Collecting and Analysing Chats and Tweets in SoNaR

Eric Sanders
CLST, Radboud University Nijmegen
E-mail: e.sanders@let.ru.nl

Abstract

In this paper a collection of chats and tweets from the Netherlands and Flanders is described. The chats and tweets are part of the freely available SoNaR corpus, a 500 million word text corpus of the Dutch language. Recruitment, metadata, anonymisation and IPR issues are discussed. To illustrate the difference of language use between the various text types and other parameters (like gender and age) simple text analysis in the form of unigram frequency lists is carried out. Furthermore a website is presented with which users can retrieve their own frequency lists.

Keywords: SoNaR, social media, chats, tweets, corpus collection, corpus analysis

1. Introduction

The computational linguistics community in the Netherlands and Belgium has built a major reference corpus of written Dutch called SoNaR (Oostdijk et al forthcoming, Reynard et al, 2012) which will expectedly boost natural language processing involving the Dutch language. The corpus, was finished 1 March 2012 and contains 500 million words from a variety of text sources, e.g. books, newspapers, manuals. Sources also include so called social media, like online chats, internet fora, blogs, twitter and text message (SMS, Treurniet et al, 2012).

In this paper the collection of online chats (henceforth: chats) and tweets are described, with a focus on the way the data was collected, IPR issues and anonymisation of the data. To illustrate an analysis of the sort of language that chats and tweets contain, some simple statistics of word use in the chats and tweets that were collected will be presented.

Both chats and tweets are relatively easy to collect by researchers. Reasons for including them in the SoNaR corpus are that the data collection can serve as a reference, it takes away the need for researchers to collect the data themselves, and it may be used as a starting point for further collections. The most important reason though 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.

2. SoNaR

The data in SoNaR is divided in a part from the Netherlands and a part from Flanders. For the tweets there is no difference in collection of the two sources, but for the chats the data is quite different in nature as will be explained below.

All data is stored in the FoLiA format[1], a xml-format developed especially for linguistic resources. Every data file is accompanied by a metadata file in CMDI format[2]. Metadata of the users is restricted to age, gender and place of residence or birth.

The chats and tweets were tokenised by UCTO[3]. The tokeniser was adapted for social media in such a way that it recognises e.g. emoticons and hashtags.

The chats and tweets will be made available together with the rest of SoNaR via the TST-centrale[4].

3. Chats

Chats are real time typed conversations over a computer network between two or more people. Internet chatting dates from 1980 (Samarajiva er al, 1997) and is practiced on many different platforms, like IRC, ICQ, MSN, Google chat, web based chatrooms, etc.

The Flemish part of the data comes from one large open chat channel. No metadata is available and anonymisation was not done. The part from the Netherlands exists of data from not publicly accessible chat sessions, especially set up for data collection, from four different sources described below.

The chats are stored as one session per file. For a few sources there is no distinction between different sessions. In that case the data was split up in one day per file with the transition to a new day at 4 o'clock in the morning, since this was expected to be the time with the least conversations.

Three possible events concerning chats are stored:
1) a user enters a message (concluded by pressing the send button or typing the enter key)
2) a user joins or leaves a chat room
3) a user changes his/her nick name

In general each event is represented with a date and time stamp, the nick name of the user and -in case of a message- the content of the message. Of some sources, date and time stamps are not available or imprecise. Also information about joining/leaving the chat room is not always available.

Although a nick name in chats seldom reflects the user's real identity, 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.

[1] http://ilk.uvt.nl/folia/
[2] http://www.clarin.eu/cmdi
[3] http://ilk.uvt.nl/ucto/
[4] http://www.inl.nl/tst-centrale/

IPR issues differ per subcorpus and are discussed in the overview of the different subcorpora below.

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

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

3.1 ChatIG

The ChatIG corpus (Charldorp, 2005) is a chat corpus of Dutch teenagers. It was collected before the existence of SoNaR by the VU Amsterdam in two school years: 2004/05 and 2005/06. Different classes from secondary schools in Amsterdam came to the VU to chat via the chat tool in Blackboard (a digital learning environment system). The chat conversations were regulated and topics were provided. Chat session were in groups consisting of boys only, girls only, or mixed.
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.

3.2 Bonhoeffer

This subcorpus was collected at a secondary school (named Bonhoeffer) in Enschede, in April 2010. In cooperation with the teachers, a class of students was divided in groups of four students, who had chat conversations in a chatbox, set up for SoNaR (an inspirdcd server on a linux system and a Mibbit webclient). In each group, there were 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.

3.3 LandS

Colleagues from the language and speech (LandS) group at the Radboud University Nijmegen used the same chat system that was set up for the Bonhoeffer chats. Collections lasted from 8 December 2010 until 17 February 2012. A reminder e-mail was sent to the participants during this period on each workday to chat after the coffee break in the morning. After entering the chatbox, a statement was shown that the data in this 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 metadata is available
The size of this subcorpus is 353.541 word tokens.

3.4 MSN

In the framework of the NEWSPEAK project[5] MSN chats were collected. The collection of the data took place between October 2009 and April 2010.
Recruitment was organised through a chain letter sent to friends and family. All participants signed a form in which they gave permission to use the data for scientific research and gave sociolinguistic information, like age, gender and region of birth.
The size of this subcorpus is 272.237 word tokens.

3.5 chat.be

The Flemish website www.chat.be gave permission to use chats from their website. Chats were (not continuously) collected (using xchat logging) from 4 March 2011 until 11 February 2012. The chats are from the main chat channel of the site (named chat.be). Participants did not give permission individually and no metadata of the participants is available. No anonymisation of the nick names or data has been carried out.
The size of this subcorpus is 11.135.664 word tokens.

4. Tweets

Tweets are messages published via twitter.com. Twitter is much younger than chat and dates from October 2006 (Java et al, 2007) and is by far the most popular micro blog with 250 million tweets per day according to Twitter blog on 27 January 2012[6].
The Twitter API[7] was used to collect the tweets for the corpus. Retweets were not collected. Of each tweet, the twitterer, date and time stamp and the message ('tweet') is stored. Tweets of one twitterer are stored in one file. Of each twitterer gender, age and residence or birth place are available in the corpus.
Only tweets that are publicly available are collected. The *Guidelines for Use of Tweets in Broadcast or Other Offline Media*[8] 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.
The tweets in SoNaR are divided over two subcorpora.

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

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

4.1 Submitted

The first subcorpus contains mainly tweets from the Netherlands. A tweet about the twitter collection in

[5] http://www.ru.nl/cls/events_news/news/@754375/knaw-subsidie_voor/
[6] http://blog.twitter.com/2012/01/tweets-still-must-flow.html
[7] http://api.twitter.com
[8] https://support.twitter.com/entries/114233

SoNaR with a request for metadata of twitterers caused a snowball effect. Attention for the tweet collection was spread over twitter via retweets and also a national news website and the Radboud University's homepage reported on the data collection. Twitterers were asked to participate and to submit the name of their Twitter account and metadata to SoNaR, either by e-mail of via a webform set up for this purpose.

The size of this subcorpus is 16.705.718 word tokens.

4.2 Found

The second subcorpus contains tweets from Dutch and Flemish (semi-) celebrities. Politicians, actors, sports people and other public figures using Twitter were searched. 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.

5. Text Analysis

The type of language used in social media is quite different from conventional texts. Chats are real time (spontaneous) conversations, causing less time to think and type for the users, which leads to more spelling errors, abbreviations, incomplete and ungrammatical sentences. Tweets are restricted to 140 characters, which also leads to more abbreviations and incomplete and ungrammatical sentences. For chats and tweets, as for social media in general, special lingo has evolved. Emoticons (smiley's) and special abbreviations are used that will seldom be found in printed texts.

Research on language used in chats and tweets is increasing (see e.g. Erik Tjong Kim Sang, 2011). The collection of chats and tweets in SoNaR is very suitable for this purpose. The metadata makes it possible to find differences between subcategories. A website (http://wwwlands2.let.kun.nl/sonar/) was created with which the user can retrieve frequency lists of the number of tokens from subcategories. The user can select parameters for gender (male, female, unknown), agegroup (0-20, 21-40, 41-60, 61-99, unknown), word type (word, emoticon, hashtag, other) and subcorpus (as described above). Any combination of parameters is possible. Also, the user has the possibility to do a case insensitive search and can choose the number of words to be returned (10, 25, 100 or all). The frequency lists that are returned are limited to unigrams, but they can serve as a starting point for further text analysis. Because there are no metadata with the Flemish tweets, they are left out of this data set.

Below a few examples of analysis of differences between subcategories of the chats and tweets are given.

5.1 Words by gender

In table 3 the most frequent words used by men and women are shown.

The words in the top-10s are either articles (*de, het, een*), prepositions (*in, van*), conjunctions (*en, dat*), a personal pronoun (*ik*), a negation (*niet*) or a conjugation of 'to be' (*is*).

The words in the top-10 are the same for men and women. The order is a bit different, but the percentages are similar.

	Men		Women	
Rank	Word	%	Word	%
1	de	2.64	ik	2.44
2	ik	1.97	de	2.44
3	het	1.71	en	1.78
4	een	1.70	het	1.70
5	en	1.60	een	1.63
6	in	1.56	je	1.48
7	is	1.53	in	1.48
8	van	1.48	is	1.45
9	je	1.35	dat	1.38
10	dat	1.26	van	1.35

Table 3: Top 10 of most frequent words from men and women of all ages in all chats and tweets.

The biggest difference is that women use the word 'ik' (I) more often (2.44% against 1.97% for men). There could be several reasons for this: women use more complete sentences and delete the word 'ik' less or maybe women talk more about themselves or women use more personal statements ('I find that') compared to men ('It is so'). Further investigation using more context is needed to find the precise reason.

5.2 Hashtags by country

Hashtags are typical for Twitter. They are used to mark keywords or topics of the tweet it appears in. It was created organically by Twitter users as a way to categorise messages[9]. Several third party websites are dedicated to the use[10] or meaning[11] of hashtags.

Table 4 shows the most popular hashtags for users from the Netherlands and from Flanders.

	Netherlands		Flanders	
Rank	Hashtag	%	Hashtag	%
1	#durftevragen	1.44	#dcln	1.37
2	#ff	0.79	#vivelevelo	0.91
3	#fail	0.61	#terzaketv	0.89
4	#dtv	0.58	#fb	0.85
5	#fb	0.53	#sbbvgc	0.83
6	#penw	0.52	#sporza	0.82
7	#twexit	0.45	#durftevragen	0.78
8	#dwdd	0.36	#in	0.72
9	#lastfm	0.31	#nogov	0.63
10	#vvd	0.30	#tdf	0.60

Table 4: Top 10 of most frequent hashtags in the Netherlands and Flanders.

[9] http://support.twitter.com/articles/49309-what-are-hashtags-symbols

[10] http://hashtags.org

[11] http://tagdef.com

Only a few hashtags are both in the top-10 of the Netherlands and Flanders. Very popular in the Netherlands is '#durftevragen' (meaning #daretoask) or its abbreviation '#dtv', which is the 7th most popular hashtag in Flanders. Also '#fb' (for facebook) is in both top-10. The international hashtags '#ff', '#fail' and '#twexit' are only in the top-10 of the Netherlands. Many hashtags refer to a television or radio programme, like '#penw' and '#dwdd' in the Netherlands and '#dcln', '#vivelevelo' (about the tour de France #tdf), '#terzaketv' and '#sporza' in Flanders. The '#nogov' (no government) refers to the long formation period that took place in Flanders. Note that hashtags are very sensitive to trends. Using data from another period might result in a very different top-10.

5.3 Emoticon by age

Emoticons are used to express the mood of the user and are used a lot in social media. The majority of the emoticons consist of eyes, a mouth and sometimes a nose, in many different forms[12]. Table 4 shows the most used emoticons for different age categories.

	0-20		21-40		41-60	
Rank	Emo	%	Emo	%	Emo	%
1	:)	20.01	:)	27.96	;-)	30.21
2	:P	18.97	;)	18.95	:-)	28.59
3	:D	10.83	;-)	18.61	:)	19.39
4	;)	9.91	:-)	12.78	;)	7.24
5	:p	7.81	:D	5.98	:-))	4.33

Table 5: Top 5 of most frequent emoticons for 0-20, 21-40, 41-60

Young people often use a P or D for the mouth reflecting the tongue sticking out, whereas older people seldom use these. The opposite goes for the use of the nose. Luckily most used emoticons are happy smiley's :).

6. Conclusions and Future Work

The collection of chats and tweets in the framework of the SoNaR corpus is described in this paper. Getting large amounts of chat or twitter data is not very difficult, but the extra value of the SoNaR data is in the metadata. This can be used in comparing use of language between different categories.

A website was presented with which users can retrieve frequency lists of the data for different (combinations of) categories.

Some examples of comparisons using this website are given, that can use as starting point for further research. This could include using context (e.g. multigrams) into account or more detailed linguistic information like part of speech tagging, which is currently not available for the social media in SoNaR.(e.g. multigrams) into account or more detailed linguistic information like part of speech tagging, which is currently not available for the social media in SoNaR.

7. Acknowledgements

This research was funded by the STEVIN programme under grant number STE07014.

8. References

Charldorp, T. van. (2005) *Building a chat corpus: ChatIG.* Master thesis, Vrije Universiteit Amsterdam

Java, A., Song, X., Finin, T., Tseng, B.. (2007) *Why we twitter: understanding microblogging usage and communities.* Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis

Oostdijk, N., Reynaert, M., Hoste, V., Schuurman, I., (forthcoming). *The construction of a 500-million-word reference corpus of contemporary written Dutch.* In *Essential Speech and Language Technology for Dutch: resources, tools and applications*. Springer, Verlag.

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

Samarajiva, R., Shields, P.. (1997) *Telecommunication networks as social space: implications for research and policy and an exemplar.* Media Culture Society 19: 535

Tjong Kim Sang, E. (2011) *Het gebruik van Twitter voor Taalkundig Onderzoek* TABU: Bulletin voor Taalwetenschap, volume 39, number 1/2, pages 62-72 (in Dutch)

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

[12] http://en.wikipedia.org/wiki/List_of_emoticons