Tunisian Arabic Corpus

Design and Progress

Karen McNeil
Georgetown University

Miled Faiza
University of Virginia
Why Corpora?

- **Language Resources**
  - Lexicography
  - Grammar
  - Authentic Teaching Materials

- **Natural Language Processing**
  - Automatic Speech Recognition (ASR)
  - Language ID
  - Machine Translation
A Linguistic Map of Arabic
## Why Spoken Arabic?

<table>
<thead>
<tr>
<th>Percentage of Overlap</th>
<th>Conv Egyptian vs. Media MSA</th>
<th>Conv British vs. Media American</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unigrams</td>
<td>10.3 %</td>
<td>44.5 %</td>
</tr>
<tr>
<td>Bigrams</td>
<td>1 %</td>
<td>19.2 %</td>
</tr>
<tr>
<td>Trigrams</td>
<td>&lt; 1 %</td>
<td>5.3 %</td>
</tr>
</tbody>
</table>

**Kirchhoff 2005:41**

<table>
<thead>
<tr>
<th>Error Rate</th>
<th>Egyptian Only</th>
<th>Egyptian + MSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Development Set</td>
<td>56.1 %</td>
<td>54.8 %</td>
</tr>
<tr>
<td>Evaluation Set</td>
<td>42.7 %</td>
<td>41.4 %</td>
</tr>
</tbody>
</table>

**Kirchhoff 2005:49**
Spoken Arabic Corpora

- CALLFRIEND Egyptian Arabic (1996)
- CALLHOME Egyptian Arabic Speech (1996)
- Fisher Levantine Arabic Conversational Telephone Speech (2005)
- Gulf Arabic Conversational Telephone Speech (2006)
- Iraqi Arabic Conversational Telephone Speech (2006)
- Levantine Arabic Conversational Telephone Speech (2006)
- BBN/AUB DARPA Babylon Levantine Arabic Speech and Transcripts (2005)
Why Tunisian?

- So-far neglected dialect
  - No resources available in English
- Linguistically interesting
  - Sociolinguistics / Codeswitching
- North African
Sources

- Traditional Written Sources
  - Folklore
  - Songs / Folk Poetry
  - Proverb collections
  - Screenplays
- New Written Sources
  - Blogs
  - Email
  - Facebook
- Transcribed Audio
  - Radio
### Manage Batches

#### Batches in progress (1)

<table>
<thead>
<tr>
<th>Batch Name</th>
<th>Created</th>
<th>Assignments Completed</th>
<th>Time Elapsed</th>
<th>Estimated Completion Time</th>
<th>Effective Hourly Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>'OCX' @ 04 Apr 2011</td>
<td>April 04, 2011</td>
<td>68 / 141</td>
<td>about 20 hours</td>
<td>April 06, 2011 9:02 PM PDT (Wednesday)</td>
<td>$1.26</td>
</tr>
</tbody>
</table>

**Average Time per Assignment:** 9 minutes 30 seconds

**Batch Progress:**
- 41% submitted
- 100% published

#### Batches ready for review (3)

<table>
<thead>
<tr>
<th>Batch Name</th>
<th>Created</th>
<th>Assignments Completed</th>
<th>Time Elapsed</th>
<th>Estimated Completion Time</th>
<th>Effective Hourly Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>'Long Audio Transcription' @ 16 Feb 18:29</td>
<td>February 16, 2011</td>
<td>1 / 1</td>
<td>7 days</td>
<td>COMPLETE</td>
<td>$7.41</td>
</tr>
</tbody>
</table>

**Average Time per Assignment:** 6 hours 44 minutes 47 seconds

**Batch Progress:**
- 100% submitted
- 100% published
### 'Post Tagging' @ 14 Feb 17:36

688 of 1161 assignments (FILTER APPLIED: only show assignments that are in 'Approved' status)

<table>
<thead>
<tr>
<th>HIT ID</th>
<th>Worker ID</th>
<th>Lifetime Approval Rate</th>
<th>Input.Text</th>
<th>Input.Forum Text</th>
<th>Comment</th>
<th>Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>100AASYU110370788DJYHNCY749P</td>
<td>ACKRAJS6LU85BE</td>
<td>100% (172/172)</td>
<td>786</td>
<td></td>
<td>Furniture Expenses Lifelong</td>
<td></td>
</tr>
<tr>
<td>100AASYU110370788DJYHNCY794U</td>
<td>A38MDQ64DOQ92L</td>
<td>100% (214/215)</td>
<td>1413</td>
<td></td>
<td>Israel problems sports TV</td>
<td></td>
</tr>
<tr>
<td>104J4QUF3R47VGNCCU6ZCJ3VUNMAJ3E</td>
<td>A212GKBAXUP5IM</td>
<td>100% (83/83)</td>
<td>102</td>
<td></td>
<td>alert reasoning answering</td>
<td></td>
</tr>
<tr>
<td>108BURZW0YLPFQZ8TD1850Y54VRNMU</td>
<td>ACKRAJS6LU85BE</td>
<td>100% (172/172)</td>
<td>715</td>
<td></td>
<td>security, system, gratitude</td>
<td></td>
</tr>
<tr>
<td>10CN45L0SJ54G4YGBAZJ8E7C8JSJL</td>
<td>A38MDQ64DOQ92L</td>
<td>100% (214/215)</td>
<td>1037</td>
<td></td>
<td>nationalism brotherhood advertisements</td>
<td></td>
</tr>
<tr>
<td>10OCXIMZNNSTG88NZY49OT9K6GX42X</td>
<td>A38MDQ64DOQ92L</td>
<td>100% (714/715)</td>
<td>1235</td>
<td></td>
<td>Quran Money poverty</td>
<td></td>
</tr>
<tr>
<td>10M4F55EFC1EYGBX4QAT7KIMOD</td>
<td>ACKRAJS6LU85BE</td>
<td>100% (214/215)</td>
<td>1048</td>
<td></td>
<td>Saudi Arabia Administrative</td>
<td></td>
</tr>
</tbody>
</table>
Genres / Balance

- Blogs: 8000
- Internet Forums: 85561
- Plays: 15662
- Poems: 404
- Political Speeches: 768
- Proverbs: 3681
- Interviews: 6370
- Songs: 1216
- News: 1748
- Sports: 765
- Folktales: 14401
- Conversations: 2652
- Drama: 142030
Spelling

برشة
برشة
برشة
Spelling

699

33

563
Normalization

Linguistic Processing

Normalization

Corpus
TAC Management Tool

- **Requirements:**
  - Allow easy management of Corpus and files
  - Allow remote collaboration
  - Flexible and expandable
  - Attractive and High Usability

- **Technical Specs**
  - Python
  - Django (Web Framework)
    - Model – View – Controller (MVC) separation
Database Structure

- Categories
- Files
- Details
- Tags
- Dates
- Authors

- Texts
تونسية

* www.tunisiya.org