Wilbert Spooren, Tessa van Charldorp

Challenges and experiences in collecting a chat corpus

Abstract

Present day access to a wealth of electronically available linguistic data creates enormous opportunities for cutting edge research questions and analyses. Computer-mediated communication (CMC) data are specifically interesting, for example because the multimodal character of new media puts our ideas about discourse issues like coherence to the test. At the same time CMC data are ephemeral, because of rapid changing technology. That is why we urgently need to collect CMC discourse data before the technology becomes obsolete. This paper describes a number of challenges we encountered when collecting a chat corpus with data from secondary school children in Amsterdam. These challenges are various in nature: logistic, ethical and technological.

1 Introduction

Present day access to a wealth of electronically available linguistic data creates enormous opportunities for cutting edge research questions and analyses. The data used in such analyses often come in systematically collected collections of texts that are, in one way or the other, representative of the population of discourses from which they are taken. Computer-mediated communication (CMC) data are specifically interesting for a number of reasons (cf. Herring 2004; 2013 for an extensive discussion of possibilities and pitfalls of so-called Computer-Mediated Discourse Analysis, or CMDA). A much-discussed issue is how classical distinctions like that between written and oral data, as discussed for example in Chafe (1982), break up in CMC, as witnessed by Baron's (2008) analyses of email and instant messages. For example, the conceptualization of how coherence is established in CMC is based on language use and routines from written language (such as coherence relations and their markers) as well as from spoken language (such as adjacency pairs and other interactional 'devices') (cf. Sanders and Spooren 2013). Similarly, technological advancements allow users to express multimodal content in ways not imagined two decades ago. This type of multimodality creates new forms of communication that may not be similar to the traditional ways in which we 'write' and 'speak.'

The rapid technological developments bring into existence both new media and new genres that are sometimes short-lived. These developments make the analysis of CMC a discipline in which we operate in a permanent laboratory for the study of genre and the role of language in it. At the same time, the rise and fall of technologies and genres like Second Life gaming and MSN chat show that it is imperative to collect these materials for scientific study before they become obsolete. A large and varied collection of CMC material allows us to answer questions not only about language change across generations or language use across various CMC modes, but also about the creativity and adaptation of language use in technologically advanced or restricted environments (for example, see Herring 2004). Furthermore, sociolinguistic questions about age, gender and technological experience in relation to language use can be explored when such data are available alongside the corpus. Once we have collected this type of discourse systematically it allows us to study the use and construction of computer-mediated language use over time and throughout various types of communication.

Before corpus linguistic analysis can be done, a corpus first needs to be created and annotated. High quality written and spoken corpora have been available for decades, and within the field of corpus linguistics standardization of corpus annotation has been the main objective of the last decade or so. However, CMC is a relatively new phenomenon, and despite its many interesting research opportunities, relatively few CMC corpora are available, especially for Dutch (cf. Oostdijk et al. 2013, for a recent corpus comprising some forms of CMC data, among which the sub corpus described in this paper).

After providing a short overview on past corpus creation literature and the technicalities of our corpus, we present the challenges we met when we put together a chat corpus. The descriptions are intended as a very practical account of what we encountered when building a chat corpus in 2004-2006.

2 Literature

The ultimate goal of any corpus research is to have naturally occurring data available in which phenomena of interest can be found to a degree that is representative of the way these data occur in the type of language use that the researcher is interested in. For quantitative corpus studies this means that the occurrence of the phenomena of interest resembles that of the population. This will only occur if the corpus is based on random sampling and if the corpus is balanced, i.e., the size of the sub registers in the corpus reflects that of the language use the researcher is interested in (cf. Gries and Newman 2013, and the references cited there). For qualitative corpus analysis representativeness means that all the varieties of language use that the researcher is interested in should occur in the corpus. It also means that we should have as much data available about the context of language use as possible. Both types of requirements pose serious challenges to building a CMC corpus. It is generally acknowledged that these requirements are ideals, and that in actual practice corpus builders often deviate from them, partially due to pragmatic reasons (also see pragmatic challenges when building an SMS corpus as explained by Tagg 2009).

CMC corpora to date – for example, the Deutsches Referenzkorpus zur internetbasierten Kommunikation (Beißwenger et al. 2013), the Dortmund Chat Corpus (Beißwenger 2013), the Queer Chat-Room corpus (King 2009), or the Netlog Corpus (Kestemont et al. 2011) – scarcely describe the challenges they met when collecting CMC materials. Most authors focus on the next phase: format and annotation challenges when building corpora (cf. Beißwenger and Storrer 2008, section 3 as an example). Ethical issues are mentioned in passing, but authors are quick to suggest that it is “unrealistic to obtain a declaration of consent for the recording and subsequent use of users’ statements for research purposes.” (Beißwenger and Storrer 2008, section 3.1.8). This is especially the case when CMC corpora are based on publicly available communication on the internet, for example public chatrooms or public fora.

This article elaborates on the often overlooked phase of corpus creation: collecting data. We describe the challenges we met when building a CMC corpus and demonstrate that CMC material can be collected with consent. We made decisions that balance between the ideal requirements for a representative corpus and the pragmatic reality. Before setting out the seven challenges, a short overview of the corpus characteristics is given.

3 The ChatIG corpus

3.1 Background

The corpus project was part of the “Vrolijke School” [“Happy School”] project – a collaboration between VU University Amsterdam and Ignatius Gymnasium Amsterdam (an Amsterdam-based grammar school) – aimed at creating awareness of the pleasures and complexities of research amongst secondary school students. In this project various faculties of VU University Amsterdam formulated subprojects for collaboration. One such subproject was ChatIG, a collaboration between the Faculty of Arts and the Ignatius Gymnasium. This project had the objective to 1) let the pupils build a corpus of chat data; 2) make the corpus available for scientific research; and 3) formulate and answer research questions based on the corpus.

3.2 Participants

In 2004-2005 four classes of Ignatius pupils participated, two classes from grade 1 (age 12/13) and two classes from grade 3 (age 14/15). In 2005-2006 three classes participated, all 3rd grade students (age 14/15). In total 188 pupils participated. The pupils were supervised by their own Dutch language teachers as well as both authors. Technical assistance was provided by the university.

3.3 The chat experiment

The pupils chatted with each other through the chat program within Blackboard, an online environment used in various schools and universities for handing in papers, storing files, emailing, having forum discussions, etc. During the chat experiment, the pupils participated in seven short chat sessions of five minutes each. The first session was a practice session to get used to the chat environment. The other six sessions were devoted to various topics, two of which were deemed ‘involved’ (i.e., it was expected that the pupils would experience involvement with the topic, e.g. the MTV awards which were recently aired on television), two of which were expected to be ‘non-involved’ (e.g. the election of a new pope), and two sessions were free choice topics. The experiment was also set up in such a way that each participant took part at least once in a session with two, three and four pupils. There was no moderator present and students did not have the option to integrate other types of media into the chat sessions (i.e. whiteboards, material on external platforms, movies, etc.).

3.4 Metadata

Apart from partaking in various chat sessions, all pupils filled in a questionnaire, which allowed the authors to collect metadata on the pupils’ gender, background of their parents, languages spoken, computer and chat usage, circle of close friends, etcetera. Both pupils and parents filled in a consent form allowing the authors to use the collected materials, both the chat output and the questionnaires, for scientific research.

3.5 The final corpus

Once the pupils had participated in the chat sessions, we had a large collection of chat interactions as archived by the Blackboard chat system. No messages were adapted or deleted. All the sessions were exported to the NoteTab Light editor and the data were cleaned up manually.

We decided that time tags after each chat contribution would not help analyses since the system was quite slow. We therefore removed the time tags, with the exception of the time tags that showed when individuals entered and left the chat session. Furthermore we removed all irrelevant information related to archiving the sessions. We also replaced the pupil information at the beginning of each submission with a more detailed line of information. In the final corpus this information has been replaced with the following string of information: 'unique pupil ID_male or female + schoolyear_pupil ID throughout chat sessions.' A new ID will look like this: '64_m1_20', meaning that this pupil has received 64 as their unique ID in the Meta Database, the pupil is male, a first grader, and used number 20 throughout the chat sessions.

The total corpus has been incorporated in the SoNaR corpus (Oostdijk et. al. 2013) in order to make it accessible for linguistic research. Here the corpus has been standardized and annotated using the SoNaR standard (cf. Sanders 2012 for details and references). The size of the ChatIG corpus is 83,806 tokens (Sanders 2012). The corpus is available through VU University or via TST Centrale (http://tst-centrale.org).

4 Challenges

4.1 Challenge 1: Finding chat participants

Since chat communication generally takes place in a private setting (whether this is in a chat room or through a chat messenger program), we required participants who would chat in similar, private conditions. At the same time we wanted to store the chat logs. Rather than visiting people in their homes while chatting or asking people to store their own private chat logs from their own computers we decided to control the chat sessions and set up a chat experiment. This way we could also be sure that the chat data were unedited. Since chat communication in 2004 was most popular amongst Dutch youth (0-24 year olds), we wanted to have access to this audience. A school seemed like the best way to reach Dutch teenagers. We found a school to collaborate with through the "Happy School" project.

Creating the chat corpus therefore depended very much on the collaboration with this Amsterdam grammar school. On the one hand this allowed us to get a unique set of data, on the other hand this created a number of practical issues that are difficult to deal with and may well hinder the creation of a large scale corpus of this type. First of all, it is imperative that the pupils are capable of making their contribution. In our case this required that appointments were made with the teachers of the various classes, so that the pupils could come to the university and chat in our computer rooms where we had the required hardware and software. In order to make this work this required very precise arrangements with the school, who are ultimately responsible for the pupils (for example, the pupils' use of the tram, checking the presence of all of the pupils, etcetera).

The benefits of working together with the school are that the chat sessions were part of the regular school program. Pupils were therefore required to attend and participate. Furthermore, the pupils were always supervised by their own teacher and a lot of the practical matters were dealt with by the teacher. However, even though the chat experiment was made part of the regular school lessons, one cannot make partaking in a university experiment a requirement for passing a Dutch class at grammar school. One cannot force anyone to partake in an experiment, especially not minors. It is therefore essential to have a good relationship with the partner school and to have consent from all relevant parties.

4.2 Challenge 2: Privacy and consent issues

The pupils' chat contributions needed to be available to the scientific community. In principle there are two ways to make these data available. The first is to use the opt-out strategy that is used for example by Google Books: the data are available unless participants explicitly request that their data be withdrawn from the corpus. We consider this option unethical, in that we feel that participants should be aware of the fact that their data are the object of research, not post hoc, but before the fact. The second option is to ask the participants beforehand for consent to participate in the research. Lewis (2002) distinguishes between consent (there is formal approval for participation) and assent (the participant is willing to participate). In the case of adult participants these two forms of approval usually coincide. In the case of pupils from the Gymnasium age this requires approval from the caretaker (usually the parent). That is why we asked both parents and pupils to sign an informed-consent form beforehand.

Informed consent implies that participants know beforehand that their data will be recorded. This is what we believe to be the most ethically responsible way to collect research data. Although this option is ethically transparent, it does cause participants to possibly behave differently than they would under 'natural' circumstances. Our corpus data show that the participants are very much aware that their data are being read by the researchers.

(1) 6 leerling: alles wordt gescreendt
6 pupil: everything is being screened

At the same time, the students treat this fact as a joke:

(2) 15 leerling: zouden ze al deze gesprekken kunnen nalezen
14 leerling: ja
11 leerling: Tuurlijk
15 leerling: shit
11 leerling: Ach
15 leerling: pas op met wat je zegt
15 leerling: !!!
11 leerling: Whahaha

15 pupil: will they be able to re-read all these conversations?
14 pupil: yes

11 pupil: Of course
15 pupil: shit
11 pupil: Owh
15 pupil: be careful with what you say
15 pupil: !!!
11 pupil: whahaha

Just like participants who are being recorded with a tape recorder forget within a few minutes that the tape recorder is actually present in the room (ten Have and Komter 1982), these pupils also quickly forget that their chat conversations are being recorded. This becomes apparent when pupils are talking about smoking drugs, a topic that will most likely not be appreciated by their school teachers. Even when one of the pupils reminds the others that the conversation is being recorded, the pupils continue talking about drugs.

(3) 18 leerling: pam[1] heeft kk veel geblowt dit weekend ehh
17 leerling: haha
18 leerling: gwn bij der huis, kapot gek
17 leerling: ja?
16 leerling: ehm.. dit wordt opgenome he.. dat jullie da ff wete\
17 leerling: heb ik ook wel is gedaan
18 leerling: ja ze is para tog
17 leerling: ma toen waren mn ouders weekend weg

*18 pupil: pam [name] f*ing smoked up a lot this weekend ehh*
17 pupil: haha
18 pupil: jst at her house, mad crazy
17 pupil: yeah?
16 pupil: ehm.. this will be recorded.. just so you guys know
17 pupil: I've done that before
18 pupil: yeah, she's para right
17 pupil: but my parents were away for the weekend

This example also shows that we need to be very careful with anonymizing the data. Pupils at this age cannot be held accountable for their actions or language use at a later stage in life. Although they do all sign a form which states that their data can be used for research, they will most likely not be aware of the implications of their talk about drugs in future circumstances. It is for this reason that the data were anonymized in all publications. Last names were never recorded or required.

4.3 Challenge 3: Technological challenges

For our data collection we made use of the Blackboard 6.2 and later 6.5.1 (full participation mode) chat system. This text-based tool is especially designed for live, synchronous interaction. This provided the opportunity to archive the data and to prevent the pupils from chat-

[1] Names in the examples have been changed.

ting with people outside of the chatroom. We also turned off the private-message setting so that students could not message each other privately, outside of the experimental setup. At the same time the use of this technology proved to be challenging. The interface was highly unfamiliar for the pupils, which we accommodated for by letting the students familiarize themselves with the system during the first five minute session. The interface was also relatively basic, compared to that of MicroSoft Network messaging service program (MSN), which was the dominating chat program at the time of the data collection. For example, the pupils could not make graphical smileys and the like in the way they were used to.

Another technical issue had to do with the sizes of the groups of pupils coming to the university. As our computer rooms cater for 18 participants the groups had to be divided over two rooms, on different floors. The synchronization between the two classrooms was not optimal during the first chat session. This meant that some pupils were waiting for their chat partner to arrive in the chat room. Especially when the pupil only had one chat partner, this meant that pupils were annoyed or waiting for too long, missing out on minutes in which they could have been chatting.

The choice for using the Blackboard system provided us with maximal control over the situation: pupils could chat only with fellow pupils of a known age and background who had also consented to participate in the data collection. Drawbacks to this approach were already mentioned above (unfamiliar interface, lack of speed, etcetera). In a later chat data collection for the benefit of the SoNaR project a dedicated tool was used to collect chat data from similar age groups (cf. Sanders 2012 on the Bonhoeffer data). In general such a dedicated tool seems preferable, but of course it implies the availability of the relevant technology.

4.4 Challenge 4: Imitating a natural chat situation

By choosing to have pupils take part in a controlled chat experiment, we also chose to work in an unnatural chat situation. The university computer facilities as well as the number of supervisors involved in the experiment allowed us to use two different classrooms. However, pupils would sometimes chat with people in the other classroom and sometimes with people in the same classroom.

In order to maintain the experimental set up but at the same time imitate a home situation, individual cabinets might offer a better solution. At the same time, it will remain difficult to resemble the home situation (van Charldorp 2005). In a home situation pupils might be listening to the radio, watching TV, be involved in a face-to-face conversation or otherwise be multitasking. Nowadays, pupils may be chatting or apping on their mobile phones and may even be on the road, at school or in the bus while chatting online with friends.

The speed of the internet connection and the chat program also contributed to the unnaturalness of the chat situation. At the time of creating this corpus, pupils were used to MSN which operates not only speedier, but also has a different interface. The speed of the Blackboard chat system was a frequently discussed topic:

(4) 12 leerling: sssssssssssssssssslllllllllllllllllllllllloooooooooooooooooooooooooooooooooooooooooooooooooooooooooooommmmmmmmmmmmmmmmmmmmmmmmmmmmmmmm

12 student: sssssssssssssssssslllllllllllllllllllllllllllooooooooooooooooooooooooooooooooooooooooooooooooooooooowwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwww

The Blackboard chat system is also specifically compared to MSN in the example below:

(5) 11 leerling: dit is vaag man, en sloom
17 leerling: isset leuk daaro?
11 leerling: jaah, man
17 leerling: jah kweet
17 leerling: msn is btr
11 leerling: waatttuh\, neej, tis saai maar beter dan sgool
17 leerling: dat zei ik dus egt een ur gelee
17 leerling: ja keej
17 leerling: maar kan neit tege sloomheid
11 leerling: jaaah,
17 leerling: van de comp

11 student: this is vague man, and slow
17 student: isit fun o there?
11 student: yeah, man
17 student: yah I know
17 student: msn is btr
11 student: whaaaat\, naah, its boring but better than scool
17 student: that's what I said like an hour ago
17 student: ya ok
17 student: but I can't take slowness
11 student: yeaaah,
17 student: of the comp

In the MSN chat program you can see who you are chatting with. Since we decided to anonymize all pupils by assigning numbers (pupil 1, etc.), pupils' first question would usually be: who are you?

(6) 16 leerling: who r u???
16 pupil: who r u???

Or, in the example below where there are more than two chat participants, pupil 10 asks the others present who pupil 16 is.

(7) 10 leerling: wies 16
10 leerling: ?????????

10 pupil: whos 16
10 pupil: ?????????

These examples show that identity is generally known beforehand when chatting with each other. This is an important issue given that identity management may crucially determine chat communication (Becker and Stamp 2005).

Soon the pupils also started playing with this anonymity issue. Pupils could for example pretend to be someone else, like in the example below:

(8) 1 leerling: he ik ben ook sacha
7 leerling: das raar

1 pupil: hey I'm also sacha
7 pupil: thats weird

This type of activity then shows that pupils quickly become aware of the technological restrictions but also opportunities that it provides for interaction.

A similar activity occurred when pupils realized that emoticons did not work the same way in the Blackboard chat program as they did in the chat programs they were used to at home.

(9) 23 leerling: wij zijn annaaa(8)
19 leerling: anna is sexy naam (HAHAHAHA)
23 leerling: oh.. geen emoticons =_=

23 pupil: we are annaaa(8)
19 pupil: anna is sexy name (HAHAHAHA)
23 pupil: oh.. no emoticons =_=

In the first line, pupil 23 uses an emoticon: (8), which should create an emoticon with sunglasses. However, the Blackboard system does not create emoticons based on characters, and thus the pupils just see the (8) on their screen. The pupil realizes that her emoticon did not work and shares her disappointment: "oh.. no emoticons" and adds the characters that create a 'bored face' emoticon. Again, the use of characters to display an emoticon, while knowing that the emoticon will not work in this chat program, shows the ability of the pupils to adapt their language use to the technological restrictions. However, students still complained about the lack of emoticons and buzzers, not only in the evaluation of the chat sessions afterwards, but also to each other in the chat sessions themselves:

(10) 8 leerling: echt rot dat hier gteen emoticons zijn en buzxzers die zijn het leukst \
*8 pupil: it sucks that there arbent any emoticons and buzxzers they are the best *

Although the natural chat situation was not exactly replicated, the data show other interesting language activities that provide insight into how pupils deal with technological restrictions or differences. As such the data provide extremely interesting information. It remains to be seen to what extent our data resemble those of naturally occurring chat data.

There was one last aspect of our experimental setup that specifically deviates from the natural situation, which were the topics that students had to chat about. For sociolinguistic purposes we wanted to see if students chatted differently when talking about an involved or a non-involved topic. We furthermore wanted to have a comparable dataset in which pupils were allowed to talk about anything at all. During the evaluation afterwards it appeared that students often did not stick to the given topics. Therefore we believe that having such topics did not specifically add to the unnaturalness of the situation:

(11) 12 leerling: we hebben het niet echt over tmf awards maar kan mij het schelen
12 student: we didn't really talk about tmf awards but I don't care

For some research communities (e.g., conversation analysts) collecting data in an experimental situation and under suboptimal technical circumstances affects the validity and hence the usefulness of the collected materials. Although we have no emprical evididence yet to support the claim, we believe that many of the phenomena that linguists look for in these type of CMC data are represented in the corpus and hence the corpus will be of use for many researchers.

4.5 Challenge 5: Ethical dilemmas

As can be expected amongst teenagers, not all pupils get along with each other. One of the classes even dealt with regular bullying which had been discussed at school with the students and parents. The fact that the pupils realize that they are being monitored does not prevent them from bullying while chatting:

(12) 3 leerling: Leerling 4 is een lul:P
[...]
3 leerling: het stinkt naar lleeerling 4
[...]
3 leerling: miriam ruitk vies
[...]
3 leerling: jaah leerling 4 g0re l*ul

3 pupil: Pupil 4 is a dick:P
[…]
3 pupil: It smells of ppupil 4
[...]
3 pupil: miriam stinks
[...]
*3 pupil: yeah pupil 4 dIrty d*ick*

Pupils dare to say a lot of things in the chat room, perhaps more so than face to face in the classroom, or when writing papers. Especially the boys-only conversations are filled with slang and curses. There are entire sessions where boys just curse at each other and use curse abbreviations. Pupils like to make fun of each other but they also make fun of (or bully) pupils who are not partaking in the session. Even though the pupils are aware that the conversations are being recorded and that their input is being used for research, they do not seem to mind using vulgar language or talking about taking drugs or crime related topics, as in example (3).

Such language use creates ethical issues. Researchers saw these data but also the pupils' own teachers. Even though the data are anonymized, should we not protect the pupils by excluding these materials from the corpus that in principle is expected to last for a long period of time? We decided to leave all of the materials in the corpus, as the pupils and their parents had consented to the data collection, but we realize that such a choice is up for debate. We are interested in how other corpus analysts look upon this issue.

4.6 Challenge 6: Understanding in-group language

The results of the questionnaire show that 77% of first graders and 88% of third graders believe chat language to be like spoken language (vs written language or both). Although this remains an unanswered and much discussed question for linguists (see the introduction), pupils themselves seem to have a more unified view on this matter. Knowing that pupils themselves feel like their language use resembles spoken language when chatting, might make understanding the data a little bit easier. For example, in the data we see contributions such as written out animal sounds:

(13) 26 leerling: Kukelekuuuuuuuuuuuuh
26 pupil: Cock-a-doodle-dooooooo

instrument sounds:

(14) 22 leerling: toeteretoeteretoet
22 pupil: tooootooootootoot (trumpet sound)

extreme use of interpunction symbols:

(15) 10 leerling: ?????????
10 pupil: ?????????

phonetic spelling:

(16) 27 leerling: nahja maar hoezo ik wist egt neit dattiej bij soon club ofzo zat…
27 pupil: nooow but why i relly didt know thathej was part of such a club or smth...

Furthermore, pupils quote songs, chat in different languages (English, Italian, German), frequently misspell words, shorten words, use abbreviations and create creative language (also see van Charldorp 2006). When cleaning up the data we learned one important lesson:

do not throw away data that look unfamiliar. The data proved to be enormously rich. It provides a great insight into the creative use of language by Dutch teenagers in a CMC environment that can be used for a great variety of research topics. At the same time, this creativity creates the challenge of how to normalize and standardize the spelling variants to make the data searchable (cf. Oostdijk and van Halteren 2013 for a similar issue in Twitter).

4.7 Challenge 7: Sustainability of the corpus

In developing our corpus its sustainability proved to be a challenge in several respects. Firstly, there is the issue of the rapid technological developments, which can make systems and even complete genres obsolete almost overnight. MSN is a good example: while being the dominant chat system in the period of our data collection, presently it has disappeared and has been succeeded by systems with different technological affordances like Whatsapp, Facebook chat and Twitter. In a sense this means that corpus linguists interested in the relationship between language use and medium should be aware of their role as archeologists of language, before the phenomenon of interest has disappeared.

A second type of sustainability involves the extension of the materials. The creation of the ChatIG corpus depended completely on the cooperation with the Ignatius Gymnasium Amsterdam, funding from the VU University, and our contacts with enthusiastic teachers. That makes data collection also a vulnerable type of operation. After funding stopped and one of the teachers involved took a different position, our data collection project was discontinued. Fortunately, researchers from the Language and Speech Technology Group at Radboud University Nijmegen have set up a chatbox to collect data and have thus added to the collection of chat data in SoNaR considerably.

A third type of sustainability is related to making the data accessible for linguistic research. To that end our data were transferred to the Language and Speech Technology Group at RU Nijmegen, which has standardized and annotated the data following the SoNaR standard (cf. Sanders 2012 for details and references).

5 Conclusion and discussion

CMC data contain a wealth of information for corpus linguists, and hence it is imperative that we collect such data before the technology by which they were produced becomes obsolete. In this paper we described the challenges we encountered in our collection of chat data in the ChatIG project. We have shown that access to the right group of people to produce the data can be difficult, and depends very much on the social network of the researcher. Especially when we are dealing with a specific target group such as secondary school pupils, data collection creates issues of privacy and consent. We firmly believe that those issues should not be taken lightly and that only a full consent (or, more specifically, consent and assent) of the contributors of the corpus is ethically acceptable.

Other ethical issues occur when contributors display socially unacceptable behavior during the chat sessions. As researchers we should realize that a vulnerable group like secondary school pupils may not be aware of the consequences when displaying such behavior in a data collection project that intends to generate sustainable data for future research; should we protect the students and eliminate the data from the corpus? But that would eradicate a type of language behavior that might very well be typical for the type of language use under investigation and hence affect the validity and usability of the resulting corpus. As mentioned above, we decided to leave these data in the corpus, thus prioritizing the validity of the materials. We welcome a debate about this issue.

Data collection also creates linguistic and logistic challenges. How do we know for sure that we understand the speech-like utterances that are full of slang and language games? Of course, this issue is not unique for collecting CMC data (viz. the analysis of so-called *straattaal* ('street language') by Appel 1999 and Nortier 2001), but given the speech-like character of CMC language use by this age group, we may well need an anthropological take on learning to understand this type of in-group language. Logistically, we need to devote attention to a timely collection of the data, before the technology is out of date, in such a way that it leads to a substantial amount of data that is suitable for corpus linguistic analysis, and hence links on to established formats. That is why we believe that our case study has important implications for the collection of data from newer technologies like Whatsapp: we urgently need to collect sufficient amounts of those data and store them in a standardized format before it is too late.

References

Appel, R. (1999). Straattaal; De mengtaal van jongeren in Amsterdam. *Toegepaste taalwetenschap in artikelen*, *62*, 39-55.

Baron, N. (2008). *Always on: Language in an online and mobile world*. Oxford etc.: Oxford University Press.

Becker, J.A.H. and Stamp, G.H. (2005). Impression management in chat rooms: A Grounded Theory Model. *Communication Studies, 56*(3), 243-260, DOI: 10.1080/10510970500181264

Beißwenger, M. and Storrer, A. (2008). Corpora of computer-mediated communication. In: A. Lüdeling and M. Kytö (Eds), *Corpus Linguistics. An International Handbook. Volume 1.* (pp. 292-308). Berlin / New York: De Gruyter.

Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L. and Storrer, A. (2013). *DeRiK*: A German reference corpus of computer-mediated communication. *Literary and linguistic computing, 28*(4), 531-537.

Chafe, W. L. (1982). Integration and involvement in speaking, writing, and oral literature. In: D. Tannen (Ed.), *Spoken and Written Language: Exploring Orality and Literacy* (pp. 35-54). Norwood, NJ: Ablex.

Charldorp, T.C. van (2005). Building a chat corpus. Unpublished manuscript, available from http://www2.let.vu.nl/documenten/corpora/Building_a_Chat_Corpus_2005.pdf

Charldorp, T.C. van (2006). Dutch teenage chat communication: a sociolinguistic perspective. Unpublished thesis VU University Amsterdam.

Gries, S.T. and Newman, J. (2013). Creating and using corpora. In Robert J. Podesva and Devyani Sharma (eds.), *Research methods in linguistics* (pp. 257-287). Cambridge: Cambridge University Press.

Herring, S. C. (2004). Computer-mediated discourse analysis: An approach to researching online behavior. In S. A. Barab, R. Kling, and J. H. Gray (Eds.), *Designing for Virtual Communities in the*

Service of Learning (pp. 338-376). New York: Cambridge University Press. Preprint: http://ella.slis.indiana.edu/~herring/cmda.pdf.

Herring, S. C. (2013). Discourse in Web 2.0: Familiar, reconfigured, and emergent. In D. Tannen and A. M. Tester (Eds.), *Georgetown University Round Table on Languages and Linguistics 2011: Discourse 2.0: Language and new media* (pp. 1-25). Washington, DC: Georgetown University Press. Prepublication version: http://ella.slis.indiana.edu/~herring/GURT.2011.prepub.pdf

Kestemont, M., Peersman, C., de Decker, B., de Pauw, G., Luyckx, K., Morante, R., Vaassen, F., van den Loo, J. and Daelemans, W., (2011). The Netlog Corpus. A Resource for the Study of Flemish Dutch Internet Language, in: *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 23-25). Istanbul: European Language Resources Association (ELRA).

King, B. (2009). Building and analysing corpora of computer-mediated communication. In: Baker, P. (ed.), *Contemporary Corpus Linguistics* (pp. 301-320). London: Continuum.

Lewis, A. (2002). Accessing, through research interviews, the views of children with difficulties in learning. *Support for Learning, 17*(3), 1467-9604. DOI: 10.1111/1467-9604.00248.

Nortier, J. M. (2001). *Murks en straattaal: Vriendschap en taalgebruik onder jongeren*. Amsterdam: Prometheus.

Oostdijk, N.H.J. and Halteren, H. van (2013). Shallow parsing for recognizing threats in Dutch tweets. In: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)* (pp. 1034-1041). New York: IEEE. http://dx.doi.org/10.1145/2492517.2500271

Oostdijk, N.H.J., Reynaert, R., Schuurman, I. and Hoste, V. (2013). The construction of a 500-million-word reference corpus of contemporary written Dutch. In P. Spyns and J. Odijk (Eds.), *Essential Speech and Language Technology for Dutch: Theory and Applications of Natural Language Processing* (pp. 219-247). Berlin: Springer.

Sanders, E. (2012). Collecting and analysing chats and tweets in SoNaR. In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani et al. (Eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul: European Language Resources Association (ELRA). Available from http://www.lrec-conf.org/proceedings/lrec2012/pdf/416_Paper.pdf

Sanders, T.J.M. and Spooren, W.P.M.S. (2013). Exceptions to rules: a qualitative analysis of backward causal connectives in Dutch naturalistic discourse. *Text & Talk, 33*(3), 399-420.

Tagg, C. (2009). A corpus linguistic study of SMS text messaging. Unpublished PhD thesis.

ten Have, P. and Komter, M. (1982). De angst voor de tape. Over bezwaren tegen het gebruik van de bandrecorder voor onderzoek. In C. Bouw, F. Bovenkerk, K. Bruin, and L. Brunt (Eds.), Hoe weet je dat? Wegen van sociaal onderzoek (pp. 228-242). Amsterdam: Uitgeverij de Arbeidspers & Wetenschappelijke Uitgeverij.