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

Zhiyi Song, Stephanie Strassel, Haejoong Lee, Kevin Walker, Jonathan Wright, Jennifer Garland, Dana Fore, Brian Gainor, Preston Cabe, Thomas Thomas, Brendan Callahan, Ann Sawyer

Linguistic Data Consortium, University of Pennsylvania

3600 Market Street Suite 810, Philadelphia, PA, 19104, USA

{zhiyi,strassel,haejoong,walker,jdwright,garjen,foredana,bgainor,cabep,sawyera}@ldc.upenn.edu,
{brendan.callahan, thomas.brooks.thomas}@gmail.com

Abstract

The DARPA BOLT Program develops systems capable of allowing English speakers to retrieve and understand information from informal foreign language sources. Phase 2 of the program required large volumes of naturally occurring informal text (SMS) and chat messages from individual users in multiple languages to support evaluation of machine translation systems. We describe the design and implementation of a robust collection system capable of capturing both live and archived SMS and chat conversations from willing participants. We also discuss the challenges recruitment at a time when potential participants have acute and growing concerns about their personal privacy in the realm of digital communication, and we outline the techniques adopted to confront those challenges. Finally, we review the properties of the resulting BOLT Phase 2 Corpus, which comprises over 6.5 million words of naturally-occurring chat and SMS in English, Chinese and Egyptian Arabic.

Keywords: SMS, chat, data collection, informal genres, multilingual corpora

1. Introduction

DARPA's Broad Operational Language Translation (BOLT) program is aimed at developing technology that enables English speakers to retrieve and understand information from informal foreign language sources including chat, text messaging and spoken conversations. The genres of interest to BOLT are characterized by inherent variation and inconsistency, motivating the development of a new breed of collection and annotation methods.

In BOLT's first phase, LDC collected and annotated large volumes of online discussion forum data using techniques adapted from prior collection efforts (Garland et al., 2012). In this paper we describe the creation of the BOLT Phase 2 Corpus, consisting of large volumes of naturally occurring informal text (SMS) and chat messages from individual users in multiple languages. We describe MCol, a robust collection system capable of collecting live SMS and chat conversations between enrolled participants in real time, and permitting willing participants to contribute existing chat and SMS messages directly from their smartphones or computers. After reviewing existing approaches to informal data collection (Section 2), we describe the BOLT Message Collection System (Section 3) including the novel technical solutions employed. In Section 4 we describe our approach to addressing the challenges of participant recruitment that arose during this collection, in particular focusing on the issue of gaining trust from individuals who have deep and growing concerns about personal privacy. Section 5 reviews the procedures used to audit collected data and select it for subsequent translation and annotation. Section 6 describes the resulting BOLT Phase 2 SMS/Chat Corpus, presenting details of the collected data, while Section 7 discusses some of the linguistic features that present particular challenges to downstream annotation tasks and to the BOLT technology itself. Finally we present concluding remarks in Section 8.

2. Prior Collection Efforts

There have been a number of previous efforts to collect SMS and/or chat messages; these have varied in collection techniques, language and data focus. The NUS SMS collection project created a large-scale SMS corpus by developing a Google Android application allowing users to automatically deposit messages to the corpus (Chen and Kan, 2011). The sms4Science project lowered the technical barrier to donation by letting users forward their messages directly to a central number, free of charge (Fairon & Paumier, 2006). The SoNaR project adopted a combination of the NUS SMS and sms4Sciencedata collection methods (Treurniet, et al, 2012). Finally, the SoNaR chat and tweets project collected online chat from one open chat channel, as well as capturing real-time chats from consented users using several chat clients (Sanders, 2012). Due to privacy concerns and ethical considerations, all three SMS corpora limited collection to messages provided by the sender, resulting in single-sided conversations. The SoNaR chat and tweets corpus contains all conversation sides.

The BOLT Phase 2 corpus required not only two-sided conversation, but also required that the corpus include primarily naturally-occurring data rather than conversations staged for the corpus building effort. Moreover, it was desirable for the corpus to comprise primarily SMS rather than chat messages. Given these constraints combined with the large data volume targets (a minimum of 2 million words in each of English, Chinese and Egyptian Arabic) and multilingual focus, a new collection approach was adopted.

3. Message Collection System

3.1 Overview

To enable creation of the BOLT Phase 2 corpus, we developed a robust message collection system, integrating live collection (real-time capture of SMS or chat messages between pairs of consented, enrolled users) with donation collection (user contribution of archived SMS or chat messages). The live collection component, known as MCol, is an extension of LDC's existing telephone speech collection system, used to build the MIXER corpora (Cieri et al, 2007; Brandschain et al, 2008). A new module was added for text and chat message collection in which a bot initiated conversations by sending an invitation to a pre-enrolled pair of users and prompting them to text or chat via a specified client. The donation collection, on the other hand, required an entirely new system capable of accepting donations of SMS and chat message archives from users' phones and computers across multiple platforms, clients and apps. Both collection modes were supported by a customized user screening and enrollment framework as well as backend databases for enrollment and collection. Figure 1 shows the architecture of the integrated system.

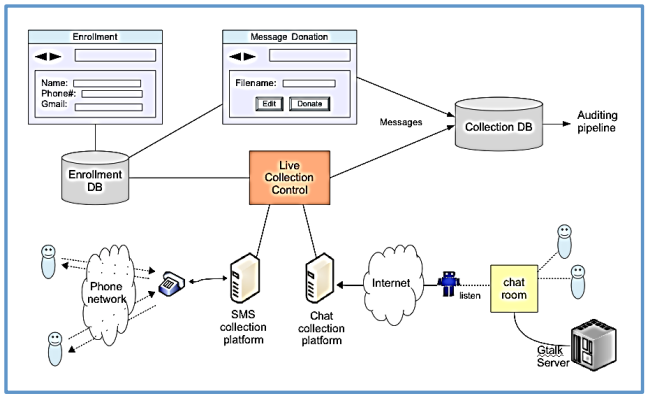

Figure 1: Message Collection System Architecture

3.2 Participant Enrollment

A customized page within LDC's WebAnn framework (Wright et al. 2012) was created for users to sign up, provide their consent to participate and enroll in the collection. This front end handled user authentication and connection to the generalized enrollment infrastructure. Users could choose whether to participate in live collection, message donation or both. Users' real names were not collected; each enrolled user was assigned a unique ID and could choose a username that would identify them to other participants. Users were asked to provide scheduling and contact information (e.g. phone number to be used for live SMS collection, user ID for live chat collection). Demographic information and other personal information could be provided if the user so desired. All personal identifying information was stored in a separate, secure database that is never linked to the corpus data.

3.3 Live Collection in Real Time

Users who enroll in the live collection indicate both their preferred conversation partners (by username, if people known to each other enroll together) and their availability for live chat or SMS conversations. The MCol system regularly checks the enrollment database to identify available participants for a given time window. For SMS, once MCol has chosen a pair of conversation partners, a session is initiated and the designated participants receive a text message from the MCol bot inviting them to start a conversation. Users reply to MCol's invitation text message by proceeding to text with one another. MCol intercepts each user's message and relays it transparently to the other party. The user's experience is the same as in normal text messaging; messages are delivered and appear in real time within the user's native texting app. However, users do not see their conversation partner's phone number; instead they see the MCol platform number and their partner's selected username. During live chat collection the MCol bot creates a chat room and adds designated participants to the room, where they can exchange messages. The bot simply "sits" in the room and records the exchanged messages.

Because BOLT is focused on conversations in specific languages, participants in the live collection were required to take a simple pre-screening test to verify their language ability. Collected live conversations are also passed through a simple Language ID system to catch potential cheating.

3.4 Collection of Existing Data

Participants who enroll in the donation collection upload their existing SMS and chat messages using a simple web interface created for the project. Because of the wide range of phone systems and chat clients used by potential participants, LDC conducted surveys prior to collection to identify the most popular systems and apps among user populations (languages and countries) of interest. As a result we focused our efforts on supporting message donation from a range of clients and apps, including:

- SMS: iMessage, Android SMS, Symbian SMS, Viber, BlackBerry
- Chat: WhatsApp, QQ, Google chat, Skype chat, Yahoo Messenger

For each client and app we created a step-by-step tutorial showing users how to locate existing SMS and chat messages on their device, create an archive file, and export the file for upload to LDC's collection platform. We also provided live, multilingual help desk support via email, phone, text message and chat for users who preferred hands-on guidance.

For each of the supported apps we developed custom parsers to process the incoming message archives. The user's data archive is first uploaded to a temporary holding tank, where an automated process detects its file format and selects the appropriate message parser. The parser then divides the archive into individual conversations and performs some simple sanity checks. Any personal identifying information contained in message metadata (e.g. phone numbers or usernames) is automatically removed during parsing. The parsed conversations are then presented back to the user in a simple web GUI that allows the user to edit or remove any part of the archive they do not wish to donate. (See Section 4 for additional details). After the user is

satisfied with the edited archive, they click a button to allow it to be uploaded to the collection database. Only conversations that the user explicitly approves are stored in the database; the original unedited archive is deleted from the temporary holding tank.

3.5 Data Validation

All collected conversations (whether live or donated) are saved to the collection database where they are subject to post-hoc automatic validation, including language identification and duplicate content detection. Conversations not in the target languages are flagged as such and are subject to manual review. Occasional duplicated conversations are found; users may accidentally upload identical or overlapping archives, some may try to game the system, and parties known to each other maybe each contribute a shared conversation. These duplicates are automatically detected and flagged. Single-sided conversations are also flagged for removal. Conversations that pass the validation stage are migrated to a centralized conversation and message database where they can be accessed for manual auditing, selection and segmentation (see Section 5).

4. User Recruitment, Participation Incentives and Privacy Protection

Previous collection efforts for chat and SMS data have relied on a variety of techniques to recruit participants, including advertising through national media and use of personal and professional acquaintance networks (Fairon & Paumier, 2006; Sanders, 2012; Taggs, 2009). We employed these same methods, but also found it necessary to extend the recruitment approach to address particular challenges associated with the multilingual nature of the collection. We recruited internationally, looking primarily to low-cost and online advertising forums including social network sites, email lists and Craigslist. We also relied heavily on professional and personal networks, with focused outreach among international college and graduate student populations, since this group of users was expected to be most familiar with text and chat messaging.

These standard recruitment methods worked well for the English collection and we quickly reached our goals; this was partially due to our initial outreach focus on people who had recently participated in other LDC data collections. Collection for Chinese and especially for Egyptian Arabic was much slower to develop and more challenging to achieve. The original compensation model for the collection was designed to encourage a high level of participation for each user. For every 50 messages contributed, the participant earned one entry in a weekly drawing for a monetary incentive; thus, a user donating 125 messages would have their name entered twice into each weekly drawing, while a user donating 875 messages would have their name entered 17 times. In order to count toward the drawing, messages were required to be primarily in the participant's target language and could not be repeats of messages already contributed by this participant. Once a participant was entered into the weekly drawing pool, they remained in the pool for the duration of the study.

English participants responded well to this incentive model and we were able to exceed the 2 million word target relatively quickly, but Chinese and Egyptian participants required a different model. For some potential participants in these languages the mere possibility of winning was insufficiently motivating, and for others, earning entries to win monetary "prizes" was culturally prohibited. In response, we adopted a tiered per-message compensation model for Egyptian participants. For Chinese, to supplement the weekly drawings we added guaranteed compensation bonuses for the most productive participants.

With added recruitment time and effort we were eventually able to achieve collection targets for Chinese. The majority of Chinese participants were friends and family of our recruitment staff; having a personal connection to the study gave participants some additional assurance that their privacy would be protected. Anonymity was an especially deep concern for participants physically located in China.

Egyptian recruitment was by far the most challenging. Despite a massive recruitment effort (sending thousands of emails to organizations and individuals, online advertising, social media, etc.), the perception of personal risk was too high for most Egyptians to feel comfortable sharing their data. The political and social instability in Egypt at the time of collection (mid-2013) created a baseline of suspicion and fear, and concurrent news reports of global government surveillance programs in the US and elsewhere made recruitment still more difficult. There were also technical barriers; most phones used in Egypt are non-smartphones and are not capable of archiving messages. Even among smartphone users the process of archiving and uploading data proved to be more difficult than for Chinese and English participants, requiring far more hands-on guidance from our support staff.

	Participants			Recruiter Hours Req'd to Yield a Productive Participant
Language	Enrolled	Productive	Yield	
Egyptian	46	26	57%	38.46
Chinese	118	77	65%	3.90
English	275	152	55%	1.64

Table 1: SMS/Chat Recruitment Yield

Along with modifying the compensation model to be more attractive to Egyptians, LDC approached multiple academic and industrial partners in Egypt to discuss potential recruitment collaboration, emphasizing the extensive privacy protections in place for the collection along with the legitimacy of collection for academic research. Ultimately, with support from our Egyptian academic partners, we managed to recruit 46 participants over a 6-month period, with 26 recruited participants actually contributing data to the study. Altogether, the collection recruited 439 participants, with more than half contributing at least some data. Table 1 summarizes participant recruitment per language.

Due to the personal nature of SMS/chat messages, protection of user privacy was one of our top concerns,

and multiple measures were implemented to address that concern for both live and donated data. During recruitment we emphasized the intended use of the data: to enable language-related research and technology development, with regulation from the University of Pennsylvania's Internal Review Board.

We also emphasized the multiple precautions taken to protect users' privacy:

- Users are never identified by name, phone number, email address, chat ID or other personal identifier in any corpora, publications or presentations.
- Any personal information required for enrollment is stored in a separate secure database and is never shared.
- Message headers that contain phone numbers, chatIDs, email addresses or other personal information are deleted before the message content is added to the corpus.
- Participants can withdraw from the study at any time, and can request removal of a message, conversation or archive at any time

To provide users with an additional measure of control over their data, we added a stage to the data upload process that allowed participants to review and edit their messages before they are added to the corpus. The participant-controlled editing process is illustrated in Figure 2.

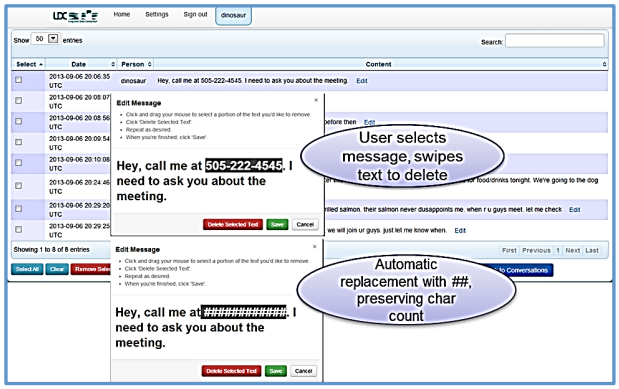

Figure 2: Participant-Controlled Editing

After uploading their archive, the participant is directed to a page that lists all conversations in the archive. Conversations can be sorted (for instance by date) and can be searched for particular content. Participants can select whole conversations for deletion, click on a conversation to view, edit and/or delete individual messages. If a user wants to redacted particular content from a message without deleting the message in its entirety, they simply swipe over the content they wish to remove. The content is replaced by hashtags, preserving the original character count. Several hundred messages were redacted by users, with many more messages and whole conversations being deleted entirely. Because message archives contained two sides of a conversation, users were also instructed to consider the preferences of their conversation partners and to remove any content that might be considered sensitive.

5. Auditing, Selection and Segmentation

After collection and automated processing, all messages were manually audited by project staff who had received special training in the protection of participants' privacy. Manual auditing was necessary for two reasons: to determine which conversations were suitable for downstream translation and annotation tasks, and to exclude any messages or conversations that contained personal identifying information or sensitive content that had not already been redacted by the participant. Auditors used a web-based GUI to screen conversations for acceptability, flag unacceptable content at the level of either individual messages or entire conversations, and split or merge messages to create message units of appropriate size and semantic integrity for translation and downstream annotation.

Auditors first made a global judgment about whether the conversation was acceptable (i.e., primarily in the target language, primarily non-offensive content, etc.). Then the auditor reviewed the entire conversation, flagging any messages that are duplicates, consist entirely of auto-generated content, contain sensitive or offensive content or personal identifying information, or are not in the target language. All messages flagged as containing personal identifying information or sensitive or offensive content are removed from the corpus. As the BOLT program's main goal is to create new techniques for automated translation and linguistic analysis, messages and conversations that are flagged as duplicate, not in the target languages or auto generated are excluded from the corpus as well.

In order to form single, coherent units of an appropriate size for downstream annotation tasks using this data, messages that were split mid-sentence (often mid-word) due to SMS messaging character limits were rejoined, and very long messages (especially common in chat) were split into two or more units, usually no longer than 3-4 sentences. Data releases to BOLT performers include information about the original message form as well as the split or merged version, in order to maintain data integrity.

6. Corpus Characteristics

Both Chinese and English surpassed the collection target of 2 million words/language, with 3.7 million words collected for Chinese and 3.3 million for English.

Chinese and English collection required approximately 15 weeks, though English hit the 2 million-word target in less time. The Egyptian Arabic collection lasted 23 weeks and yielded only 690,000 words. On average, each productive participant contributed more than 30,000 words to the collection. Approximately 33% of all collected messages are removed from the corpus, either during automated validation processes that flag single-sided conversations and duplicates, or during manual auditing. While 7.69 million words were collected, the final corpus contains just over 6.5 million words, comprising 19,139 conversations and 637,000 messages.

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

Table 2: Summary of Corpus Properties

Messages are typically quite short; on average, an Egyptian Arabic message contains only 4 words, a Chinese message contains 7 and an English message 11 words. Across languages, conversations contain 39 messages and 254 words on average. Table 2 presents summary information about the collection for each language.

The collection effort focused primarily on donated data rather than live real-time collection, in keeping with the BOLT program's desire for naturally-occurring content. The corpus reflects this emphasis, with the vast majority of data coming from donations. However, despite the program's preference for SMS, most of the Egyptian and Chinese data was in the chat genre since participants were far more willing to donate chat archives. These results are summarized in Table 3.

	Collection Method		Genre	
Language	*Live*	*Donation*	*SMS*	*Chat*
Egyptian	4%	96%	39%	61%
Chinese	1%	99%	4.40%	95.60%
English	6%	94%	94%	6%

Table 3: Collection Method and Genre

7. Linguistic Features of the Data

Because the BOLT Phase 2 corpus consists of messages exchanged primarily between acquaintances, the language is very conversational and rich in discourse elements and interjections. Laughter and filled pauses are common. Messages are characterized by non-standard use of punctuation and, quite often, punctuation is completely missing. Typographical errors, misspellings and missing spaces are also common. These non-standard usages pose particular challenges to downstream translation and annotation.

Similar to observations made of other English SMS corpora (Tagg, 2009), the English collection contains semi-conventional spellings and contractions such as *'sup* (what's up), *2day* (today); misspellings and missing spaces as in *tonigjt* (tonight) and *wheredoes* (where does); combinations of features as in *lolhopefully* (LOL hopefully); and plentiful use of emoticon and emoji.

Simple and complex initialisms are prevalent in all languages, such as *idk*(I don't know), *hmu* (hit me up), *LD* (ling dao as in 领导 in Chinese), *isa* (insha Allah in Arabic). In Chinese, prevalent pro-drop and missing punctuation combined can make anaphora resolution very difficult, as in this example.

(1) 是保姆休假半个月 兔兔病了 老公出差 公婆来了 感觉比以前更忙累了 工作又不轻松 还好 挺过来了

Gloss: It's that the baby sitter is taking half a month off. Tutu is sick. Husband is on a business trip. In-laws have come (to visit). Feel I'm busier and more tired than before. Work is not easy either. But it's OK. I have survived.

Across all three languages, the nature of the communication means that there is a great deal of implicit context which can make interpretation of messages very difficult, especially when combined with variable surface features (like unconventional spellings and typos). For instance:

(2) Or irp our govener .corbert' whoops).

(3) 收到，就这了，礼拜六

Gloss: I got it. This is the one. This Saturday.

The situation for Egyptian Arabic is further complicated by the combination of a lack of standardized spelling conventions for Egyptian (which is primarily a spoken rather than written language), and the prevalent use of Romanized Arabic, or Arabizi (Yaghan, 2008; Palfreyman and Khalil, 2003) for SMS and chat since most phones do not support Arabic script input. Most Egyptian conversations in our collection contain at least some Arabizi; only 13% of conversations are entirely written in Arabic script, while 63% are entirely Arabizi. The remaining 24% contain a mixture of the two. Switching between the two forms mid-conversation is common, as shown in this example:

(4) A: kano bykhraboki wala eh?
B: ana asfa
B: knt me7tasa awi el ayam ele fatet
B: makntsh 3rfa a3abar 3an masha3ry
A: يعني ايه؟
B: 3shan ba2aly ktir makalemtekish
A: معلش معلش

Gloss:
A: Were they sabotaging you, or what?
B: I am sorry
B: I was very messed up in the last exam
B: I wasn't able to express my feelings
A: Meaning what?
B: Because I haven't spoken to you in a long time
A: Never mind, never mind

The adoption of Arabizi for SMS and online chat may also explain the high frequency of code mixing in the Egyptian Arabic collection. While the corpus eliminated messages that were entirely in a non-target language, many of the acceptable messages contain a mixture of Arabic and English, e.g.

(5) Bas eh ra2yak I have the mask

Gloss: But what do you think? I have the mask.

Neologisms are also prominent in the Egyptian Arabic collection, especially in Arabizi messages, where users conjugate English words using Arabic morphology, as in *el anniversary* (the anniversary). Sometimes English words are spelled in a way that is closer phonetically to the way an Egyptian would pronounce them, for example *lozar* for "loser", or *beace* for "peace". Such features pose great challenges for translation and downstream annotation tasks like word alignment and TreeBank.
Given the particular challenges posed by the prevalence of Arabizi in the data, LDC and researchers at Columbia's Arabic Dialect Modeling Group are collaborating to transliterate the collected Arabizi text into Arabic orthography, and to normalize spelling to the CODA standard (Habash et al, 2012 & 2012) to facilitate morphological analysis and subsequent annotation.

8. Conclusion

To support the BOLT Program's goal of improved multilingual machine translation and information retrieval technologies for informal genres, the Linguistic Data Consortium has produced the world's first publicly-available large-scale multilingual collection of two-sided, naturally-occurring SMS and chat data. The predominance of donated data in the corpus satisfies the goal of "naturalness", providing data rich in the features of informal written conversation. The robust and extensible infrastructure developed for this collection can support live collection and/or donation of a variety of data types and languages in future.
The BOLT Phase 2 SMS/Chat Corpus has already been distributed within the BOLT Program for Machine Translation system training and evaluation. The Chinese and Arabic data has been translated into English, and various stages of annotation are in progress, including word alignment, TreeBank, PropBank and coreference. Portions of the data are slated for use in other sponsored research programs, which will result in additional annotation layers. As these resources are distributed to BOLT researchers, they will also wherever possible be prepared for broader distribution to LDC members and non-member licensees, through our usual mechanisms including publication in the LDC catalog. The BOLT Phase 2 SMS/Chat Collection will be the first such publication, and is expected to appear in the LDC catalog in 2015.

9. 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.

10. References

Brandschain, L., Cieri, C., Graff, D., Neely, A., Walker, K. (2008) Speaker Recognition: Building the Mixer 4 and 5 Corpora. In Proceedings of LREC 2008, Marrakech, Morocco.

Fairon, C. and Paumier, S. (2006). A translated corpus of 30,000 French SMS. In Proceedings of LREC 2006, Genova.

Cieri, C., Corson, L., Graff, D., & Walker, K. (2007). Resources for new research directions in speaker recognition: The mixer 3, 4 and 5 corpora. In Proceedings of Interspeech 2007 (pp. 950–953), Antwerp, Belgium.

Chen, T. & Kan, M.-Y. (2012). Creating a Live, Public Short Message Service Corpus: The NUS SMS Corpus. In Language Resources and Evaluation.

Garland, J., Strassel, S., Ismael, S., Song, Z., Lee, H., (2012). Linguistic Resources for Genre-Independent Language Technologies: User-Generated Content in BOLT. In Proceedings of LREC 2012, Istanbul, Turkey.

Habash, N., Diab, M., and Rambow, O. (2012).Conventional Orthography for Dialectal Arabic: Principles and Guidelines – Egyptian Arabic. Technical Report CCLS-12-02, Columbia University Center for Computational Learning Systems.

Habash, N., Diab, M., and Rambow, O. (2012). Conventional Orthography for Dialectal Arabic. In Proceedings of LREC 2012, Istanbul.

Palfreyman, D. and Al-Khalil, M. (2006). "A Funky Language for Teenzz to Use": Representing Gulf Arabic in Instant Messaging. Journal of Computer-Mediated Communication, 9(1).

Sanders, E., (2012). Collecting and Analyzing Chats and Tweets in SoNaR. In Proceedings of LREC 2012, Istanbul, Turkey.

Tagg, C., (2009). A corpus linguistics study of SMS text messaging. Ph.D. thesis, University of Birmingham

Treurniet, M., De Clercq, O., Oostdijk, N., Heuvel, H. vanden, (2012) Collecting a Corpus of Dutch SMS. In Proceedings of LREC 2012, Istanbul, Turkey.

Wright, J., Griffitt, K., Ellis, J., Strassel, S., Callahan, B., (2012). Annotation Trees: LDC's Customizable, Extensible, Scalable Annotation Infrastructure. In Proceedings of LREC 2012, Istanbul, Turkey.

Yaghan, M.A. (2008) "Arabizi": A Contemporary Style of Arabic Slang. Design Issues, 24(2), pp. 39-52.