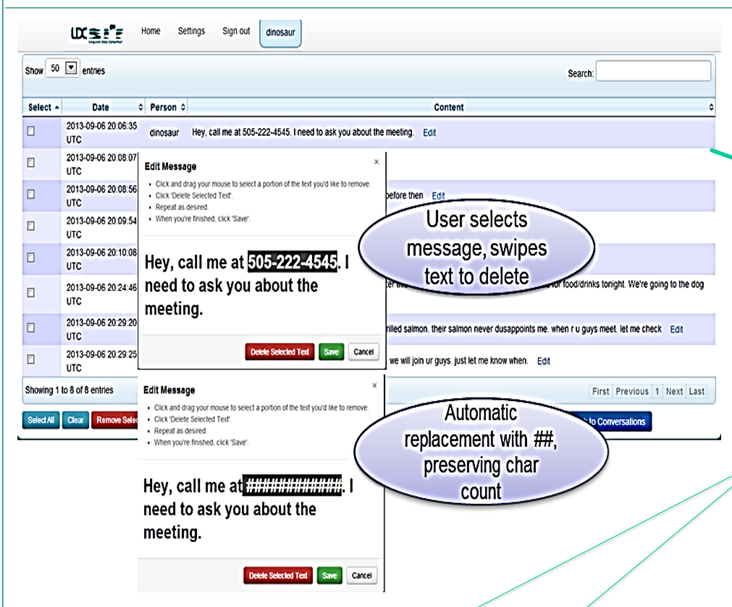
Collecting Natural SMS & Chat Conversations in Multiple Languages: BOLT Phase 2 Corpus

Z. Song, S. Strassel, H. Lee, K. Walker, J. Wright, J. Garland, D. Fore, B. Gainor, P. Cabe, T. Thomas, B. Callahan, A. Sawyer

Linguistic Data Consortium

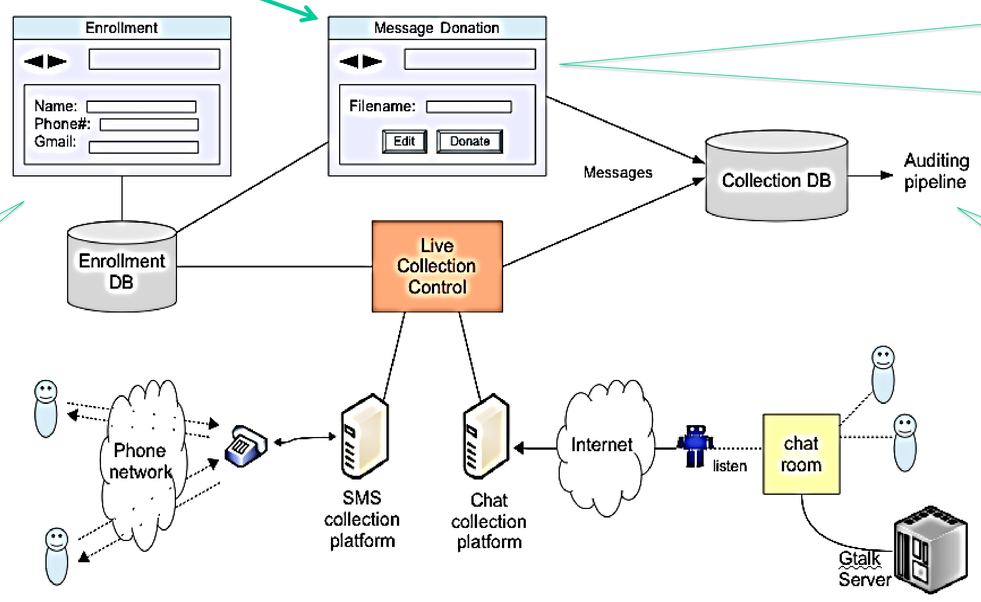
- **The Broad Operational Language Translation (BOLT) Program** is developing technology that enables English speakers to retrieve and understand information from informal foreign language sources including chat, text messaging (SMS) and spoken conversation
- **BOLT Phase 2 SMS/Chat Collection Requirements**
  - **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



- Conventional recruiting techniques including advertising, networking
- Collaboration with in-country partners, especially local Universities
- Participation is entirely anonymous

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

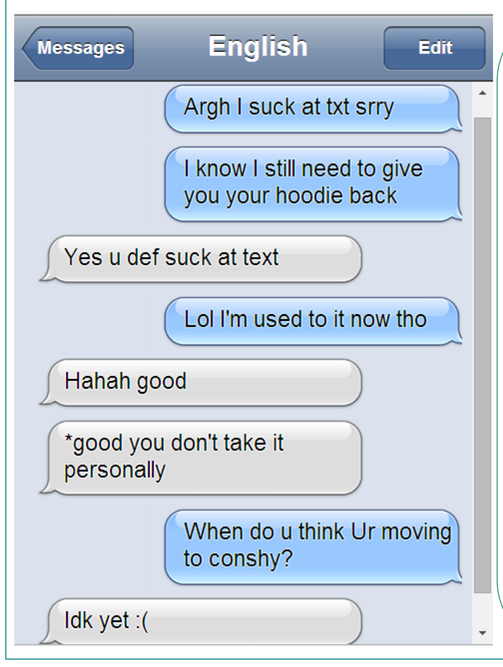
- Messages manually audited by certified staff to exclude PII and sensitive content prior to inclusion in corpus
- Messages, threads are selected and segmented prior to downstream translation and annotation

Collection Result

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

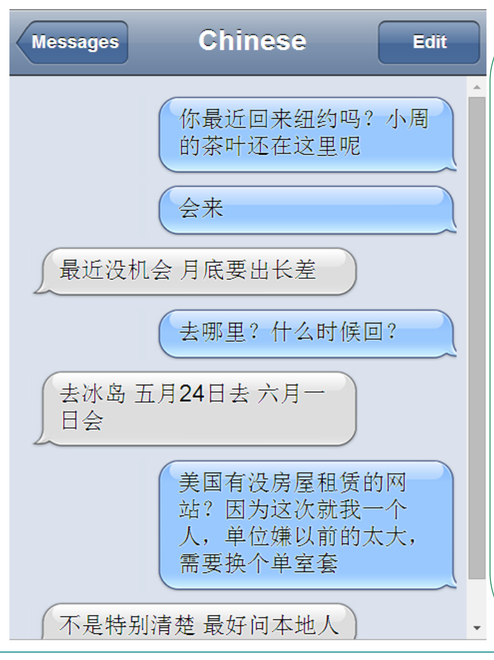
Language	Collection Method		Genre	
	Live	*Donation*	*SMS*	*Chat*
Egyptian	4%	96%	39%	61%
Chinese	1%	99%	4%	96%
English	6%	94%	94%	6%

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.