Chats, Tweets and SMS in the SoNaR Corpus: Social Media Collection

Maaske Treurniet, Eric Sanders

CLST, Centre for Language and Speech Technology
Radboud University Nijmegen
Nijmegen, The Netherlands
[m.treurniet, e.sanders]@let.ru.nl

Abstract—In this paper we discuss the compilation of a social media corpus with chats, tweets and SMS text messages as part of the SoNaR corpus, a 500-million word reference corpus of written Dutch, comprising many different text categories. Social media are more and more becoming part of everyday life, which makes the need for social media corpora an urgent matter for research. Special focus was addressed to the collection of metadata and intellectual property rights (IPR). IPR was obtained both through licenses with platform owners, and by consent of individual contributors. Recruitment of participants was done by means of *free publicity*. The data will be available for research and commercial use.

Keywords: *SoNaR, social media, chats, tweets, SMS, corpus building*

I. INTRODUCTION

SoNaR ([9], [11]) is a major reference corpus of written Dutch collected by a cooperation of several linguistic research groups in the Netherlands and Flanders. The corpus was finished 1 March 2012 and contains 500 million words from a variety of text sources, e.g. books, newspapers, manuals. Social media, like online chats, internet fora, blogs, twitter and text messages (SMS) are also included in the corpus. The chats, tweets and SMS will be made available together with the rest of SoNaR via the HLT centre¹, the Dutch-Flemish centre for storage, maintenance and distribution of digital language materials in the Dutch language.

In this paper the collection of chats and tweets [14] and SMS [15] are described. The (collection of social media in the) SoNaR corpus serves a couple of purposes: the data collection will serve as a reference (e.g. for testing), researchers don't need to collect the data themselves. But the most important added value is that the data are accompanied by reliable metadata, i.e. gender, age and residency of the users, which are usually difficult to acquire for this type of data. Ensuring free availability was one of the main prerequisites of the SoNaR project. This presumes that all Intellectual Property Rights (IPR) are cleared to the fullest possible extent, turning the data collection process into a challenging task.

In the remainder of this paper we first describe other social media corpus compilation projects and their way of collecting

metadata in section 2. IPR issues are discussed in section 3. Then the collection of chats, tweets and SMS in the SoNaR corpus is described in section 4, followed by a description of the processing and anonymisation in section 5. In section 6, we finish with a discussion on using social media for linguistic research.

II. RELATED WORK

A. Chats

Chats are real time typed conversations over a computer network between two or more people. Internet chatting started in 1980 [12] and is can be done on many different platforms, like IRC, ICQ, MSN, Google chat, web based chatrooms, etc.

Several collections of chat data for (socio)linguistic studies have been carried out. The methods of collecting the data varies. The easiest way of collecting a corpus of chat data, is by joining one of the widely available online chat boxes with a programme that can keep a log of all data and so storing the chats. This way, no information about the identity of the chatters will be available. Reference [3] describes such a collection. Although she was not able to ask permission for from the participants, she decided to use the data anyway after considering the possible harm for the participants. The Navel Postgraduate School [5] collected a larger chat corpus from various chat services in accordance with the terms of privacy from the platforms. The corpus is distributed for non-commercial, non-profit educational and research use. A privacy marking was done by hand and the corpus is annotated with part-of-speech tags and dialogue-act tags. The data are from age-related chat rooms, but besides that, there are no metadata about the participants available.

Other linguistic researchers set up their own chat platforms in different ways, depending on the nature of their research questions, often resulting in project-related chat corpora. Reference [9] recruited 62 participants, mainly from the incoming first semester psychology students at the executive university, who held group discussions using ICQ computer chat in different conditions. A similar method was used by [1] to collect a Dutch chat corpus, called ChatIG, who collected chats from 114 secondary school students using the Blackboard Collaboration Tool (a digital learning environment system). The ChatIG data was included in the SoNaR corpus.

¹ <http://www.inl.nl/tst-centrale/>

B. Tweets

Tweets are messages published via twitter.com. Twitter started relatively recently; the first tweet dates from October 2006 [8]. Twitter is growing fast and has now 250 million tweets daily according to Twitter blog on 27 January 2012². Twitter provides an API³, which enables to collect Twitter messages with username, time and date stamp among other information. Although very easy to create with the API, publicly available corpora with tweets are scarce. Reference [18] compiled a corpus of 45,000 tweets after the announcement of Barak Obama's victory in the 2008 US presidential elections, in order to examine the evaluative language used to affiliate in tweets. The aim was not to construct a representative corpus and metadata about the twitterers is not available.

C. SMS

Since the first SMS service was offered to consumers in 1993 it has become one of the most widespread means of communication, especially among youngsters. In 2010, 6.1 trillion text messages were sent worldwide.⁴

In general, SMS corpora are scarce and the data are often not publicly available [2]. This is mostly because of the private character of SMS. Existing SMS corpora differ in size, language and collection method. Two notable SMS collection projects are the sms4science project⁵ and the NUS SMS Corpus Project⁶. Sms4science started in Belgium and over the years the same techniques have been carried out in other countries (Switzerland, France, Greece, Spain and Italy). A free of charge telephone number was opened to which participants could forward their text messages [4]. A relatively large budget was spend on the recruitment of participants through large-scale publicity through posters, newspapers and other types of media. At the National University of Singapore (NUS) English and Mandarin text messages have been collected. The corpus is collected by employing a combination of methods, to enable as many people as possible to contribute their messages. An application on the Google Android platform was developed, that allowed users to automatically send messages to the corpus [2]. Other possibilities of contribution included manually copying messages from the mobile device to an online website form or sending or instructions to generate an SMS back-up file from the mobile device and send it to the researchers. The same strategies were adopted for SoNaR.

III. IPR

A main consideration in creating a social media corpus is the need to protect the rights and interests of the authors and other persons mentioned in the text, while still preserving the original text and gathering sufficient metadata information.

For the SoNaR corpus, permission is needed for the data to be used by both commercial and non-commercial institutions.⁷

Roughly two ways of acquiring permission for incorporating social media into a language corpus can be distinguished. The first way of obtaining permission, is to contact the keeper of a large amount of data from different users. When the IPR of the data between the keeper and the users is settled, the data from different users can be obtained through a single license. In practice, it proved to be hard to obtain this permission, because e.g. chat room owners tend to protect the rights of their users more than to support the interests of the researcher.

The second way of acquiring permission is by contacting each individual owner of social media data. The advantage is that together with permission for usage and redistribution, also metadata such as gender, age and residency can be asked. A major disadvantage is of course the large effort that is needed to obtain and administrate the permission from all the individual owners of data.

IV. COLLECTION

A. Chats

In SoNaR, several ways to collect chat data were applied. Firstly, an existing Dutch chat corpus was included in SoNaR. The ChatIG corpus mentioned above was incorporated. The chat conversations in this collection were regulated and topics were provided. The chats were collected at a secondary school in Amsterdam in the school years 2004/05 and 2005/06. The parents of the pupils gave permission for the chats to be published. Metadata (sex, age, residence) of the pupils is available for the data of 2004/05 but not for the 2005/06 data. The size of this subcorpus is 83,806 word tokens.

Secondly, an IRC platform was set up to collect chat data. Two different sets of chat data were collected this way. The Bonhoeffer subcorpus was collected at a secondary school in April 2010. In each group of four students had two chat sessions with two persons participating and one chat session with four students. Each session lasted 10 minutes and topics for the chats were provided, although students were allowed to choose their own topic as well. The parents of the students gave permission for the chats to be published and metadata is available of all students. The size of this subcorpus is 27,936 word tokens. Colleagues from the language and speech (LandS) group at the Radboud University Nijmegen used the same set up that was used for the Bonhoeffer chats. Collections lasted from 8 December 2010 until 17 February 2012. Participants were asked daily by email to join the chat session after the coffee break in the morning. After entering the chatbox, a statement was shown that the data in the chatbox would be used in the SoNaR corpus and that by participating one would give permission hereto. In total the chatbox was used by 30 participants. From all users the usual

² <http://blog.twitter.com/2012/01/tweets-still-must-flow.html>

³ <http://api.twitter.com>

⁴ <http://www.itu.int/ITU-D/ict/material/FactsFigures2010.pdf>

⁵ <http://www.sms4science.org/?q=en>

⁶ <http://wing.comp.nus.edu.sg:8080/SMSCorpus/>

⁷ For some parts of the data, a different license is used, that cleared only non-commercial use. Moreover, the HLT centre is distributing the corpus for non-commercial purposes for free, but a payment is charged in case of commercial use.

metadata is available. The size of this subcorpus is 353,541 word tokens.

A third collection method was using the Microsoft Network (MSN) chats. Participants were recruited via a chain letter among friends and family to use MSN and send the chats. Participants filled in a form in which they gave permission for use of the data and to give their age, gender and residence. The size of this subcorpus is 272,237 word tokens.

The final subpart of chats, originates from a public chat platform. The Flemish website www.chat.be gave permission to use chats from their website. Chats were collected (using logging with the xchat programme) from 4 March 2011 until 11 February 2012. Users did not give permission individually and metadata of the users is not available. The size of this subcorpus is 11,135,664 word tokens.

Table 1. Percentages of the distribution of the 737,520 word tokens in the chat corpus from the Netherlands.

Age\Sex	M	F	?	Total
0-20	12.4	17.3	0.0	29.7
21-40	9.4	24.2	0.0	33.6
41-60	18.8	0.9	0.0	19.7
61-99	0.0	0.0	0.0	0.0
?	5.4	7.6	4.0	17.0
Total	46.0	50.0	4.0	100.0

B. Tweets

Tweets were collected using the Twitter API. Retweets were left out the collection and only tweets that are publicly available are collected. Two ways of getting metadata were used, dividing the tweets in SoNaR in two subcorpora; submitted and found. The submitted part contains tweets, resulting from a request for metadata of twitterers on Twitter, that caused a snowball effect resulting in retweets and news articles on the internet. Twitterers were asked to participate by submitting the name of their Twitter account and metadata to SonaR. The size of this subcorpus is 16,705,718 word tokens.

The second subcorpus contains tweets from Dutch and Flemish (semi-) celebrities that were searched on Twitter. On public websites such as Wikipedia and home- or fanpages the corresponding metadata was collected. The size of this subcorpus is 6,491,493 word tokens. Table 2 shows the percentages of the distribution in the corpus.

Table 2. Percentages of the distribution of the 23,197,211 word tokens in the Twitter corpus.

Age\Sex	M	F	?	Total
0-20	7.7	2.9	0.0	10.6
21-40	34.5	20.3	0.0	54.8
41-60	17.4	9.9	0.0	27.3
61-99	0.8	0.9	0.0	1.7
?	3.9	1.8	0.0	5.7
Total	64.3	35.7	0.0	100.0

C. SMS

For IPR reasons and to make sure metadata would be present, only messages that were sent (and not received) by the participant were included in the corpus. Contributors were recruited mainly by so-called *free publicity*: news articles on through the universities' communication offices and social media. The methods of SMS collection, were developed in cooperation with the NUS SMS team and provided the following options for contribution:

- Smartphone users, using the Android platform, could download an application that automatically uploads all sent SMS messages to their Gmail mailbox account. Afterwards, this list could be sent to the SoNaR SMS project;
- Apple iPhone⁸ and Nokia users could find instructions on the project website on how to obtain the SMS back-up file when connecting their phone to a computer;
- All other mobile phone users could fill in an online submission form and manually retype some text messages.

Putting metadata at disposal was not obligatory for smartphone users and other mobile phone users, but metadata is available for all but two contributors.

In total 52,913 SMS messages have been submitted by 272 contributors (147 Dutch, 125 Flemish). Dutch donators contributed 31,586 text messages (i.e. on average 215 SMS messages per contributor) and Flemish donators 21,327 (on average 171 SMS messages per contributor). The lower average number of Flemish contributors can be explained by a lower number of contributors using the Android app. More details about the distribution in the SMS corpus can be found in table 3.

Table 3. Percentages of the distribution of the 723,876 word tokens in the SMS corpus.

Age\Sex	M	F	?	Total
0-20	4.56	1.52	-	6.08
21-40	37.55	27.203	-	64.75
41-60	1.41	0.26	-	1.67
61-99	-	0.01	-	0.01
?	23.45	3.43	0.62	27.49
Total	66.97	32.41	0.62	100

V. PROCESSING AND ANONYMISATION

The data in SoNaR is divided in a part from the Netherlands and a part from Flanders. All tweets, chats and SMS were processed and incorporated in the appropriate part of the SoNaR corpus. All data is stored in the FoLiA format⁹, a xml-

⁸ Due to stricter security rules in the design of Apple software, building a similar App for iPhones proved to be much more complicated. Exporting the SMS messages from an iPhone is only possible after connecting it to a computer.

⁹ <http://ilk.uvt.nl/folia/>

format developed especially for linguistic resources. Every data file is accompanied by a metadata file in CMDI format¹⁰.

Metadata of the users is restricted to age, gender and place of residence or birth. The chats, tweets and SMS were tokenised by UCTO¹¹. The tokeniser was adapted for social media in such a way that it recognises e.g. emoticons and hashtags. In the remainder of this section, anonymisation procedures are discussed for each social media subcorpus.

A. Chats

All nick names in the data from the Netherlands are anonymised, both in the field that indicates the sender of the message, as in the messages themselves. No further anonymisation has been done of e.g. real names, addresses, telephone numbers. In an internal study carried out to investigate the possibilities of (automatic) anonymisation, this seemed not feasible. The data from Flanders has not been anonymised.

B. Tweets

The *Guidelines for Use of Tweets in Broadcast or Other Offline Media*¹² state that it is allowed to republish tweets, but only unchanged. Therefore there are no IPR issues and no anonymisation or alteration of the tweets was done.

C. SMS

In the SMS data, all phone numbers of the sender and (if present) the receiver of the messages were replaced by a unique identifier. Contributors were informed about the purpose of the SMS corpus and selecting appropriate text messages was left to the contributor's proper judgment. Contributors were free to remove messages before uploading them. They were given the instruction not to modify the messages, but we cannot guarantee that contributors did not change the content of the messages.

VI. LINGUISTIC RESEARCH OF SOCIAL MEDIA

Social media data can be seen as examples of new, widely used everyday computer-mediated communication. The study of social media is a contribution to our understanding of human social communication [6].

Companies offer more and more services via online chats and Twitter to their customers, supported by mobile devices, and social media is popular for marketing purposes as well. Understanding natural language use in social media is vital to understand and implement such services optimal in terms of efficiency and credibility. In this section a few examples are given of linguistic studies using social media.

The relationship between gender and chats has been investigated for many years. Reference [9] investigated how the new media affect communicative gender construction under different degrees of gender salience in German, using chat data. Other gender studies using chats and SMS are done in different languages. Reference [7] for example, studied

gender-related language use in SMS. Examples of linguistic research for tweets can be found in [16] and [14].

Applications of knowledge about text authorship on social media include forensic linguistic research, such as tracing pedophiles from chat rooms, a valuable application for criminal investigation [11]. Yet another relevant application of social media investigation is to enhance the efficiency of mobile technology. These studies are driven largely by attempts to improve the keyboards of mobile devices (refer to [15] for an overview).

VII. CONCLUSIONS

In this paper we discussed the construction of a social media corpus with chats, tweets and SMS in the framework of the SoNaR project. The promotion and collection methods varied slightly, but promotion principles such as using free publicity are proven useful in the collection of both tweets and SMS. Conclusion of an IPR license from the owner of a platform turned out to be difficult, but possible for some online chats. The collection of metadata rested upon consent from individual contributors, which made the collection of SMS and to a lesser extent also tweets and chats, relatively labour-intensive.

The availability of metadata and the permission of the participants for usage of social media data makes the presented corpus quite unique. The corpus can be an important resource for (socio)linguistic research for the Dutch language.

ACKNOWLEDGMENTS

The App for Google Android was translated and modified for Dutch in collaboration with Tao Chen, PhD student at NUS. The SMS collection in Flanders, Belgium was coordinated by Orphée DeClercq. We would like to thank everyone who contributed to our social media corpus.

REFERENCES

- [1] Charldorp, T. van (2005). *Building a Chat Corpus: ChatIG*, unpublished master thesis, Vrije Universiteit Amsterdam.
- [2] Chen, T. & Kan, M.-Y... *Creating a Live, Public Short Message Service Corpus: The NUS SMS Corpus*. In *Language Resources and Evaluation journal*, unpublished.
- [3] Del-Teso-Craviotto, M. (2008). Gender and sexual identity authentication in language use: the case of chat rooms. *Discourse Studies* 10 (2), 251-270
- [4] Fairon, C. & Paumier, S. (2006). *A translated corpus of 30,000 French SMS*. Proceedings of LREC 2006. Geneva, Switzerland.
- [5] Forsyth, E.N. & Martell, C.H. (2007). *Lexical and Discourse Analysis of Online Chat Dialog*. Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007), pp. 19-26
- [6] Hancock, J. T. & Dunham, P.J. (2001). Language use in computer-mediated communication: The role of coordination devices. *Discourse Processes* 31: 91-110.
- [7] Herring, S.C. & Zelenkauskaitė, A., (2009). Symbolic Capital in a Virtual Heterosexual Market : Abbreviation and insertion in Italian iTV SMS. *Written communication*, 26(1), 5-31.
- [8] Java, A., Song, X., Finin, T., Tseng, B.. (2007) *Why we twitter: understanding microblogging usage and communities*. Proceedings of

¹⁰ <http://www.clarin.eu/cmdi>

¹¹ <http://ilk.uvt.nl/ucto/>

¹² <https://support.twitter.com/entries/114233>

the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis.

- [9] Koch, S., Mueller, B., Kruse, L. & Zumbach, J. (2005). *Constructing Gender in Chat Groups*. Sex Roles: A Journal of Research, 53(1,2), 29-41.
- [10] Oostdijk, N., Reynaert, M. Monachesi, P., Noord, G. van, Ordelman, R. Schuurman, I. & Vandeghinste, V., (2008). *From D-Coi to SoNaR: A reference corpus for Dutch*. Proceedings of LREC 2008, Marrakech, Morocco.
- [11] Peersman, C., Daelemans, W. & VanVaerenberg, L. (2011). Predicting Age and Gender in Online Social Networks. 3rd International Workshop on Search and Mining User-generated Contents. SMUC.
- [12] Reynaert, M., Schuurman, I., Hoste, V. and Oostdijk, N. (2012) *Beyond SoNaR: towards the facilitation of large corpus building efforts*, Proceedings of LREC 2012 Istanbul, Turkey
- [13] Samarajiva, R., Shields, P. (1997) Telecommunication networks as social space: implications for research and policy and an exemplar. Media Culture Society 19: 535
- [14] Sanders, E., (2012). Collecting and Analysing Chats and Tweets in SoNaR, Proceedings of LREC, in press.
- [15] Tagg, C., (2009). *A corpus linguistics study of SMS text messaging*. Ph.D. thesis, University of Birmingham.
- [16] Tjong Kim Sang, E. (2011). *How to Use Twitter for Linguistic Research*. Tabu, 39(1,2), 62-71 (In Dutch).
- [17] Treurniet, M., De Clercq, O., Oostdijk, N., Heuvel, H. van den, (2012) *Collecting a Corpus of Dutch SMS*. Proceedings of LREC, in press.
- [18] Zappavigna, M. (2011). Ambient affiliation: A linguistic perspective on Twitter. New Media & Society 13(5). 788-806.