The Conundrum of New Online Languages: Translating Arabic Chat
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)
1. Hi
   Welcome to this group...

2. I want your opinion
   Hi Ana Guitta, Hebe 2a3rif chou ra2ykon bi hal group. W kamen hebe kil wahad, iza baddo, y3ote ra2yo bi balado libnan....

6. Re: I want your opinion
   Libnen ya out3it sama, chou 3imlo fik ya balade, 7abbe min trabak biknouzidini, b7ibbak ya libnen ya watane. W ba3ed..... Ana bwefi2 3ala fikrit inno lmessages...

8. Re: I want your opinion
   ana bshaj3ak ya jean w bou2af ma3ak, leizim totra7 "a poll" 3inweino inta ma3 aw dod il ra2abi 3al mail heik minshouf ra2y il aktariyi w min koun 7afazna 3a...
‘Arabizi is destroying the Arabic language’

By RENAD GHANEM | ARAB NEWS

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.
Problems Understanding Chat
LOL! AFAIK he was the last one to see his girlfriend. She is a cutie, in my humble opinion, but a real pain in the abdomen. Later. [wink]

Laughing out loud! As far as I know, 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=l  3=E  5=S  6=b  8=B  9=g

Absolutely
ab50lutely
ab50lutely
a6501ute1y
BASIS TECHNOLOGY
8A515 TECHNOLOGY
• 7H15 M3554G3 53RV35 7O PR0V3 H0W OUR 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 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
- Dialectical 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

<table>
<thead>
<tr>
<th>Arabic</th>
<th>Arabic Chat</th>
<th>Latin</th>
</tr>
</thead>
<tbody>
<tr>
<td>ع</td>
<td>3</td>
<td>A</td>
</tr>
<tr>
<td>ح</td>
<td>7</td>
<td>H</td>
</tr>
<tr>
<td>ط</td>
<td>6</td>
<td>T</td>
</tr>
<tr>
<td>ص</td>
<td>9</td>
<td>S</td>
</tr>
<tr>
<td>م</td>
<td>2</td>
<td>A</td>
</tr>
<tr>
<td>خ</td>
<td>5</td>
<td>KH</td>
</tr>
<tr>
<td>ق</td>
<td>8</td>
<td>Q</td>
</tr>
</tbody>
</table>
Hello Abu Masud. Hey, it’s been a long time since I’ve seen you, sheikh.
Arabic is spoken by 450 million people, and is the official language in 25 countries.
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 $\rightarrow$ One

One $\rightarrow$ Many

$ش + ه$ $\rightarrow$ \textit{sh}

$ق + ج$ $\rightarrow$ $\textit{g}$, $\textit{k}$, $2$, $8$
Resolving the Problems
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
Disambiguation

 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
• 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 alkhabar 3an albarnamaj aljadeed min Basis Technology
Imad: 6ab3an! Basis Technology min afdhal alsharikat fe haza almajal
Ahmad: 3endahum barnamaj jaded idtarjamat aldardasha al3arabiyya
Imad: ana jarabta, fantastique!

Convert chat to Arabic

Basis Technology

Ahmad: كيفك عماد؟ أنا سمعت الخبر عن البرنامج الجديد من Basis Technology
عماد: طبعًا! من أفضل الشركات في هذا المجال
Ahmad: 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 - SPE -
FACILITY - NATIONALITY - RELIGION - TEMPORAL
IDENTIFIER - TITLE - OTHER

لحمد: كيفك عماد؟ أنا سمعت الخبر عن البرنامج الجديد من Basis Technology
عماد: طبعًا! من أفضل الشركات في هذا مجال
عماد: عندهم برنامج جديد لترجمة الدردشة العربية
عماد: أنا جربته، fantastique!

BASIS TECHNOLOGY
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
For more information
Visit www.basistech.com
Write to joelr@basistech.com
Call 617-386-2090 or 800-697-2062