Lang Resources & Evaluation (2013) 47:299–335
DOI 10.1007/s10579-012-9197-9

ORIGINAL PAPER

Creating a live, public short message service corpus: the NUS SMS corpus

Tao Chen · Min-Yen Kan

Published online: 24 August 2012

Abstract Short Message Service (SMS) messages are short messages sent from one person to another from their mobile phones. They represent a means of personal communication that is an important communicative artifact in our current digital era. As most existing studies have used private access to SMS corpora, comparative studies using the same raw SMS data have not been possible up to now. We describe our efforts to collect a public SMS corpus to address this problem. We use a battery of methodologies to collect the corpus, paying particular attention to privacy issues to address contributors' concerns. Our live project collects new SMS message submissions, checks their quality, and adds valid messages. We release the resultant corpus as XML and as SQL dumps, along with monthly corpus statistics. We opportunistically collect as much metadata about the messages and their senders as possible, so as to enable different types of analyses. To date, we have collected more than 71,000 messages, focusing on English and Mandarin Chinese.

Keywords SMS corpus · Corpus creation · English · Chinese · Crowdsourcing · Mechanical turk · Zhubajie

1 Introduction

Short Message Service (SMS) is a simple and global form of communication, facilitated by the ubiquitous nature of mobile phones. As mobile handsets

T. Chen (✉)
School of Computing, National University of Singapore, AS6 04-13, Computing 1, 13 Computing Drive, Singapore 117417, Singapore
e-mail: taochen@comp.nus.edu.sg

M.-Y. Kan
School of Computing, National University of Singapore, AS6 05-12, Computing 1, 13 Computing Drive, Singapore 117417, Singapore
e-mail: kanmy@comp.nus.edu.sg

increasingly become cheaper to manufacture, and as the secondhand handset market distributes them more widely downstream, SMS emerges as the most widely-used form of digital communication next to voice telephony. As of 2010, there were 5.3 billion active users of SMS, globally sending 6.1 trillion messages, as estimated by the International Telecommunication Union (ITU).[1] In the same year, Americans alone sent 1.8 trillion messages.[2] Even in China, 93.8 % of phone owners use SMS service, averaging 39.1 received messages per person weekly.[3]

The 160 character limit of SMS was designed by Hildebrandt[4] to accommodate the "communication of most thoughts" presumably in English and other European languages. The limits as well as the early difficulties with text input pushed users to be inventive and create shortenings for communication. The resulting genre of communication has its own name: "texting" (Crystal 2008), and has been debated as to whether it is a friend or foe of language creation, use and understanding. With the advent of the smartphone, other informal communications (e.g., tweets and instant messages), which were previously computer mediated, are now enabled on mobile phones as well. The SMS length restriction has been circumvented by standards that allow multiple messages to be concatenated together. They share some characteristics with SMS, such as the frequent use of emoticons, unconventional short forms and misspelled words. Tweets, for instance, have been dubbed as "SMS of the Internet."[5]

1.1 Difficulties of SMS collection

SMS messages have been studied since the late 1990s, when SMS communication became widely available among cellular carriers. The ubiquitous, personal and informal nature of SMS attracted researchers' attention, resulting in sociographic, linguistic and usability research topics—e.g., the impact of SMS on social culture, linguistic analysis, text-entry improvement, named entity recognition, normalization, authorship identification and spam message detection.

With such a large societal impact, however, SMS seems underwhelmingly studied, especially when compared with tweets. Munro and Manning (2012) pointed out there is a bias in recent research of SMS, Twitter and email, that is, Twitter makes up almost 75 % of research and SMS only accounts for 14.29 %, [6] while the actual global usage are 0.16 and 37.83 %, respectively. It is clear why—gathering a large corpus to study is difficult and fraught with confidentiality problems, as messages are often personal or contain confidential information. Unsurprisingly, there are thus few publicly available SMS corpora, which discourages comparative studies. The unavailability of corpora also becomes an onus on the aspiring researcher, forcing them to gather messages themselves.

[1] http://www.itu.int/ITU-D/ict/material/FactsFigures2010.pdf.

[2] http://www.ctia.org/media/press/body.cfm/prid/2021.

[3] http://12321.cn/pdf/sms1102.pdf.

[4] http://latimesblogs.latimes.com/technology/2009/05/invented-text-messaging.html.

[5] http://www.wisitech.com/blog/?p=57.

[6] In terms of publications in IEEE, ACM and ACL between June 2010 and June 2011.

We now discuss three issues related to SMS corpus collection:

- Why are there so few public SMS corpora?
- What factors make collecting SMS so difficult?
- Can other corpora of communication vehicles, such as tweets, replace SMS in studies?

SMS messages are primarily stored in mobile phone carriers' database; only a small portion of them are stored in users' phones given limited phone storage. For legal and privacy reasons, carriers cannot release their SMS databases for research, as users' messages are often personal, privileged and private. Even when SMS corpora are collected by researchers from phone owners, such researchers have to maintain exclusive access due to the same privacy concerns. Thus, the private nature of SMS results in the scarcity of public SMS corpora.

In this article, *SMS collection* in particular refers to gathering messages directly from phone owners, largely for research purposes. There are two major explanations for the difficulty of SMS collection. As researchers want the text and the metadata for study, privacy is a major concern, as the majority of phone owners are uncomfortable with having their private messages revealed, even when the purpose is for research and academic pursuits. Additionally, in collecting SMS directly from users, the target messages to be collected are stored on the users' mobile phones, which means that the collection of a large corpus requires the cooperation of many phone owners, and the support of software to ease the collection over many potential mobile phone platforms.

In recent times, Twitter and social networks such as Facebook have made the genre of the short message even more ubiquitous. Tweets and status updates are closely related to SMS, sharing the characteristic terse style. However compared with SMS, tweets are remarkably easy to gather since Twitter releases API for accessing data. So a natural question arises: can tweets replace SMS for related studies? Perhaps for some purposes, the answer is "yes", but for other purposes they clearly cannot. They have fundamental differences which affirm SMS as an important medium of study. First, SMS is a private communication between two parties, which may contain very sensitive topics or information (e.g., bank account and email address), hence its difficulty for collection. In contrast, tweets and a large portion of social network messages and comments are decidedly broadcast media, and hence far less private and sensitive. Second, though SMS and tweets have similar characters restriction, 160 and 140 respectively, they still differ in length. Ayman reported that the bulk of tweets are around 40 characters long in a corpus of 1.5 million tweets,[7] while Tagg (2009) mentioned that average length of SMS is 17.2 words in her corpus of 11,067 messages. Bach and Gunnarsson (2010) validated this, pointing out that SMS messages were more likely to be very short—containing only one word—compared with tweets, due to the more personal and conversational aspects of SMS. Moreover, tweets tend to be more standard and formal than SMS, using more standard punctuations and less number of logograms and pictograms, as observed by Denby (2010). In this sense, if the understanding of

[7] http://www.ayman-naaman.net/2010/04/21/how-many-characters-do-you-tweet.

personal informal communication and how it is evolving is important, then SMS deserves to be collected and studied for the public good.

1.2 Our contributions

A public SMS corpus is needed to fill this gap for research material which will benefit all the researchers who are interested in SMS studies. In 2004, our former project established an SMS collection project for this aim, gathering and publishing a corpus of 10,117 English SMS, mainly from students in our university (How and Kan 2005). The corpus was released freely online. It was the largest publicly available English SMS corpus until 2010, and was used in a number of SMS studies. However, mobile technology has developed significantly in the past 8 years. The advent of the smartphone, which features newer text input technologies, has influenced people's texting habits. SMS use has penetrated other countries and languages of interest in this period. A language of interest to us in our region is Mandarin Chinese, of which there is an extreme scarcity of public domain SMS corpora.

Considering this, we resurrected the SMS collection project in October 2010, reviving it as a live corpus collection project for both English and Chinese SMS. The "live" aspect of our project emphasizes our aim to continually enlarge and publish the corpus for end users. Our SMS collection makes use of an array of collection methodologies, leveraging current technology trends, with the aim of making the resultant corpus more representative of SMS worldwide (rather than a specific community[8]), to enable more general studies. Our current corpus also features improved collection methodology, making the collection process more accurate with fewer transcription errors, and is continuously released publicly with an effective anonymization process.

As of June 2012, we have collected 41,790 English messages and 30,020 Chinese messages, resulting in the largest public SMS corpora (to our knowledge), in terms of both English and Mandarin Chinese languages, independently.

Our article reports on the contributions of our corpus and its collection methods. In particular, we:

- use a battery of methodologies to collect the corpus, paying particular attention to privacy issues to address contributors' concerns (Sect. 3);
- create a website to document and disseminate our gradual achievement, enabling direct, online browsing of the messages. We also release the resultant corpus on a regular monthly schedule as XML and as SQL dumps, along with salient corpus statistics (Sect. 4); and
- exploit a good Chinese crowdsourcing website for language data collection and compare it with its more well-known, US counterpart (Sect. 5)

[8] In contrast, our 2004 corpus was collected locally within the University in Singapore, not representative of general worldwide SMS use.

2 Related work

2.1 Comparison of SMS corpora

While the scope of related work to texting in general is vast, for the purposes of this article, we limit the scope of our review to scholarly publications concerning SMS studies. This makes the task feasible and allows us to further break the review down into particular facets of interest. In particular, we pay attention to the size and language of each collection, the characteristics of its contributors, how the respective authors have collected their corpus, and the respective corpus' availability. An important goal of this chapter is to provide a comprehensive inventory of SMS corpora that have been collected and pointers to their associated studies.

A survey of the literature shows that publicly-available SMS corpora are scarce. The resulting scarcity motivates subsequent researchers to collect their own SMS corpora for their specific projects; but as these projects are often one-off, the resulting collections are also not publicly available, creating a vicious cycle. For ease of reference, we have also compiled and listed the existing corpora in Table 1 (publicly available corpora are indicated by an asterisk). In addition to the availability, size and language of the corpora, we are interested in the underlying facets of their contributors' identity and the collection methods used.

- The *Size* of existing SMS corpora is tiny when compared with corpora that subsample Twitter. For instance, Ritter et al. (2011) built an automatic post response generation system based on 1.5 million tweets, and Wang et al. (2012a) studied the linguistic characteristics of retweets and created a retweetability predictor with a corpus of 9.5 million tweets. In contrast, the largest SMS corpus consists of a mere 85,870 messages (Liu and Wang 2010). 50 % of the corpora contain less than 1,000 messages, and only five corpora comprise of more than 10,000 messages. We attribute the small scale of these corpora to the aforementioned difficulty of collecting SMS. However when the corpus is small, the resultant findings of the studies are often not statistically significant (Dürscheid and Stark 2011).
- The *Language* of the corpora ranges from European languages (English, French, Germany, Italian, Polish, Swedish, etc), Asian languages (Mandarin Chinese), to African ones (Kiswahili and Xhosa). However, European languages dominate, with only two Chinese corpora and two African corpora being the exceptions. A corpus can be classified as monolingual or multilingual, describing how many languages are exhibited in its component messages. Most corpora are monolingual, and to our knowledge, only five existing corpora are multilingual (Deumert and Oscar Masinyana 2008; Elvis 2009; Bach and Gunnarsson 2010; Barasa 2010; Bodomo 2010; Dürscheid and Stark 2011).
- *Contributors* to the SMS corpora can be categorized as either known or anonymous. Families and friends are the most common known contributors (Segerstad 2002; Žic Fuchs and Tudman Vukovic 2008; Gibbon and Kul 2008; Tagg 2009; Barasa 2010). Others include colleagues (Ju and Paek 2009; Bach

Table 1 Existing SMS corpora

Researcher(s)	Size	Language(s)	Contributors	Collection method(s)
Pietrini (2001)	500	Italian	15–35 years old	*Not mentioned*
Schlobinski et al. (2001)*	1,500	Germany	Students	*Not mentioned*
Shortis (2001)*[a]	202	English	1 male student, his peers and family	Transcription
Segerstad (2002)	1,152	Swedish	4 paid and 16 volunteers	Transcription, Forwarding
Kasesniemi and Rautiainen (2002)	7,800	Finnish	Adolescents (13–18 years old)	Transcription
Grinter and Eldridge (2003)	477	English	10 teenagers (15–16 years old)	Transcription
Thurlow and Brown (2003)	544	English	135 freshmen	Transcription
Ogle (2005)	97	English	Nightclubs	Subscribe SMS promotion of nightclubs
Ling (2005)	867	Norwegian	Randomly select 23 % of 2003 respondents	Transcription
How and Kan (2005)*	10,117	English	166 university students	Transcription
Fairon and Paumier (2006)*	30,000	French	3,200 contributors	Forwarding
Choudhury et al. (2007)*	1,000	English	Anonymous users in treasuremytext[b]	Search the SMS from the website
Rettie (2007)	278	English	32 contributors	*Not mentioned*
Ling and Baron (2007)	191	English	25 undergraduates	Transcription
Žic Fuchs and Tudman Vukovic (2008)	6,000	Croation	University students, family and friends	*Not mentioned*
Gibbon and Kul (2008)	292	Polish	University students and friends	*Not mentioned*
Deumert and Oscar Masinyana (2008)	312	English, isiXhosa	22 young adults	Transcription, Forwarding
Hutchby and Tanna (2008)	1250	English	30 young professionals (20–35 years old)	Transcription
Walkowska (2009)	1700	Polish	200 contributors	Forwarding, Software
Herring and Zelenkauskaite (2009)	1452	Italian	Audiences of an iTV program	Online SMS archives
Tagg (2009)	10,628	English	16 family and friends	Transcription
Elvis (2009)	600	English, French, etc.	72 university students and lecturers	Forwarding

Table 1 continued

Researcher(s)	Size	Language(s)	Contributors	Collection method(s)
Barasa (2010)	2,730	English, Kiswahili, etc.	84 university students and 37 young professionals	Forwarding
Bach and Gunnarsson (2010)	3,152	Swedish, English, etc.	11 contributors	Software
Bodomo (2010)*	853	English, Chinese	87 youngsters[c]	Transcription
Liu and Wang (2010)	85,870	Chinese	Real volunteers	*Not mentioned*
Sotillo (2010)	6,629	English	59 participants	Software
Dürscheid and Stark (2011)*	23,987	Germany, French, etc	2,627 volunteers	Forwarding
Lexander (2011)	496	French	15 young people	Transcription
Elizondo (2011)	357	English	12 volunteers	Transcription

An asterisk (*) indicates that the corpus is publicly available

[a] The corpus was largely assembled by Jon Stevenson, one of author's students

[b] http://www.treasuremytext.com

[c] The contributors and collection method are for the 487 messages collected in 2002; later, another 366 messages were collected from 2004 to 2006 without mentioning the contributors and collection methods

and Gunnarsson 2010), students in a specific university (Thurlow and Brown 2003; How and Kan 2005; Gibbon and Kul 2008), and recruited teenagers (Kasesniemi and Rautiainen 2002; Grinter and Eldridge 2003).

Anonymous contributors are those that the researchers do not personally know or need to make direct contact with. The methods to carry out the collection are also more varied than in known contributor cases. An example is the corpus collected by Ling (2005), which involved 2,003 anonymous respondents via a telephone survey. Another example is the corpus collected by Herring and Zelenkauskaite (2009), whose participants were anonymous audience members of an Italian interactive television program. Perhaps the most important instance is the distributed effort by the sms4science project, an international collaboration aiming to build an SMS corpus alongside corpus-based research on the resulting corpus. The sms4science subprojects have been carried out in nine countries, such as Belgium (Fairon and Paumier 2006), Switzerland (Dürscheid and Stark 2011), France,[9] Greece, Spain and Italy. Anonymous contributors were recruited from across the subproject's country or region.

The difference between the known versus anonymous contributor corpora affects the corpora' representativeness. Known contributors usually share similar demographic features, such as similar age (teenagers or college students), same career (colleagues), and/or same geographic location (living in a same city). Hence, the corpora from known contributors may not be representative of the general landscape of SMS usage. This characteristic may be perfectly acceptable or desired for the particular project for which the corpus is collected, since the project may be restricted in a particular purpose or study. For instance, Grinter and Eldridge (2003) collected 477 messages from 10 teenagers for studying how messages have been incorporated into British teenagers' lives. Corpora from anonymous contributor projects, such as the sms4science project, are more broad and aim to satisfy general studies. We note that both aims can be achieved in an anonymous collection process, as when suitable demographics are taken per message and the corpus is sufficiently large, an identifiable subset of the corpus can still serve for specialized studies.

- *Collection methods*. For our purpose, the most interesting facet of the studies is how they each collect their corpus. We observed three primary methods to collect SMS. The simplest approach is to simply transcribe messages from the mobile phone, by typing them into a web based submission form (Segerstad 2002; How and Kan 2005), into a word processing or other electronic document (Segerstad 2002; Deumert and Oscar Masinyana 2008; Tagg 2009; Elizondo 2011), or even the simple method of writing them down on paper (Kasesniemi and Rautiainen 2002; Grinter and Eldridge 2003; Thurlow and Brown 2003; Deumert and Oscar Masinyana 2008; Bodomo 2010). Transcription can also happen later to facilitate collection speed—Lexander (2011) took photos of messages stored in participant's phone, for later transcription by researchers. A second method is to export or upload SMS messages via software. The corpus

[9] http://www.alpes4science.org.

collected by Jonsson et al. (2010) is such an example. They implemented a phone widget for providing location-based SMS trend service and collecting SMS messages as well, since messages will be uploaded to a server when using the service. Sotillo (2010) and Walkowska (2009) gathered messages by collecting SMS exported from contributors' mobile phone using software suites such as Treo Desktop (supporting PalmOS) and Microsoft's My Phone. The third class of methods is to have contributors forward messages to a collection number. Messages usually are forwarded to researcher's own mobile phone (Segerstad 2002; Walkowska 2009; Barasa 2010), which may incur cost for the contributors. They are typically compensated for their cost, thus the large-scale collection can be costly. The studies done by (Fairon and Paumier 2006; Dürscheid and Stark 2011), were in collaboration with mobile phone operators, such that contributors could forward their messages to the operator-central number for free, further lowering the barrier for potential contributors.

Aside from these common methods, we observed other one-off methods used in particular studies. Ogle (2005) collected broadcasted SMS by subscribing to the SMS promotion of several nightclubs; Ling (2005) asked respondents to read their SMS aloud during a telephone survey; Herring and Zelenkauskaite (2009) downloaded the audience's SMS from an SMS archive of an interactive TV program; and finally, Choudhury et al. (2007) collected SMS from an online SMS backup website.

Which method is best? Each method has its own merits and drawbacks. If the scale needed is small, transcribing a few messages is the easiest, needing virtually no effort in preparation and cost. However for medium to large corpus projects, this methodology is not scalable, being time-consuming. Also, it is prone to both transcription and deliberate correction errors, despite any instructions to transcribe messages exactly as displayed.

Exporting via software support preserves the originality of messages, and in certain cases, gathers valuable metadata (sender and receiver's telephone number and sending timestamp). Exporting also enables batch submission, easily enabling a stream of valid and accurate messages to be collected. It also encourages continual SMS contribution, especially when the export-and-collection process can be automated. The key drawback with this method is that it is tied to the phone model, and thus creates a selection bias in the possible contributors. Exacerbating this problem is that some phone models do not even have (free) software to export SMSes.

Forwarding messages is also effective to maintain the original message, but may be costly if the sending cost needs to be recouped by the researcher. Forwarding may be easy for newly-sent messages as the collection number can be added as a second recipient. However, many phone models allow only forwarding single messages, such that forwarding lots of individual messages may be tedious. This discourages the collection of many messages from a single contributor.

- *Availability*. In terms of availability, the existing corpora range from private, partially public, to completely public. As displayed in Table 1, most existing SMS corpora are private access. Without a doubt privacy and non-disclosure issues are the underlying reasons. On one hand, it is the responsibility of

researchers to protect the contributors' privacy and the easiest way to achieve the aim is by not making the corpus public (Dürscheid and Stark 2011). On the other hand, researchers may not be able to get the consent from contributors to release the corpus or be restricted by the rules of their institution's Institutional Review Board (IRB) (Sotillo 2010).

We define partially public corpora as corpora that are not immediately freely accessible, and require initiative on the part of the scholar to secure. The two partially public corpora are the Belgium Corpus (Fairon and Paumier 2006) and Swiss Corpus (Dürscheid and Stark 2011), collected by two sub-projects of sms4science. The former was distributed as a Microsoft Access database on a CD-ROM and is purchasable but restricted to bona fide research purposes only. The latter corpus is browsable online for researchers and students, but not downloadable as an entire corpus. Online browsing is very convenient for reading a few SMS without the need to parse the raw data, but makes it difficult to obtain all the SMS for serious corpus study. We feel that this limits the potential studies that could employ the corpus.

Completely public corpora are freely, immediately and wholly accessible. Shortis (2001) published 202 English SMS as a webpage[10] Other public corpora were released as a file for freely downloading but vary in the file format. Both the German Corpus (Schlobinski et al. 2001)[11] and HKU Corpus (Bodomo 2010)[12] were released in Portable Document Format (.pdf), while the IIT Corpus (Choudhury et al. 2007)[13] and our aforementioned 2004 NUS SMS Corpus[14] were released as text and XML files, respectively. The HKU Corpus is the only corpus containing about 140 Chinese language messages, but these mix English words, as may often be the case in Hong Kong. Strictly speaking, there is no pure, Mandarin Chinese SMS corpus in the public domain.

Another public corpus is 9/11 pager messages released by Wikileaks in 2009[15] with over half million messages. The intercepts cover a 24 h surrounding the September 11th, 2001 attacks, ranging from exchanges among office departments, to fault reporting of computers as the World Trade Center collapsed. A few research studies have been conducted on this 9/11 corpus. Back et al. (2010) investigated the emotional timeline of the messages and analyzed the negative reaction to 9/11 terrorist attacks. Back et al. (2011) also used automatic algorithms and human judgment to identify 37,606 social messages from the original corpus, and rated the degree of anger in the timeline. Although 37,606 messages is considered quite a large corpus, most of the intercepts are limited to people's reaction to that terrorism event. Such topically-focused pager messages

[10] Available at http://www.demo.inty.net/app6.html. Although the corpus is not directly downloadable as a file, we still consider it as public as all of the messages are displayed on the single web page.

[11] http://www.mediensprache.net/archiv/corpora/sms_os_h.pdf.

[12] http://www0.hku.hk/linguist/research/bodomo/MPC/SMS_glossed.pdf.

[13] http://www.cel.iitkgp.ernet.in/~monojit/sms.html.

[14] http://www.comp.nus.edu.sg/~rpnlpir/downloads/corpora/smsCorpus.

[15] http://mirror.wikileaks.info/wiki/911.

cannot replace a general collection SMS messages, thus the corpus is not suitable for most SMS related studies.

- *Release time*. For the seven public SMS corpora mentioned above, all of them were released after the completion of data collection. The static release favors protection of contributors' privacy, since a global and thorough anonymization could be conducted (Fairon and Paumier 2006; Dürscheid and Stark 2011). A live release, in which the corpus is continually published and updated during the collection process, faces greater challenge and risk in anonymization.
  Due to the individual and private nature of SMS, the resulting collected corpora also contain private details, which may make it easy to discover the identity of the sender or recipient. For these reasons, such resulting corpora also cannot be made public. As mentioned, this creates a vicious cycle, erecting a barrier to SMS research, making SMS seem less significant than it is for understanding our current era of communication. It is clear that a publicly available, large-scale corpus of SMS could lower this barrier, and make the study of SMS more widely available to scholars of all disciplines.

2.2 Crowdsourcing SMS collection

From the above summary, we can see that a variety of approaches have been employed to collect SMS for study. In collecting any large-scale corpora, it is necessary to distribute the task among a large group. This is aptly illustrated by the sms4science project which involves thousands of contributors. As we aim to create an authoritative SMS corpus to enable comparative studies, it is vital that the corpus also be large. Thus our methodology should also follow this distributive paradigm.

Crowdsourcing, the strategy of distributing a task to a large "crowd" of human workers via some computer-mediated sources, has emerged as a new driver of computation. In tasks where raw compute power cannot succeed, but where an aggregate of human judgments or efforts can, crowdsourcing can be used. It uses the computing medium to connect many workers to a task necessitating human processing. For our particular instance of SMS collection, we can employ crowdsourcing to connect many potential contributors of SMS to a collection framework.

The term "crowdsourcing" actually subsumes several forms (Quinn and Bederson 2009), of which the Mechanized Labor form is most relevant to our project. This form is defined by its use of a (nominal) monetary to motivate distributed workers to do their task. The most notable example of mechanized labor in practice is embodied by Amazon's Mechanical Turk (hereafter, MTurk),[16] an online marketplace for employers (a.k.a. requesters) to publish small tasks, and workers (a.k.a. Turkers) to choose and complete the tasks. MTurk has become a popular crowdsourcing platform for its low cost and diverse workforce.

Of late, MTurk has been employed for many uses within scholarly work. We only focus on works concerning data collection, in particular, the collection of language-

[16] https://www.mturk.com/mturk/welcome.

related data. In 2010, a special workshop was held with the North American Annual meeting of the Association of Computational Linguistics (NAACL), entitled "Creating Speech and Language Data With Amazon's Mechanical Turk". Callison-Burch and Dredze (2010) categorized the data collected in this workshop into six types: Traditional NLP tasks, Speech and Vision, Sentiment, Polarity and Bias, Information Retrieval, Information Extraction, Machine Translation. While SMS collection is not subsumed by any of the six types, the success and variety of the corpora created in the workshop as well as in other studies convince us that MTurk is a suitable platform for our project.

3 Methodology

Our project focuses on collecting both English and Chinese SMS messages, to expand our 2004 English SMS corpus and address the need for a public Chinese SMS corpus. Our aim is to create an SMS corpus that is: (1) as representative as possible for general studies, (2) largely free of transcription errors, (3) accessible to the general public without cost, and (4) useful to serve as a reference dataset. Several strategies were used in our collection process to achieve these goals.

First, we did not restrict our collection to any specific topics, to encourage diversity among the messages. Also we did not limit to known contributors, but instead tried to diversify contributor backgrounds to fulfill the first aim of making contributor demographics similar to the general texting population. We used three different technical methods to collect SMS: (1) simple transcription of an SMS into a collection web site, (2) exporting of SMS directly from phone to a file for submission, and (3) uploading lists of SMS as an email draft, for editing and eventual submission via email initiated by the contributor. The latter two collection strategies also favor the collection of whole SMS streams during an interval, favoring an unbiased collection of messages. They also collect the messages as-is from the phone's memory, minimizing the chance of transcription or entry errors, satisfying the second aim. To achieve the third aim, we created a program (discussed below) to automatically replace any identifiers and sensitive data with placeholders and to encrypt identifiable metadata with each SMS. With these minor modifications, contributor's privacy issues are mollified and allow us to release the corpus to the general public. Finally, to ensure that the corpus satisfies our fourth aim of being a viable reference corpus, we release archived, static versions of our continually-growing corpus on a monthly basis. In the following, we present these strategies in more detail.

3.1 SMS collection requirements

We did not restrict contributors to send only SMS on certain topics. This helps to keep the collected messages representative of actual content (Barasa 2010), and diversify the corpus in content. Moreover, we required contributors to fill out a demographic survey about their background (e.g., age, gender, city, country),

texting habits (input method, number of SMS sent daily, years of using SMS) and information about their phones (brand, smartphone or not). Such answers form a profile associated with the bulk of the SMSes collected in our corpus, which we feel can facilitate sociolinguistics studies.

We did, however, require that the submitted messages be personal, sent messages only. The "sent-message" restriction is required for two important reasons. Ethically speaking, the submission of received messages is disallowed as the consent of the sender is not guaranteed, and may violate the trust and rights of the original sender. As we also aim to have as complete demographics on the SMSes collected, we would also be unable to contact the senders to have them complete the same demographic survey, which makes received messages less appealing to collect. The "personal" restriction means the messages are typed by the contributors themselves and not of artificial or commercial nature; i.e., chain messages to be forwarded (e.g., blessings, jokes, quotes) that may be available on the Internet.

3.2 Source of contributors

As we aim to create a corpus which reflects the general characteristics of SMS messages, we want contributors to have diverse backgrounds, of a wide range of ages, and living in various geographic locations. As crowdsourcing methods pull from a variety of sources, we deemed this strategy as the most suitable for SMS collection.

Probably the most well known crowdsourcing platform is Amazon's Mechanical Turk (henceforth, MTurk), which allows users to publish tasks and for the members of the general public to do the tasks, usually for a nominal fee. The use of MTurk as a crowdsourcing technique has been widely documented in the computer science literature. It is also an ideal place for conducting our English SMS collection. We also published a few tasks in another mechanized labor site, ShortTask,[17] to diversify the background of contributors. Considering its similarity to MTurk and the limited usage in our current collection methods, we do not discuss it further in this paper.

A demographic survey of MTurk workers (known colloquially as Turkers) conducted by Ipeirotis (2010a) reveals that the respondents are from 66 countries with a wide distribution in age and education levels, but that the majority of them are from English-speaking countries (46.8 % American and 34.0 % Indian). However, the study also suggests the scarcity of Chinese workers, which has been validated by other researchers (Gao and Vogel 2010; Resnik et al. 2010) who have pointed out that there are few Chinese-speaking Turkers and thus difficult to recruit. We also performed a pilot study in MTurk, publishing two batches of tasks to collect Chinese SMS messages, but received few submissions, validating the earlier reports of shortage of Chinese workers. So while MTurk is a good platform for collecting English SMS, we have to find a more suitable platform for gathering Chinese SMS.

[17] http://www.shorttask.com.

In China, the same crowdsourcing form of mechanized labor goes by the name of "witkey" (威客, *Wēi Kè* in pinyin) , short for "key of wisdom", described as using the wisdom of the masses to solve problems. Among such Chinese websites, Zhubajie (猪八戒)[18] currently stands out for its dominant market share (over 50 %) and huge workforce (over 50 million).[19] Zhabajie also categorizes tasks within its own ontology, and one specific category relates to SMS (more details in Sect. 5) Therefore, we chose Zhubajie as the crowdsourcing platform for collecting Chinese SMS.

Besides anonymous workers in MTurk, ShortTask and Zhubajie, we also leveraged the local pool of potential contributors in Singapore. English and Chinese are two of the official languages of Singapore, making it an ideal place to recruit contributors for our corpus collection. We recruited contributors by emailing students in our department. They were requested to submit either English or Chinese SMS. Participants from all four above sources were reimbursed a small sum of money for their contributions.

Finally, we also wanted to explore whether people would be willing to contribute SMSes purely for the sake of science (without remuneration). To test this, we sent email invitations to Internet communities of linguists and mobile phone owners. These communities comprised of the well-known *corpora-list*[20] (an international mailing list on text corpora for research and commercial study), *corpus4u*[21] (a Chinese corpora forum), *52nlp*[22] (a Chinese collaborative blog in natural language processing), and two Chinese popular mobile phone forums – *hiapk*[23] and *gfan*.[24]

In summary, we tried to diversify our contributor pool to ensure we collected from a wide variety of sources. This makes the overall corpus more diverse, as each one of the contributor sources has a certain bias. However, even with our diverse contributor pool, we cannot claim to have a fully representative sample of SMS, as our collection methods target a rather wealthy subset of SMS users, who have access to computers (only such potential contributors would be reachable through crowdsourcing methods, forums, etc.). It is important to recognize this limitation of our work, but we feel that this does not limit the applicability of our corpus for most study purposes.

3.3 Technical methods

Our collection methods can be categorized into three separate genres. We want our methods to be simple and convenient for the potential contributors and allow us to collect SMS accurately without transcription errors.

[18] http://www.zhubajie.com.

[19] According to the China Witkey Industrial White Paper 2011 conducted by iResearch.

[20] Corpora@uib.no.

[21] http://www.corpus4u.org.

[22] http://www.52nlp.cn.

[23] http://bbs.hiapk.com.

[24] http://bbs.gfan.

- *Web-based transcription*. The simplest collection method is transcribing messages from phone. We designed a web page for contributors to input their messages. Contributors were asked to preserve the original spelling, spaces and omissions of SMS, and standardized emoticons by following a transcription code table of our design (e.g., any simple smiling emoticon should be rendered as ":)").

  As it is simple to implement, we adopted this transcription method as the collection method in our pilot, when we restarted our collection in 2010. We published a series of small tasks in MTurk to collect Chinese and English SMS, to test the waters and refine problems with our instruction set. However, when reviewing the submitted messages before Turker payment, we found a serious problem: a high rate of cheating. Some of the English Turkers had just typed messages such as blessings, jokes and quotes that were verbatim copies of ones that were publicly available in some SMS websites. These messages would not be representative of personally-sent messages, as required in the task's documentation. For Chinese SMS, it was apparent that some English speakers pretended to qualify as Chinese speakers, copying Chinese sentences from Internet websites or completing the task without actually submitting any content. For both languages, we spent a non-trivial amount of manpower (about 3.5 hours for 70 English submissions, and half an hour for 29 Chinese submissions) to inspect such messages, to validate the submissions as original, checking against identical or very similar messages publicly available in the web. For our pilot tasks, our rejection rate of English and Chinese messages was 42.9 and 31.0 %, respectively—clearly suboptimal considering the time and effort needed to review submitted SMSes and the potential ill-will generated with Turkers who performed our tasks but whom we deemed as cheating.

  A final problem is that transcription is prone to typos and deliberate corrections, and discourages contributors from inputting a lot of messages, since the re-typing is tedious. From these pilot tasks, we learned that we needed better collection methods that ensured message accuracy and demanded less validation.
- *SMS export*. With supporting software, some mobile devices can export received or sent SMS as files in various formats such as TXT, CSV, XML, HTML. This capability allows us to define this second collection method, SMS Exporting. It involves two steps. First, contributors export SMS from their phone as a readable archive, and optionally, censor and delete SMS that they do not want to contribute. Second, contributors upload the archive and answer a web-based demographic survey.

  We recruited contributors via both crowdsourcing websites and by regular, email invitations. Our unified description (for both emailed invitations and the crowdsourcing tasks) asked contributors to participate if they can export messages as a readable file (e.g., CSV, XLS). While such exporting capabilities are becoming more prevalent, not all phone models have such software. Even when available, it is not always free nor easy to use. Noting these difficulties, we thus prepared notes to ease contributors' potential difficulties for popular platforms.

Demographics from our web-based transcription task fielded in MTurk shows that 60 % of English SMS workers and 47 % of Chinese SMS workers were Nokia owners. This phenomenon is in accord with Nokia's large market share and penetration in China and India (74 % of English SMS workers in the pilot task are from India). Fortunately, Nokia provides the Nokia PC Suite,[25] a free software package for SMS export and backup via computer, which works on most Nokia models and meets our requirements. In the task description, we therefore linked to the download site for Nokia PC Suite and we offered a webpage-based tutorial on how to export SMS using the software.

Besides the advantage of high accuracy, and ease of batch submissions of SMS as mentioned in Sect. 2, SMS Export greatly helps us in validation. Since the archive has a specified format—which includes the telephone numbers of the sender and receiver, the send timestamp of the message—it significantly lowers the barrier for submitting valid data and significantly raises the barrier for submitting false data. For this reason, we expend significantly less effort in validating SMS submitted by this process.

- *SMS upload*. With the growing popularity of smartphones, which have added functionality, we felt it would be a good idea to implement mobile applications ("apps") that can contribute SMS directly. At the current juncture, we have implemented an app for the Android platform. Inspired by another app, SMS Backup,[26] which is an open-source Android app for backing up SMS to Gmail,[27] we adapted the code to create a new app, which we called *SMS Collection for Corpus*, as a pilot software for smartphones. We have released it as a free application in Google Play.[28] Figure 1 shows a snapshot of our app.

  SMS Collection for Corpus works by uploading sent SMSes from the Android device to the user's Gmail[29] as a draft email. To allow the user to censor and delete message[30] that she does not deem suitable for the corpus, the app purposely does not send the messages directly to our collection web site. We do not receive the SMSes for the corpus until contributors act to send out the draft email to us. The app also automatically anonymizes the metadata (telephone numbers) and replaces sensitive identifiers with placeholders. The detail of the anonymization process is described later in this section.

  As with the SMS Export collection method, this method also reduces the possibility of cheating while preserving the originality of messages. One advantage over SMS Export is its convenience; there is no need to connect to a separate computer to perform the submission. Most importantly, the automatic anonymization may assure potential contributors that their privacy is protected. For SMS collected in the other two methods, we employ the same anonymization process, but after receiving the original SMS; in this method the

[25] Now replaced by Nokia Suite http://www.comms.ovi.com/m/p/ovi/suite/English.

[26] http://code.google.com/p/android-sms.

[27] http://mail.google.com.

[28] https://play.google.com/store/search?q=pname:edu.nus.sms.collection.

[29] Hence a Gmail account is a prerequisite to this collection method.

[30] But we do not encourage users to edit messages since we feel it may destroy the originality.

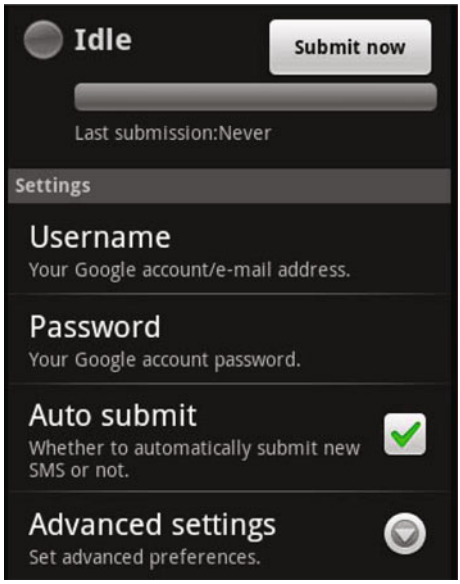

Fig. 1 A screen capture of the *SMS Collection for Corpus* mobile application

anonymization procedure runs on the client smartphone, even before it reaches our collection server.

To the best of our knowledge, *SMS Collection for Corpus* is the first application designed to collect a corpus from original sources. It is also easy to adapt the software to support nternationalization, so that the user interface can support new languages for potential submitters in other languages. In July 2011, we did exactly this, extending the user interface to support prompts in Dutch, at the request of the investigators of the SoNaR34[31] project for SMS collection in the Netherlands.

3.4 Anonymization

Since SMS often contains both personal and confidential information, such as telephone numbers and email addresses, we need to anonymize these data when included in submitted SMS.

For messages collected by SMS Export and SMS Upload, the SMS metadata—telephone number of sender and receiver, and send time—is also collected. However, for the sender and recipient SMS metadata, we need to replace the original data with a unique identifier for each phone number so that privacy can be maintained, while preserving the fact that multiple messages linked to the original, same source are attributed correctly in the anonymized versions. To solve this problem, we adopt DES encryption to create a one-way enciphering of the phone numbers, which replace the originals in the corpus.

[31] http://www.sonarproject.nl.

Table 2 Replacement codes used by our anonymization process

Original content	Example	Replaced code
Email address	name@gmail.com	⟨*EMAIL*⟩
URL	http://www.google.com	⟨*URL*⟩
IP address	127.0.0.1	⟨*IP*⟩
Time	12:30	⟨*TIME*⟩
Date	19/01/2011	⟨*DATE*⟩
Decimal	21.3	⟨*DECIMAL*⟩
Integer over 1 digit long	4,000	⟨#⟩
Hyphen-delimited number	12-4234-212	⟨#⟩
Alphanumeric number	U2003322X	*U*⟨#⟩X

For the SMS message body, sensitive data include dates, times, decimal amounts, and numbers with more than one digit (telephone numbers, bank accounts, street numbers, etc.), email addresses, URLs, and IP addresses. Ensuring the privacy for these types of data is paramount, and as such, we adopt a stricter standard in dealing with sensitive data in the message itself. Such information is programmatically captured using regular expressions and replaced by the corresponding semantic placeholders, as shown in Table 2. For example, any detected email address will be replaced by the code ⟨*EMAIL*⟩. This process gives a level of protection against publishing sensitive data. While we remove such confidential information that fits our set of regular expression patterns, in general it is impossible to remove all sensitive data with only a simple set of textual regular expressions. In particular, as person names are varied, language-specific, and often confusable with common words, we do not try to remove or replace personal names.

All contributors were informed about the intended publishing of the resultant corpus, its public availability and the above anonymization procedure. This process also aided our internal review board application for exemption, as it was deemed that through this method, that our collection procedure did not collect personally identifiable information and was granted exemption from full review. However, the contributors may still not be entirely clear about the automatic anonymization procedure after reading the description. To eliminate their uncertainty and skepticism, we need a straightforward and compelling way to show the anonymization in action. As we mentioned before, our Android app integrates the anonymization process internally, so potential submissions can be previewed as a draft email before sending the SMS batch to the collection server. This manner allows the actual collection data to be previewed, and more likely to convince the contributor of the veracity of the project and collection process.

We created and deployed the website for corpus in January 2011, at the very beginning of our data collection process. The website allows users to browse the current version of the corpus. In calls to contributors, we also link to this live website so that potential contributors can view the scope, volume and content of the current corpus. We feel this is a strong factor in both lowering the anxiety of potential submitters and raising awareness of the corpus in general.

3.5 Live corpus

A few words about the notion of a live corpus. We feel that a live corpus is an emerging concept, in which the corpus grows, is maintained and released on regular, short intervals. A truly live corpus connotes that as soon as a new text is created, it becomes part of the distributed corpus. Such an interpretation can cause replicability problems, as different researchers may use different versions of corpus. Due to this problem, we have chosen to release a new version of the corpus on a regular, monthly basis. This strategy of having regular updates promotes interested parties to stay up to date with the corpus development while allowing the easy identification of a particular version, for papers that wish to use the corpus for comparative benchmarking. The release cycle further helps to demonstrate the trend of our gradual achievement in SMS collection, which, in turn, also may spur more contributors to help in our project. It also allows us to batch corpus administrative duties, such as proofchecking submitted SMS and re-computing demographic statistics, which we describe later.

4 Properties and statistics

Given the variety of methods in our current corpus collection, and its previous history within another former project within our group, it is worthwhile to describe some properties of the resultant SMS collection thus far.

The original corpus was collected by an honors year undergraduate project student, Yijue How, over the course of her project from the fall of 2003 to the spring of 2004 (How and Kan 2005). The main collection method was by SMS transcription to a publicly-accessible website, largely from volunteers that How contacted directly. Hence most volunteers were Singaporeans in the young adult age range. A few contributors gave a large set of messages directly from their phones, foregoing the transcription process. This led to a distinctly bimodal message distribution of having a small "head" of few contributors that submitted many messages that represented depth in the collection, as well as a long "tail" of many contributors that submitted less than 100 messages (The transcription website allowed the submission of 10 messages at a go). Each contributor was assigned an identifier such that the number of messages by each contributor could be tracked. Further details about the demographics of the collection are available from How's complete thesis (How 2004).

We embarked on our current SMS collection project in October 2010. At the time of writing (June 2012), we have collected more than 71,000 messages, with the number still growing. In the remainder of this section, we give statistics on the current corpus (June 2012 version), the demographics of contributors, furnish a cost comparison of the major three sources we used, and wrap up with a discussion on matters related to the corpus release.

4.1 Specification

As of June 2012, our corpus contains 41,790 English messages and 30,020 Chinese messages. In total, 150 contributors contributed English SMS and on average, each

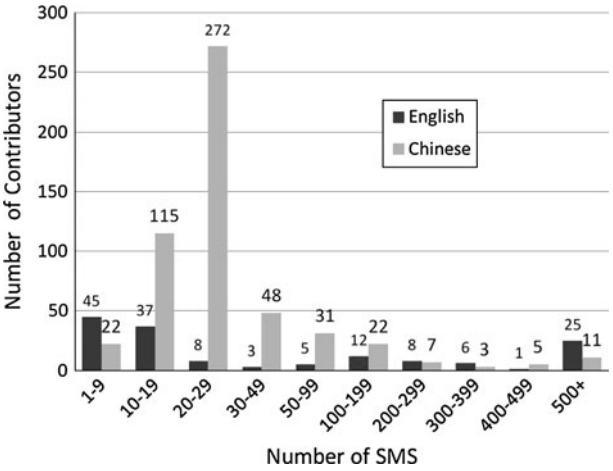

Fig. 2 Distribution of SMS by contributor in the June 2012 corpus version

individual submitted 278.6 messages. In comparison, the total number of Chinese contributors is 536, with an average contribution rate 56.0 messages per person. Detailed histograms showing the number of messages contributed per contributor are given in Fig. 2. Similar to the previous project, both histograms show a peak at the very left side, meaning that only a small proportion of people contributed the bulk of the messages—"the short head". Specifically, 60.0 % of English contributors submitted less than 30 messages. This figure is 76.3 % for Chinese contributors, which explains why the per-contributor Chinese SMS figure is much less than its English counterpart.

The cause of this difference is related to our collection methods. As discussed previously, due to its simplicity, Web-based Transcription is an effective way to obtain mass participation but makes it difficult to collect large amounts of SMS from a single contributor, while both the SMS Export and SMS Upload methods have the opposite characteristics. We fixed the number of SMS per submission in our Web-based Transcription method to 2 or 5 English messages, and 10 or 20 Chinese messages. A small number was used in the initial experiment for exploring a good ratio between monetary reward and workload (number of messages). Using MTurk, we published two batches of tasks to recruit 40 workers to follow our Web-based Transcription to collect English SMS. Unfortunately, due to the resulting high level of cheating and effort expended in verification for English messages, we felt the utility of this method was not tenable, so we stopped using this collection method for English SMS. In contrast, Web-based Transcription was much more effective in Zhubajie, perhaps due to the unavailability of sources to cheat on the task. Up to now, we have retained the use of Web-based Transcription in Zhubajie for the resulting high-quality SMS. This leads to the ascribed difference in demographics in recruiting more Chinese contributors with a resulting smaller per capita figure.

Table 3 demonstrates the number of messages collected by each of our methods. We see that 97.9 % (40,896) of English messages and 46.1 % (13,847) of Chinese

Table 3 Number of SMS collected, broken down by the collection method, in the June 2012 corpus version

Method	English SMS	Chinese SMS
Web-based transcription	894	16,173
SMS export	12,017	12,855
SMS upload	28,879	992

Table 4 Number of SMS and contributors by source in the June 2012 corpus version

Source	English SMS	English contributors	Chinese SMS	Chinese contributors
MTurk	11,274	75	55	19
ShortTask	650	41	0	0
Zhubajie	0	0	24,209	503
Local	28,227	24	3,524	10
Internet community	1,639	10	2,232	4
Total	41,790	150	30,020	536

messages were collected by SMS Export and SMS Upload methods, which are free of typos and contain metadata (ownerships of sender and receiver, and the timestamp of when it was sent). Table 4 then shows the number of SMS and contributors per source. For the English SMS, 27.0 % (11,274) were from workers in MTurk, 67.5 % (28,227) were from local contributors, 1.6 % (650) were from workers in ShortTask and the remaining 3.9 % (1,639) were from the Internet community. For the Chinese SMS, workers from Zhubajie contributed 80.6 % (24,209) of the SMS, local and Internet contributors submitted 11.7 % (3,524) and 7.4 % (2,232), respectively. Currently, only 55 Chinese SMS were contributed by users from MTurk, about 0.2 %.

While the focus of this work is not on the analysis of the corpus, it is instructive to give some basic lexical statistics on the corpus. English messages comprise 33,596 unique tokens, containing 10.8 tokens per message on average; while Chinese messages consist of 3,173 unique tokens, containing 10.3 tokens per message on average.[32]

We also extracted the most frequent words from the English SMS, and compared them with the most frequent words in the Brown Corpus, which is a traditional, wide-domain English corpus, as displayed in Table 5. Not surprisingly, common words, such as “I”, “to”, “the”, “it”, “a”, “is”, “for”, “and”, “in”, and “at”, appear in both wordlists. However, words like “u” (“you”), “haha”, “lol” (“laugh out loud”), exclusively appear in SMS, manifesting its informal and speech-like nature.

Interestingly, we noticed that about 0.5 % of English messages and 3.1 % of Chinese messages are a mixture of English and Chinese language. With further investigation, we found these English messages are mainly from local contributors,

[32] Since we replace sensitive data with pre-defined codes in the anonymization process, the unique token count of the original messages is likely to be higher than what we calculated.

Table 5 The top 20 most frequent words (listed in descending frequency) in the English portion of our corpus and the Brown corpus

Corpus	Words
SMS	I, to, you, the, u, haha, it, a, me, is, for, my, and, in, so, at, can, lol, 's, not
Brown	the, of, and, to, a, in, that, is, was, he, for, it, with, as, his, on, be, at, by, I

which reflects Singapore's multilingual nature. Meanwhile most English words that appears in Chinese message are the common words, such as "hi", "ok", "sorry", "happy", which may indicate that these English words have been frequently used in some young Chinese's daily life, since the majority of contributors are young adults (discussed in more detail later).

4.2 Demographics

Aside from submitting SMS, all contributors were required to fill out an online demographic survey about their background (e.g., age, gender, country, native speaker or not, etc.), texting habits (input method, number of SMS sent daily, years of experience using SMS) and their phone (brand, smartphone or not). Such answers form a user profile for the contributor which is linked to each of their contributed SMS. We accept and report the data for the demographic survey as-is, assuming that contributors answered these questions to the best of their knowledge.

99.6 % of English messages and 94.0 % of Chinese messages thus have associated user profiles. The incompleteness arises from the separation of submitting SMS and filling out the survey in the two collection modes of SMS Export and SMS Upload. Some contributors submitted the messages but later did not do the survey. The phenomenon was more prevalent in the Chinese Zhubajie. During our initial usage pilots of Zhubajie, we approved contributors' SMS immediately, and trusted them to do the survey later on. To stem this problem, we later changed our protocol to only approve the task after receiving both SMSes and the demographic survey. We also had updated the survey once, adding some questions to reveal more detail on some aspects. The user profiles formed from the first version of the survey thus lack answers to a few questions. To make all sets of the demographic data comparable, we inserted "unknown" values to these missing questions as well as to questions that were skipped by contributors.

While it is not possible to conclusively say anything about the demographics of SMS senders in general, our demographic survey allows us limited insight into the characteristics of our corpus' contributors. The confounding factor is that our contributors come largely from crowdsourcing sources, so both the self-selection of participating in crowdsourcing and of SMS use contribute to the demographic patterns we discuss below.

We report both the country of origin, gender and age of contributors, subdivided by the English or Chinese portion of the current corpus. Our English SMS contributors are from 23 countries (in decreasing order of number of contributors): India, Singapore, USA, Pakistan, UK, Bangladesh, Philippines, Malaysia, China, Sri

Lanka, Canada, France, Serbia, Spain, Macedonia, Slovenia, Kenya, Romania, Ghana, Indonesia, Nepal, New Zealand and Nigeria. 71.3 % of them are English native speakers. The top three countries, in terms of the number of contributed messages, are Singapore (62.5 % of SMS), India (19.6 %) and USA (9.9 %). Our Chinese SMS contributors are from 4 countries: China, Singapore, Malaysia and Taiwan. However, the messages are overwhelming from China: China mainlanders contributed 98.4 % of the messages, resulting in 99.2 % of messages originating from native speakers.

For the English portion of the corpus, 23.8 % come from females, 67.4 % are from males, and the remaining 8.8 % are unknown. For the Chinese portion, 35.8 % come from females, 58.0 % come from males, and 6.2 % are unknown. The age distribution shows that the majority of contributors in both portions of the corpus are young adults, as displayed in Fig. 3. In particular, contributors aged 21–25 make up the largest portion, taking up 38.0 % of the English and 56.9 % of the Chinese SMS contributors, respectively. They submitted 39.7 % of the English and 66.5 % of the Chinese SMS, respectively. There is an even higher skew towards the 16–20 age group among English contributors, largely due to the fact that 67.5 % of the English SMS originate from local students in our university.

Our survey also reveals other pertinent details on the texting habits and phone models of contributors. We display the distribution of contributors' years of experience with SMS in Fig. 4. The largest English SMS portion lies in 5–10 years (46.7 % of contributors), and the second largest portion is 3–5 years (27.3 %). Similarly, most Chinese contributors have used SMS for 3–5 years (31.0 %) and 5–10 years (31.0 %). This phenomenon is in accord with the age distribution: young adults represent the majority of the contributors. We may also posit that more of our Chinese contributors have acquired their SMS-capable phone more recently than our English contributors, as we see a smaller proportion of users in the 5–10 year range.

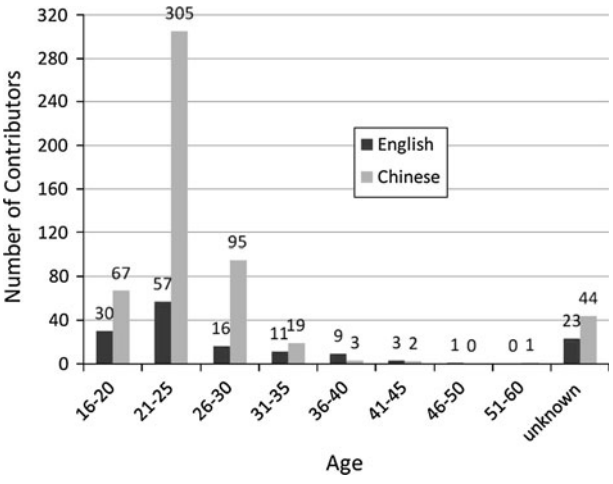

Fig. 3 Distribution of contributor's age in the June 2012 corpus version

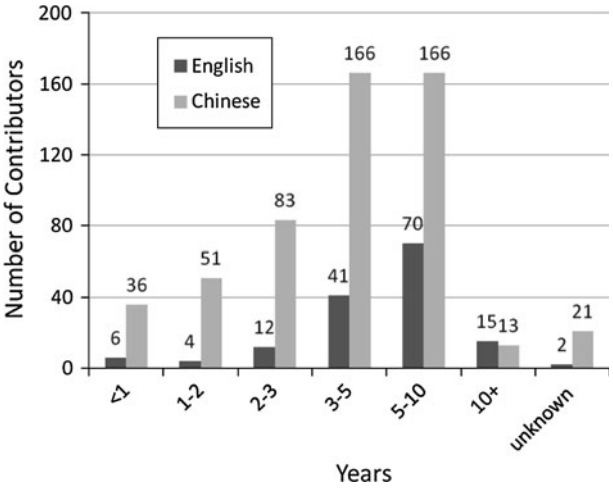

Fig. 4 Contributors' SMS experience (in years) in the June 2012 corpus version

How often our contributors' send SMS is presented in Fig. 5. We observe an interesting phenomenon that there is a general increasing trend—from 16.7 % (2–5 SMS daily), 16.0 % (5–10), to 34.0 % (10–50)—for English contributors; while for Chinese contributors there is a general decreasing trend—from 32.8 % (2–5), 25.4 % (5–10), to 22.2 % (10–50). For the English portion, 18.0 % of contributors send more than 50 SMS everyday. We posit these frequent texters are likely to use SMS to carry on conversations (thus needing more messages), rather than for sending one-off messages.

Differences in SMS input methods were also revealed in our survey. Three common input methods, multi-tap, predictive and full keyboard, account for 35.3, 28.7 and 26.0 % of English SMS contributors respectively. The remaining 8.6 % of contributors used other input methods. Chinese input methods also largely consist of three input methods, *pinyin* (拼音) , *wubi* (五笔) and *bihua* (笔画) ; these account for 86.4, 3.7 and 4.1 % of contributors' usage. As the pronunciation-based input method, *Pinyin* is very easy for Chinese speakers to learn and use, because it matches the way they think (Zhou et al. 2007). The only requirement of using *pinyin* is familiarity with Mandarin pronunciation of characters (pinyin), which is easy to meet since all children in the mainland China are required to learn pinyin in school. The latter two, subsumed to shape-based input methods, are not limited to any particular Chinese dialect but require additional memory of mapping key codes to strokes or shapes. These are the reasons why *pinyin* is the dominant Chinese input method, while the other two only have a small user base.

As displayed in Fig. 6, the majority of contributors own a Nokia phone, which accounts for 43.3 % of English SMS contributors and 51.3 % of Chinese SMS contributors. However, these figures cannot fairly present the general popularity of phone brands, for the reason that we only provide links to Nokia utility software for SMS Export collection method. In addition, 49.3 % of English SMS contributors,

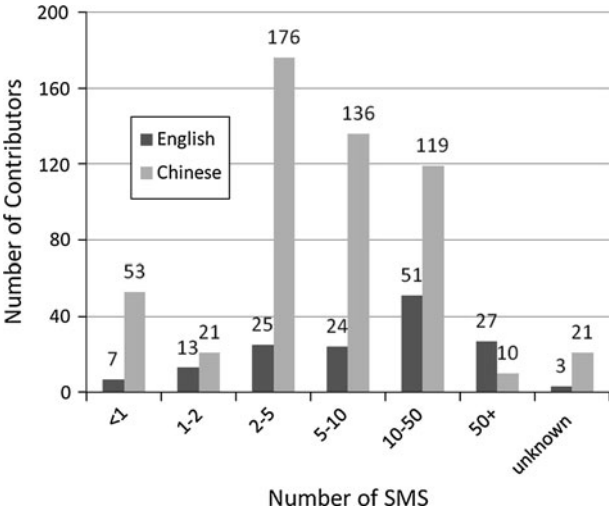

Fig. 5 Distribution of our contributors' frequency in sending SMS (daily) in the June 2012 corpus version

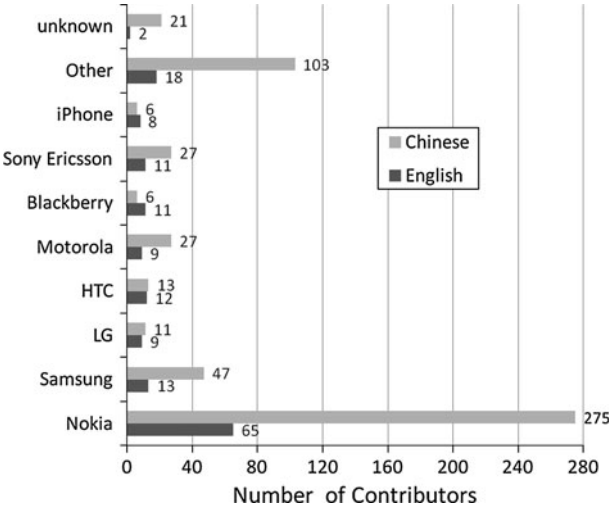

Fig. 6 Contributors' phone brand in the June 2012 corpus version

and 45.9 % of Chinese SMS contributors were users of a smartphone. This question was not included in the first version of survey used in MTurk, which resulted in "unknown" for 38.0 % of English SMS contributors and 7.5 % of Chinese SMS contributors.

For a more in-depth analysis beyond the scope of our article, we invite you to visit the corpus website where the full per-question answers for our demographic data is available.

4.3 Cost

As MTurk, ShortTask and Zhubajie are crowdsourcing sites where participants are motivated by profit, we compensated contributors monetarily. The same was true for the local collections that we ran. For the calls for participation via the Internet to linguistic, text corpus and phone manufacturer's communities, we felt that their self-motivation to see the corpus' success would be a strong motivator to participate. Thus, for these communities we did not provide any compensation.

Table 6 shows the reward scheme used in our MTurk collection runs using the SMS Export and SMS Upload methodologies. For example, a contribution of 400 messages is remunerated with USD 4.50 in total (USD 4.00 base pay with a USD 0.50 bonus). In ShortTask, we only published tasks using Web-based Transcription method, with the reward of USD 0.08 for 10 messages per submission. Since Zhubajie was a new venue for linguistic data crowdsourcing, we tuned our rewards scheme based on a pilot (as shown in Tables 7, 8). Our pilot showed that workers were very eager to participate, thus we decreased the reward amount in our subsequent runs, which did not dampen workers' enthusiasm for completing the contribution tasks. For the four batches of English (two batches) and Chinese (remaining two batches) SMS Web-based Transcription, we instead aimed for breadth of participants. In each task, we recruited 20 workers with USD 0.10 as the reward for individual workers. Finally, we also recruited contributors locally, whose reward scheme is displayed in Table 9.

Table 6 Reward scheme in MTurk (in USD)

Total number	Base eeward	Bonus
10–400	0.10	1/100 Msg
401–1,000	4.00	1/200 Msg
≥1,000	7.00	0

Table 7 Reward scheme 1 in Zhubajie (in CNY)

Total number	Base reward	Bonus
10–100	1.00	1/10 Msg
101–400	10.00	1/20 Msg
401–1,000	25.00	1/40 Msg
≥1,000	40.00	0

Table 8 Reward scheme 2 in Zhubajie (in CNY)

Total number	Base reward	Bonus
20–100	1.00	1/20 Msg
101–400	5.00	1/25 Msg
401–1,000	17.00	1/40 Msg
≥1,000	32.00	0

Table 9 Reward scheme for local collection (in SGD)

Total number	Base reward	Bonus
20–50	2.00	1/10 Msg
51–200	5.00	1/30 Msg
201–600	10.00	1/40 Msg
≥600	20.00	0

Table 10 Cost comparison (in unit USD)

Source	Total number	Total cost	Cost per message
MTurk	11,330	USD 92.30	USD 0.0081
ShortTask	650	USD 6.00	USD 0.0092
Zhubajie	24,209	CNY 888.50	CNY 0.0367 (∼USD0.0058)
Local	31,751	SGD 420.00	SGD 0.0132 (∼USD0.0106)

Since the labor cost of collecting additional SMS over the first few is small (arguably even negligible in the Upload and Export methodologies), we incentivize additional submitted messages with a bonus reward, based on the number of messages that exceed the base requirement. For all reward schemes, the bonus amount diminishes with the amount of additional messages submitted, with a maximum payment capped to allow us a measure of budgetary control. Table 10 shows the cost comparison between sources. We spent 92.30 US dollars in MTurk, 6.00 US dollars in ShortTask, 888.50 Chinese yuan in Zhubajie (all inclusive of commission fees) and 420.00 Singapore dollars for our local collection. Standardizing all currency amounts to US dollars[33] allows us to compute the cost per message for a clear comparison. In a nutshell, local collection is the most expensive method, while crowdsourcing remains an effective and economical method to collect our corpus data. We also note that Zhubajie is a little cheaper than MTurk, but that it may only be applicable to certain languages (i.e., Chinese). Due to our very limited usage of ShortTask, it doesn't make much sense to compare cost in ShortTask with other sources.

4.4 Towards a publicly-available live corpus

Our corpus, consisting of both messages and associated user profiles, has been released publicly since February 2011. To achieve our goal of making an accessible dataset, we have pursued an open license, public domain development methodology that involved first the convincing and later the blessing of our university's intellectual property department. For the aim of making a general purpose dataset, we have tried to incorporate a balanced approach for user profiling; by requiring contributors to answer a set of demographics and including them with the dataset. So as to make the corpus as large as possible, we incorporate all messages that we collected through all of the methodologies used, although this means varying levels

[33] On 21 April 2012 when most payments were made. 1 SGD = 0.8015 USD, 1 CNY = 0.1585 USD.

of quality among subportions of the corpus (e.g., some SMS may not have an associated user profile).

To make the corpus convenient for researchers to access, we also pioneer the distribution of the corpus both as an XML file as well as a database dump in MySQL. Potential SMS researchers or contributors can also browse and download the corpus directly on the corpus website, and access dataset statistics, all without the need to handle the raw corpus files or compute the statistics themselves.

Our statistics help prospective users grasp a general understanding about the demographic and representativeness of our corpora. The corpus and statistics are updated on a monthly basis, since our collection is still in progress. Moreover, the SMS can be directly browsed on our website, which provides a convenient way to learn about our corpus without the need to process raw files and helps potential contributors to understand our anonymization strategy by viewing other's messages.

5 Discussion

We now comment on three open questions surrounding the crowdsourcing of our public corpus of SMS. First, what do workers think about contributing SMS to a public corpus? Second, how does the Chinese crowdsourcing site of Zhubajie compare with Amazon's Mechanical Turk? Third, as some crowdsourcing is motivated by altruism, how feasible is it to collect SMS without offering any monetary reward?

5.1 Reactions to our collection efforts

The corpus was collected under our university's institutional review board (IRB) exemption policy and important identifiers in the corpus have been replaced by placeholder tokens for deidentification purposes. However, our experience through on this project over the last year has shown us that the privacy issues surrounding the collection of SMS is still very much a concern. Even among our research group, members were largely unwilling to donate SMS even with the safeguards in place. This may have been partially due to the fact that the authors need to manually review the submitted SMS, and that the review process may identify the contributor. This fear was further validated in our local collection drive, where potential contributors worried that their friends may review the corpus and identify their messages, especially through the mention of certain names in SMS (Personal names are not replaced by any code as given in Table 2, as many personal names are also common words).

Privacy concerns were also paramount in our crowdsourcing work with Amazon Mechanical Turk, and ultimately caused our account with MTurk to be closed. Amazon sent us several notices that our tasks violated their terms of service. Through private correspondence with Panagiotis Ipeirotis, whose research involves detecting spam in MTurk, we found out our tasks were routinely classified by Turkers as a spam or phishing task, despite our attempts to give credibility to the

project through the creation of the corpus webpage and browsing interface.[34] Unfortunately, even with repeated attempts to contact Amazon to clarify the nature of our notice of breach of service, our MTurk account was suspended without further detail from Amazon.

Similar concerns surfaced on our calls for SMS contribution in the Chinese realm. On the mobile phone forums that we solicited participation from, we encountered a few skeptic replies. For these forums, we had advertised our Android SMS uploader application, with the appropriate links to an explanation of the project and our corpus' web page. Several posters voiced their concern that the software might be malware looking to steal private information (especially given the inclusive permissions set that our application needs access to). These were valid concerns as some previous mobile application recommendations did turn out to be malware, so readers were being cautious before installing any software.

5.2 Zhubajie compared with MTurk

Zhubajie is one of a growing number of middleware websites in China offering crowdsourced labor. Forecasted online transactions on the site are expected to surpass 500 million CNY ($\sim$ 78 million USD) in 2011 alone.[35]

We found Zhubajie to be a good platform to recruit Chinese contributors. However, unlikely its western counterparts, Zhubajie, as well as other Chinese "witkey" websites, has not been widely exploited for research data collection in computer science community. Most existing academic work involving witkey have focused on their business and economic aspects, studying the user behavior (Yang et al. 2008; DiPalantino et al. 2011; Sun et al. 2011), analyzing the participation structure (Yang et al. 2008), exploring the anti-cheating mechanism (Piao et al. 2009), and investigating the business model (Zhang and Zhang 2011). For these reasons, we feel it would be useful to give a more comprehensive overview of Zhubajie, focusing on five aspects: its conceptualization, characteristics of typical tasks, cost, completion time and result quality. We compare Zhubajie against the now familiar Amazon Mechanical Turk (MTurk) when appropriate.

- *Concepts*. While both Zhubajie and MTurk can be characterized as mechanized labor, to be more accurate, Zhubajie's form of crowdsourcing originates from "witkey"—an abbreviation of the phrase "the key of wisdom". The concept of Witkey was put forward by Liu in 2005, published later in Liu et al. (2007), where he defines its core concept was to trade knowledge, experience and skill as merchandise.
- *Tasks*. MTurk focuses on tasks that take a small or moderate amount of time to complete, while most tasks of Zhubajie require expertise and can take longer to complete. Designing logos, software development and revising resumé are typical Zhubajie tasks. Zhubajie also classifies tasks into a detailed hierarchical classification system with major 9 top categories, and 2–13 secondary categories

[34] In fact, these were some of Ipeirotis' suggestions to ameliorate the problem, so credit is due to him.

[35] http://economictimes.indiatimes.com/tech/internet/idg-backed-chinese-website-zhubajie-to-list-in-us-in-3-years/articleshow/9478731.cms.

per top-level category and 1–22 third level categories.[36] It requires requesters to specify a third level category for each task. This is unlike MTurk, which eschews task classification altogether. Zhubajie's detailed browsable task hierarchy reflects its Chinese base, as the Chinese population often prefers selection and browsing over searching (as browsing only requires clicking but searching requires inputting Chinese characters, which is still somewhat difficult).[37]

This task classification leads to different service characteristics in Zhubajie and MTurk. MTurk provides keyword search and 12 different sorting options for results display (newest/oldest tasks, most/fewest available tasks, highest/lowest reward, etc.). The survey results of Chilton et al. (2010) shows that the newest and most available tasks are the most popular sorting strategies employed by workers, and that the first two pages of listings are most important. Ipeirotis (2010b) pointed out that if a task is not executed quickly enough, it will fall off these two preferred result listings and is likely to be uncompleted and forgotten. This is in accord with our experience in trying to recruit workers. In Zhubajie, even 10 days after publishing the task, we still received new submissions from workers. This is contrary to our experience with MTurk, where we did not receive many new submissions after the first two days. Due to the detailed task hierarchy in Zhubajie, potential workers can easily target specific tasks matching their expertise and interests, ameliorating the recency-listing concerns prevalent in MTurk. A few outstanding workers, based on reputation and total earnings, are also featured as top talents for each task category. Requesters can invite and recruit talents to fulfill the task. These properties all help aid matching workers to tasks in comparison with MTurk.

- *Cost*. The demand of expertise in Zhubajie also impacts the payment distributions in two websites. In MTurk, the lowest payment is just USD 0.01 and 90 % of tasks pay less than USD 0.10 (Ipeirotis 2010b). Compared with the tiny rewards offered in MTurk, the rewards in Zhubajie are significantly higher, with about USD 0.15 (CNY 1.00) as the lowest reward and about USD 182 (CNY 1181) as the average reward for the year 2010. Also, though both services make profit by collecting commission fees, they differ as to whom the commission is charged from: MTurk charges the requester 10 %, but Zhubajie charges the worker 20 % commission. Furthermore, In Zhubajie, task rewards come in two flavors: they can be set by requesters or they can be bid on by workers, which term as *contract tasks* and *contest tasks*, respectively. In this sense, our task—and MTurk tasks in general—are thus contract tasks. For our SMS collection thus far, Zhubajie has turned out to be more economical by 28.4 %; we spent USD 0.0058 and USD 0.0081 per message in Zhubajie and MTurk, respectively.

[36] As of 18 June 2012.

[37] http://www.smashingmagazine.com/2010/03/15/showcase-of-web-design-in-china-from-imitation-to-innovation-and-user-centered-design.

- *Completion time*. Here we look at the task completion time with respect to collection methodology. With the SMS Upload method, it took 2 full days to receive 3 English submissions via MTurk; and worse, there were no submissions at all from Zhubajie. This may be explained by the current low popularity of Android smartphones among Chinese SMS contributors. In contrast, under the SMS Export collection method, we received 16 Chinese submissions from Zhubajie in 40 days, and 27 English submissions from MTurk in 50 days. The collection in this method was slow in both platforms.
  Web-based Transcription offers the most telling demographic difference. Our MTurk tasks took 2 days to complete, collecting 40 valid English submissions and 20 days for 20 valid Chinese submissions (each submission having 2 or 5 individual SMSes). In contrast, the same Chinese SMS task, when published to Zhubajie, usually took less than 30 min to complete to collect for 20 submissions. We ascribe the quick completion in Zhubajie to two reasons. First, Zhubajie has a specific task category for SMS tasks—the *creative greetings* category. This category typically asks workers to compose a creative SMS and send it to bless a designated recipient (i.e., write a poem to wish someone to get well soon), as it is uplifting in China to receive lots of blessing from the general public. It is also a relatively popular category among workers for its short completion time and low expertise requirements. Second, among the tasks in the creative blessing category, our task is easier, faster and more profitable. Other tasks require workers to design or craft a creative blessing and send it to an actual recipient which incurs cost, but the payment is usually identical to ours.
- *Quality* has emerged as a key concern with crowdsourcing, and it is clear that this is a concern for our task as well. MTurk employs several strategies to help requesters control for quality: a requester can require certain qualifications based on the worker's location and approval rate, publishing a qualification test prior to the real task and blocking poorly performing workers from a task. To attract the maximal number of potential contributors, we did not set any qualifications in MTurk. In contrast, Zhubajie does not provide built-in quality control system. Tasks, when completed in either MTurk or Zhubajie, can be rejected by the requester if it does not meet their criteria for a successful task. Table 11 shows our approval rate of completed tasks for each collection method in the two crowdsourcing websites.
  As we have previously described the problems with Web-based Transcription (in that contributors can enter anything they want, including SMS copied from SMS sites on the web), we expected this poorly-performing methodology to have the highest rejection rate. Surprisingly, in fact, it was quite the opposite:

Table 11 Comparison of the approval rate among crowdsourced venues

Collection method	MTurk (%)	Zhubajie (%)
Web-based transcription	62.50	85.03
SMS export	16.58	57.14
SMS upload	42.86	No submissions

Web-based Transcription tasks enjoyed a higher approval rate than the other methods, across both sites. We believe the difference in financial incentives of the methodologies explains this. While the payment for Web-based Transcription was only USD 0.10 in MTurk and CNY 1.00 in Zhubajie, the payment of the other two methods can be as high as USD 7.00 in MTurk and CNY 40.00 in Zhubajie. Intrigued by the high reward, some workers attempted to cheat on these higher-yield methods. This validates findings by Mason and Watts (2009), who states that increased financial compensation may not improve task quality, and sometimes may even result in poorer quality.

SMS Upload approval rates were also relatively better than those for SMS Export. Workers using SMS Upload needed to type a unique code generated by the mobile application, which may discourage errant workers from cheating since they would have not known how to generate the correct code without doing the task through the application. In contrast, the SMS Export method allowed contributors to upload any files (even those not containing SMS at all), making it easier for potential cheaters to try their luck.

Finally, we judged the overall quality of work done by Zhubajie workers to be much higher than that of MTurk. We attribute this to the open worker reputation system of Zhubajie. In MTurk, the worker's approval rate is the sole figure to judge whether a worker's work is good or bad. Zhubajie stores comments on workers as well as calculating a positive comment rate (similar to MTurk's approval rate) and a reputation rank based on earned income. In some cases, if many workers compete for one task, the requester can pick over the potential, qualified workers based on these positive comments and reputation rank. Finally, Zhubajie's administrators will warn serious cheaters and even lock their account. MTurk metes out no official punishment for cheaters (unlike our experience for requesters), and requesters have to manually blacklist poorly performing workers in their tasks.

5.3 Altruism as a possible motivator

Mechanized labor is just one possible form of crowdsourcing that can result in workers performing a task. There have been a number of surveys on crowdsourcing, including a recent survey on finding optimal the method for crowdsourcing corpora and annotations (Wang et al. 2012b). Are other, non-profit oriented approaches feasible for collecting sensitive data? Could altruistic motivational factors work?

To answer these questions, we emailed calls for participation to the natural language and corpora community,[38] and two Chinese mobile phone forums[39] for voluntary, non-compensated contributions of SMS. This was a part of our methodology from the beginning as described earlier. Unfortunately, the results were not promising. As shown in Table 4 of Sect. 4, we received only 10 anonymous contributions, totalling 206 English and 236 Chinese SMS, respectively

[38] Via the Corpora List, corpus4u forum (Chinese), the 52nlp blog (Chinese).

[39] hiapk and gfan.

by these methods.[40] Compared with the rest of the for-reward collection methods, this method was a failure, and we do not recommend this method for collection in its current guise.

Our findings are contrary to the sms4science project, which succeeded in gathering a large number of messages through pure voluntary contribution. Though a small portion of contributors were randomly selected by lottery to win prizes,[41] we still deem their collection method as a purely voluntary contribution, as there is no monetary compensation. However, we note two key differences between our call and theirs. The sms4science project obtained support from phone operators, making it free for potential contributors to forward their SMS to the project's service collection number. This lowers the difficulty of contributing messages as no software installation or tedious export is necessary (but note that it does destroy some message metadata that we can collect through our other methods). Probably more important was that the sms4science project conducted large-scale publicity; its call for participation was broadcast in national media including press, radio, television (Fairon and Paumier 2006; Dürscheid and Stark 2011). For example, the Belgium project was reported in five newspapers and six websites within two weeks.[42] For our project, due to the limited publicity vehicles and technical constraints, attracting people to contribute SMS only for the sake of science was difficult.

Our appeal to the research community did not yield many SMS for the corpus, but did give us further convictions that we were performing a necessary and useful task. Several researchers supported our project by writing words of encouragement and sharing their personal difficulties with gathering SMS for research.

6 Conclusion

In order to enlarge our 2004 SMS corpus and keep up with the current technology trends, we resurrected our SMS collection project in October 2010 as a live corpus for both English and Mandarin Chinese SMS. Our aim in this revised collection is fourfold: to make the corpus (1) as representative as possible for general studies; (2) accurate with fewer transcription errors; (3) released to the general domain, copyright-free for unlimited use; (4) useful to serve as a reference dataset. To achieve these four goals, we adopt crowdsourcing strategies to recruit contributors from a wide spectrum of sources, using a battery of methodologies to collect the SMS. As SMS often contains sensitive personal data, privacy and anonymization issues have been paramount and have influenced the resulting design of the collection methods and the corpus data itself.

We are very encouraged by the results so far. As of the June 2012 version of the corpus, we have collected 41,790 English SMS and 30,020 Chinese SMS, with a

[40] From additional personal contacts, we obtained an additional 1,433 English and 1,996 Chinese SMS respectively.

[41] http://www.smspourlascience.be/index.php?page=14.

[42] http://www.smspourlascience.be/index.php?page=16.

cost of 574 US dollar equivalent and approximately 310 human hours of time (inclusive of the Android app implementation, website creation and update). As the project is a live corpus project, these figures are growing as the collection continues. To the best of our knowledge, our corpus is the largest English and Chinese SMS corpus in the public domain. We hope our corpus will address the lack of publicly-available SMS corpora, and enable comparative SMS related studies.

A novel aspect of our collection is the implementation of mobile phone applications for collection. We adapted an SMS backup software to also serve as a platform for contributing SMS. It is the first application for such purpose and is easily adapted for other SMS collection purposes. We also reported on the first use of Chinese crowdsourcing (also known as "Witkey") for collecting corpora and have discussed the significant differences between Chinese and traditional crowdsourcing in the English-speaking world, as embodied by Amazon's Mechanical Turk. Finally, we explored the possibility of calling for SMS contribution without compensation, but found that altruistic motivation is not sufficient for collecting such data. Rather, our lessons learned indicate that large-scale publicity is the key to success.

We continue to enlarge our SMS collection, as part of our interpretation of what a live corpus project means. Given the importance of SMS in carrying personal communication in our society, and the low-cost methods we have found to collect message contents (suitably scrubbed) and demographic data, we are encouraged to continue to fund this work internally, to encompass more languages and a wider population of users. We also plan to explore other SMS collection methods, such as an iOS (i.e., iPhone, iPad) application, and benchmark their efficacy against the methods we have analyzed in this article. One prominent limitation of our work is the current problem of "sent-message" restrictions, that excludes the possibility of studying texting conversations in our corpus. Collecting such bi-directional messages are very useful for studies like discourse and conversation analyses. To address the issue, we plan to employ pairs of participants who are willing to share their messages, and collect the texting conversations.

As for downstream use, with further funding, we may annotate the corpus (either automatically or manually) with part-of-speech, translations into other languages, or other semantic markup. The resulting corpus may then be used in other natural language studies and applications (e.g., machine translation). Other downstream projects that we know of may also make their annotations and additional collection of SMS available as collaborative or sister projects to our NUS SMS corpus.

7 Data

The corpus described in this paper is publicly available at our corpus website (http://wing.comp.nus.edu.sg/SMSCorpus).

Acknowledgments We would like to thank many of our colleagues who have made valuable suggestions on the SMS collection, including Jesse Prabawa Gozali, Ziheng Lin, Jun-Ping Ng, Kazunari Sugiyama, Yee Fan Tan, Aobo Wang and Jin Zhao. The authors gratefully acknowledge the support of

the China-Singapore Institute of Digital Media's support of this work by the "Co-training NLP systems and Language Learners" grant R 252-002-372-490.

References

Bach, C., & Gunnarsson J. (2010). *Extraction of trends is SMS text*. Master's thesis, Lund University.

Back, M. D., Küfner, A. C., & Egloff, B. (2010). The emotional timeline of September 11, 2001. *Psychological Science, 21*(10), 1417–1419.

Back, M. D., Küfner, A. C. P., & Egloff, B. (2011). Automatic or the people?: Anger on September 11, 2001, and lessons learned for the analysis of large digital data sets. *Psychological Science, 22*(6), 837–838.

Barasa, S. (2010). *Language, mobile phones and internet: A study of SMS texting, email, IM and SNS chats in computer mediated communication (CMC) in Kenya*. Ph.D. thesis, Leiden University.

Bodomo, A. B. (2010). *The grammar of mobile phone written language, Chap. 7* (pp. 110–198). Hershey: IGI Global.

Callison-Burch, C., & Dredze M. (2010). Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's mechanical turk*, CSLDAMT '10, Stroudsburg, PA, pp. 1–12. Association for Computational Linguistics.

Chilton, L. B., Horton, J. J., Miller, R. C., & Azenkot, S. (2010). Task search in a human computation market. In *Proceedings of the ACM SIGKDD workshop on human computation*, HCOMP '10, pp. 1–9. New York, NY: ACM press.

Choudhury, M., Saraf, R., Jain, V., Sarkar, S., & Basu A. (2007). Investigation and modeling of the structure of texting language. In *In Proceedings of the IJCAI-workshop on analytics for noisy unstructured text data* (pp. 63–70) .

Crystal, D. (2008). *Txtng: The Gr8 Db8*. Oxford: Oxford University Press.

Denby, L. (2010). *The language of twitter: Linguistic innovation and character limitation in short messaging*. Undergraduate thesis, University of Leeds.

Deumert, A., & Oscar Masinyana, S. (2008). Mobile language choices–the use of English and Isixhosa in text messages (sms): Evidence from a bilingual South African sample. *English World-Wide, 29*(2), 117–147.

DiPalantino, D., Karagiannis, T., & Vojnovic M. (2011). Individual and collective user behavior in crowdsourcing services. Technical report, Microsoft Research.

Dürscheid, C., & Stark, E. (2011). *SMS4science: An international corpus-based texting project and the specific challenges for multilingual Switzerland, Chap. 5*. Oxford: Oxford University Press.

Elizondo, J. (2011). *Not 2 Cryptic 2 DCode: Paralinguistic restitution, deletion, and non-standard orthography in text messages*. Ph. D. thesis, Swarthmore College.

Elvis, F. W. (2009). The sociolinguistics of mobile phone sms usage in Cameroon and Nigeria. *The International Journal of Language Society and Culture*, (28), 25–41.

Fairon, C., & Paumier, S. (2006). A translated corpus of 30,000 French SMS. In *Proceedings of language resources and evaluation*.

Gao, Q., & Vogel, S. (2010). Consensus versus expertise: A case study of word alignment with mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's mechanical turk*, CSLDAMT '10, Stroudsburg, PA, pp. 30–34. Association for Computational Linguistics.

Gibbon, D., & Kul, M. (2008). Economy strategies in restricted communication channels: A study of polish short text messages. In *Proceedings of 5th Internationale Tagung Perspektiven der Jugendspracheforschung*.

Grinter, R., Eldridge, M. (2003). Wan2tlk?: Everyday text messaging. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 441–448. ACM.

Herring, S., & Zelenkauskaite, A. (2009). Symbolic capital in a virtual heterosexual market: Abbreviation and insertion in Italian iTV SMS. *Written Communication, 26*(1), 27.

How, Y. (2004). Analysis of sms efficiency. Undergraduate thesis, National University of Singapore.

How, Y., & Kan, M. (2005). Optimizing predictive text entry for short message service on mobile phones. In *Proceedings of HCII*. Lawrence Erlbaum Associates.

Hutchby, I., & Tanna, V. (2008). Aspects of sequential organization in text message exchange. *Discourse & Communication, 2*(2), 143–164.

Ipeirotis, P. (2010a). *Demographics of mechanical turk*. New York University, Working Paper No: CEDER-10-01.

Ipeirotis, P. G. (2010b). Analyzing the amazon mechanical turk marketplace. *XRDS 17*, 16–21.

Yang, J., Adamic, L.A., & Ackerman, M. S. (2008). Competing to share expertise: The taskcn knowledge sharing community. In *Proceeding of the international AAAI conference on weblogs and social media*.

Jonsson, H., Nugues, P., Bach, C., & Gunnarsson J. (2010). Text mining of personal communication. In *2010 14th international conference on intelligence in next generation networks*, pp. 1–5. IEEE.

Ju, Y., & Paek, T. (2009). A voice search approach to replying to SMS messages in automobiles. In *Proceedings of Interspeech* (pp. 1–4). Citeseer.

Kasesniemi, E. -L., & Rautiainen, P. (2002). Mobile culture of children and teenagers in finland. In *Perpetual contact*, New York, NY, pp. 170–192. Cambridge: Cambridge University Press.

Lexander, K. V. (2011). Names U ma puce : Multilingual texting in Senegal. Working paper.

Ling, R. (2005). The sociolinguistics of sms: An analysis of sms use by a random sample of norwegians. *Mobile Communications Engineering: Theory and Applications, 26*(3), 335–349.

Ling, R., & Baron, N. S. (2007). Text messaging and im. *Journal of Language and Social Psychology, 26*(3), 291–298.

Liu, F., Zhang, L., & Gu, J. (2007). The application of knowledge management in the internet—Witkey mode in China. *International journal of knowledge and systems sciences, 4*(4), 32–41.

Liu, W., & Wang, T. (2010). Index-based online text classification for sms spam filtering. *Journal of Computers, 5*(6), 844–851.

Mason, W., & Watts, D.J. (2009). Financial incentives and the "performance of crowds". In *Proceedings of the ACM SIGKDD workshop on human computation* (pp. 77–85). HCOMP '09, New York, NY. ACM.

Munro, R., & Manning, C. D. (2012). Short message communications: users, topics, and in-language processing. In *Proceedings of the 2nd ACM symposium on computing for development*, ACM DEV '12, New York, NY, pp. 4:1–4:10. ACM.

Ogle, T. (2005). Creative uses of information extracted from SMS messages. Undergraduate thesis, The University of Sheffield.

Piao, C., Han, X., & Jing, X. (2009). Research on web2.0-based anti-cheating mechanism for witkey e-commerce. In *Second international symposium on electronic commerce and security, 2009. ISECS '09, Volume 2* (pp. 474–478).

Pietrini, D. (2001). X'6 :-(?": The sms and the triumph of informality and ludic writing. *Italienisch 46*, 92–101.

Quinn, A. J., & Bederson, B. B. (2009). A taxonomy of distributed human computation. Technical report, University of Maryland, College Park.

Resnik, P., Buzek, O., Hu, C., Kronrod, Y., Quinn, A., & Bederson, B. B. (2010). Improving translation via targeted paraphrasing. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, EMNLP '10, Stroudsburg, PA, pp. 127–137. Association for Computational Linguistics.

Rettie, R. (2007). Texters not talkers: Phone call aversion among mobile phone users. *PsychNology Journal, 5*(1), 33–57.

Ritter, A., Cherry, C., & Dolan, W. B. (2011). Data-driven response generation in social media. In *Proceedings of the conference on empirical methods in natural language processing*, EMNLP '11, Stroudsburg, PA, pp. 583–593. Association for Computational Linguistics.

Schlobinski, P., Fortmann, N., Groß, O., Hogg, F., Horstmann, F., & Theel, R. (2001). Simsen. eine pilotstudie zu sprachlichen und kommunikativen aspekten in der sms-kommunikation. *Networx*.

Segerstad, Y. (2002). *Use and adaptation of written language to the conditions of computer-mediated communication use and adaptation of written language to the conditions of computer-mediated communication*. Ph.D. thesis, University of Gothenburg.

Shortis, T. (2001). 'new literacies' and emerging forms: Text messaging on mobile phones. In *International literacy and research network conference on learning*.

Sotillo, S. (2010). *SMS texting practices and communicative intention, chapter 16* (pp. 252–265). Hershey: IGI Global.

Sun, Y., Wang, N., & Peng, Z. (2011). Working for one penny: Understanding why people would like to participate in online tasks with low payment. *Computers in Human Behavior, 27*(2), 1033–1041.

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

Thurlow, C., & Brown, A. (2003). Generation Txt? The sociolinguistics of young people's text-messaging. *Discourse Analysis Online, 1*(1), 30.

Walkowska, J. (2009). International joint conference intelligent information systems (IIS 2009). *Recent advances in intelligent infomation systems*, Warsaw, 2009. ISBN: 978-83-60434-59-8.

Wang, A., Chen, T., & Kan, M.-Y. (2012a). Re-tweeting from a linguistic perspective. In *Proceedings of the second workshop on language in social media*, Montréal, Canada, pp. 46–55. Association for Computational Linguistics. Accesses Mar 2012.

Wang, A., Hoang, C., & Kan, M.-Y. (2012b). Perspectives on crowdsourcing annotations for natural language processing. *Language Resources and Evaluation*. doi:10.1007/s10579-012-9176-1.

Yang, J., Adamic, L. A., & Ackerman, M. S. (2008). Crowdsourcing and knowledge sharing: Strategic user behavior on taskcn. In *Proceedings of the 9th ACM conference on electronic commerce, EC '08*, New York, NY, pp. 246–255. ACM.

Zhang, L., & Zhang, H. (2011). Research of crowdsourcing model based on case study. In *8th international conference on service systems and service management (ICSSSM), 2011*, pp. 1–5.

Zhou, K.-l., Lv, Q., Zhang, Y.-h., Pan, J.-s., & Qian, P.-d. (2007). Towards evaluating chinese character digital input system. *Journal of Chinese Information Processing, 21*(1), 67–73.

Žic Fuchs, M., & Tudman Vukovic, N. (2008). Communication technologies and their influence on language: Reshuffling tenses in Croatian SMS text messaging. *Jezikoslovlje, 2*(9.1-2), 109–122.