**y*

a6501ute1y

***8A515** TECHNOLOGY*

8A515 TECHNOLOGY



BASIS
TECHNOLOGY

Numerals as Letters

- 7H15 M3554G3 53RV35 7O PR0V3 H0W 0UR M1ND5
C4N D0 4M4Z1NG 7H1NG5! 1MPR3551V3 7H1NG5!
- THIS MESSAGE SERVES TO PROVE HOW OUR MINDS
CAN DO AMAZING THINGS! IMPRESSIVE THINGS!



Problems With Arabic Chat



Arabic Chat ≠ Standardization

Arabic Chat differs from Modern Standard Arabic (MSA), the standard used in most news media and government reportage.

MSA

- Standardized across Arabic world; no regional terms, grammar
- Literary form of Arabic

Arabic Chat

- Colloquial Arabic
- Dialectal influences – regional vocabulary, grammar, pronunciation
- Switching between languages, dialects, and alphabets
- Chat specific expressions, abbreviations, and emoticons
- Exaggerations in punctuation and letters
- Typos



Numerals as Letters

Arabic	Arabic Chat	Latin
ع	3	A
ح	7	H
ط	6	T
ص	9	S
ء	2	A
خ	5	KH
ق	8	Q



Arabic Chat Example

Hello Abu Masud. Hey, it's been a long time since I've seen you, sheikh.

marhaba Abu Masud Wallahi mudda laweela Imma shuftak ya sheikh

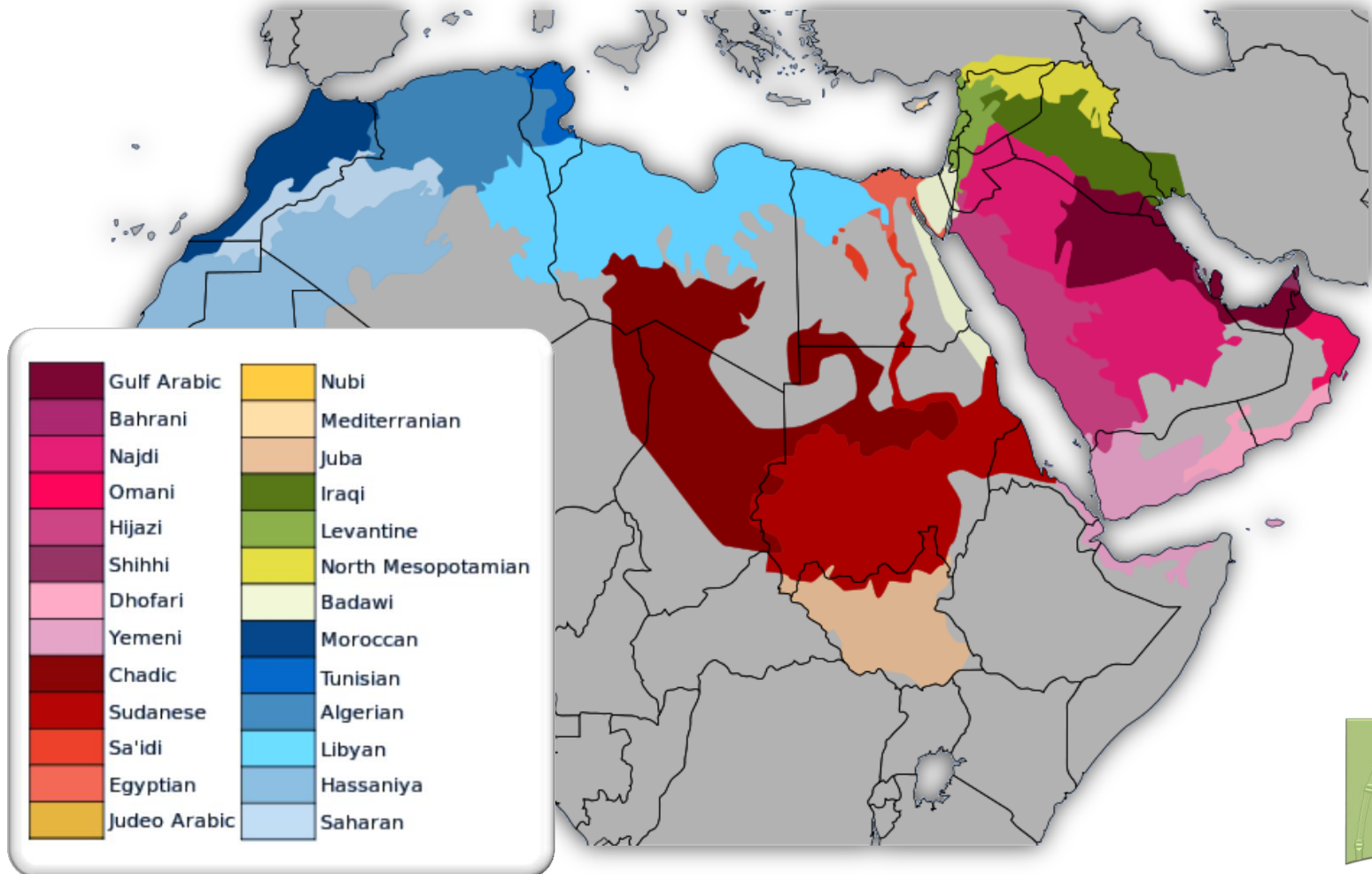
مرحبا أبو مسعود واللهي مدة طويلة لما شفتك يا شيخ



BASIS
TECHNOLOGY

Arabic Dialects

Arabic is spoken by 450 million people, and is the official language in 25 countries.



Dialectal Problems

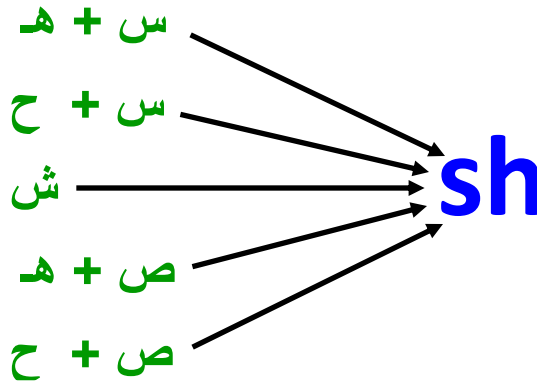
The same word is often pronounced differently in various dialects. Therefore, they are spelled differently as well.



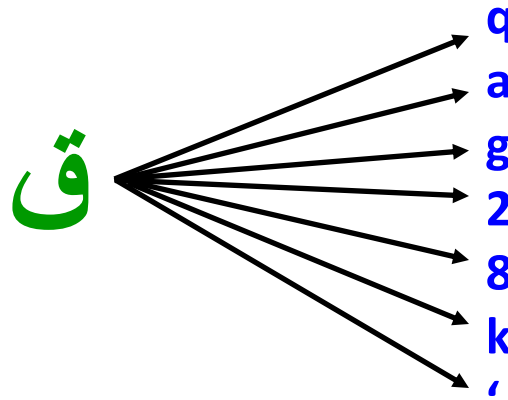
Ambiguous Transliteration

Rendering an Arabic character in a Romanized form can cause ambiguity problems.

Many → One



One → Many



Ambiguous Transliteration

Qasim	=	قاسم
Gasim	=	قاسم
8asim	=	قاسم
Kasim	=	قاسم
Asim	=	قاسم
2asim	=	قاسم
'asim	=	قاسم



Resolving the Problems



What Must a Language Tool Do?

Romanized Arabic text must be converted into Arabic script.

rayi7 2abilhom → رايح قابلهم

It needs to handle:

- Variant spellings based on dialects or typos
- Numerals substituting for letters
- Insertion of non-Arabic words
- Dialectal slang



❖ Algorithms

- Break down words into morphological components and phonemes, producing transliteration candidates
- Candidates ranked by metrics (popularity of phoneme mappings, frequency of output in Arabic text, etc.).

❖ Statistics

- Use a large database of Roman alphabet spellings generated from the input of millions of Arabic speakers online



Language- and Alphabet-Switching

- List of most popular non-Arabic words found in chat statistics (primarily English and French)
- Filter transliteration candidates through list to create a “do not transliterate” choice
- User can add ad hoc words to list



Results Combined With Other Tools



Rosette Chat Translator

Ahmad: Keefak Imad? ana sme3t alkhbar 3an albarnamaj aljadeed min Basis Technology
Imad: 6ab3an! Basis Technology min afdhal alsharikat fe haza almajal
Ahmad: 3endahum barnamaj jadeed letarjamat aldardasha al3arabiyya
Imad: ana jarabtah, fantastique!

Convert chat to Arabic

MT Enhancement

أحمد : كيفك عماد ؟ أنا سمعت الخبر عن البرنامج الجديد من Basis Technology
عماد : طبعاً ! Basis Technology من أفضل الشركات في هذا المجال
أحمد : عندهم برنامج جديد لترجمة الدردشة العربية
عماد : أنا جربته , ! fantastique

English Translation

Ahmed: Hi Emad? I heard the news about the new program from Basis Technology
Emad: Of course! Basis Technology of the best companies in this area
Ahmad: They have a new program to translate the Arabic chat
Emad: I tried it, fantastique!

Named Entity Extraction Results

PERSON - ORGANIZATION - LOCATION - GPE -
FACILITY - NATIONALITY - RELIGION - TEMPORAL
IDENTIFIER - TITLE - OTHER

أحمد : كيفك عماد ؟ أنا سمعت الخبر عن البرنامج الجديد من Basis Technology
عماد : طبعاً ! Basis Technology من أفضل الشركات في هذا المجال
أحمد : عندهم برنامج جديد لترجمة الدردشة العربية
عماد : أنا جربته , ! fantastique

Text Analysis

Entity Extraction

name Translation



Automated Arab Chat Conversion: Benefits

- Prepares Arab chat to be analyzed by existing data processing programs
- Increases production (no need to rely on human translators only)
- Enables identification of writer's origins
- Standardizes chat text



انہلا بسا!

For more information

Visit www.basistech.com

Write to joelr@basistech.com

Call 617-386-2090 or 800-697-2062



BASIS
TECHNOLOGY