Collecting Natural SMS & Chat Conversations in Multiple Languages: BOLT Phase 2 Corpus

The Broad Operational Language Translation (BOLT) Program is developing technology that enables English speakers to retrieve and understand information from informal foreign language sources including chat, text messaging (SMS) and spoken conversation.

**BOLT Phase 2 SMS/Chat Collection Requirements**
- **Multilingual**: English, Chinese, Egyptian Arabic
- **Large volume**: at least 2 million words/language
- **Conversational**: naturally-occurring, two-way, informal SMS and chat messages

Prior SMS/Chat Corpora are primarily one-sided conversations or public chat room transcripts with limited content.

**SMS/Chat Collection System**
Handles real-time collection and message archive donation.

- Users can delete or redact words, messages, full threads prior to upload
- Personal identifying information (PII) in message headers automatically stripped upon upload
- Users can withdraw data at any time

**Collection Result**

<table>
<thead>
<tr>
<th>Language</th>
<th>Active Users</th>
<th>Collect Words</th>
<th>Aud Count Rate</th>
<th>Final Words</th>
<th>Final Msgs</th>
<th>Final Convs</th>
<th>Avg Messages/Conv</th>
<th>Avg Words/Message</th>
<th>Avg Words/Conv</th>
<th>Avg Words/User</th>
</tr>
</thead>
<tbody>
<tr>
<td>Egyptian</td>
<td>26</td>
<td>690K</td>
<td>69%</td>
<td>475K</td>
<td>119,00</td>
<td>2140</td>
<td>56</td>
<td>4</td>
<td>222</td>
<td>48,051</td>
</tr>
<tr>
<td>Chinese</td>
<td>77</td>
<td>3.7M</td>
<td>54%</td>
<td>2M</td>
<td>306,99</td>
<td>7844</td>
<td>39</td>
<td>7</td>
<td>255</td>
<td>21,710</td>
</tr>
<tr>
<td>English</td>
<td>152</td>
<td>3.3M</td>
<td>79%</td>
<td>2.6M</td>
<td>212,000</td>
<td>9155</td>
<td>23</td>
<td>12</td>
<td>284</td>
<td>27,600</td>
</tr>
<tr>
<td>Total/Average</td>
<td>255</td>
<td>7.69M</td>
<td>67%</td>
<td>5.075M</td>
<td>212,000</td>
<td>19,139</td>
<td>39.33</td>
<td>7.67</td>
<td>253.67</td>
<td>32453.67</td>
</tr>
</tbody>
</table>

**Linguistic Features of Collected Data**

### General Features
- Natural and conversational
- Internet slang and novel abbreviations
- Nonstandard use or complete absence of punctuation
- Frequent typographical errors, misspellings and missing spaces
- Widespread use of simple and complex initialisms
- Prevalent pro-drop

### Arabic Features
- Code mixing, neologisms
- Multiple orthographies
- 63% Arabizi
- 13% Arabic
- 24% Mixed
- Arabizi data semi-automatically transliterated into Arabic script for use in BOLT
- Joint work with Arabic Dialect Group at Columbia University

**Acknowledgement**: This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-11-C-0145. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.