COLLECTING INFORMAL PRIVATE CMC-DATA PRODUCED BY MINORS:

practical and ethical challenges

REINHILD VANDEKERCKHOVE

LISA HILTE

CLiPS

Universiteit Antwerpen

STRUCTURE

1. Introduction: CLiPS CMC-corpora

2. Challenges

   2.1. Ethical issues

   2.2. Getting all parties on board

   2.3. Data issues

1. INTRODUCTION: CLiPS CMC-Corpora

1. INTRODUCTION: CLiPS CMC-Corpora

INFORMAL
+
mainly PRIVATE
+
FLEMISH adolescents

	CORPUS 2007-2013	CORPUS 2015-2016
Size (tokens)	2 066 521	2 531 354
Media	MSN Netlog Facebook Messenger	Facebook Messenger WhatsApp
Metadata	Age Gender Medium Region	Age Gender Medium Educational track *parental profession* *home language* (Region)

Oh zaaaaalig 😍 **= 3 tokens**

	CORPUS 2007-2013	CORPUS 2015-2016
Size (tokens)	2 066 521	2 531 354
Media	MSN Netlog Facebook Messenger	Facebook Messenger WhatsApp
Metadata	Age Gender Medium Region	Age Gender Medium Educational track *parental profession* *home language* (Region)

	CORPUS 2007-2013	CORPUS 2015-2016
Size (tokens)	2 066 521	2 531 354
Media	MSN Netlog Facebook Messenger	Facebook Messenger WhatsApp
Metadata	Age **(13-16 / 17-20)** Gender Medium Region **(central provinces + eastern & western periphery)**	Age **(13-16 / 17-20)** Gender Medium Educational track **(general, technical, vocational)** *parental profession* *home language* (Region)

- **Corpus 2007-2013**: data collection via networks linguistics students University of Antwerp and via Netlog project CLiPS

- **Corpus 2015-2016**: data collection via secondary schools

→ Students in computer classes

→ Data are sent on the spot via website

→ Stronger grip on procedure data collection and profile informants

2. CHALLENGES

2.1 How to deal with ethical issues?

Funding most recent project (corpus 2015-2016):

dependent on ethical clearance by

Ethical Advisory committee Social and Human Sciences
University of Antwerp

Conditions for ethical clearance:

- consent adolescent
- consent parent (for minors)
- anonymization
- secure storage ➔ no dessimination

➔ No data exchange with other researchers working on CMC

- destruction data in 20 years

➔ No long-term diachronic research

Additional tricky issue:

Are requirements local ethical advisory boards compatible with legal requirements?

E.g.:

- Destruction of the link between an informant code and the identity (name) of the informant is approved by EASHW Antwerp
- But informants have the right to request removal of their data from database (at any time)

➔ Link data-informant has to be reconstructable

EU GDPR will be in force from 25th May 2018:

Art. 17: “Right to erasure (’right to be forgotten’)”

https://www.eugdpr.org/eugdpr.org.html
https://gdpr-eu.be/wat-is-gdpr/

“Right to be Forgotten
Also known as Data Erasure, the right to be forgotten entitles the data subject to have the data controller erase his/her personal data, cease further dissemination of the data, and potentially have third parties halt processing of the data. The conditions for erasure, as outlined in article 17, include the data no longer being relevant to original purposes for processing, or a data subjects withdrawing consent. It should also be noted that this right requires controllers to compare the subjects' rights to "the public interest in the availability of the data" when considering such requests”

https://www.eugdpr.org/key-changes.html

Practical issue:

automatic anonymization is challenging:

- The use of capital letters is not a workable criterion
- Automatic selection and replacement based on name lists leads to unintended data loss + retention of names that are accidentally or deliberately (creatively) ‘misspelt’

Practical issue:

automatic anonymization is challenging:

- The use of capital letters is not a workable criterion
- Automatic selection and replacement based on name lists leads to unintended data loss + retention of names that are accidentally or deliberately (creatively) ‘misspelt’

E.g.: name ‘Ben’ leads to erasure of verb form ‘ben’ (of verb ‘zijn’ - ‘to be’)

Practical issue:

automatic anonymization is challenging:

- The use of capital letters is not a workable criterion
- Automatic selection and replacement based on name lists leads to unintended data loss + retention of names that are accidentally or deliberately (creatively) ‘misspelt’

Sooooooofie (for ‘Sofie’) which contains letter flooding = relevant expressive marker

2.2 How to get all of the parties involved on board?

SCHOOL BOARDS AND TEACHERS:

Creating goodwill by offering ‘return’:

a mini computational and sociolinguistic course on the online writing practices of adolescents for their students

Win-win situation

- Teachers appreciated the course (adapted to the educational level of their students)
- We could make clear what type of research we are doing with the data → reassure both teachers and students

➔ Teachers were willing to:

- Motivate students to donate data (but no pressure!)
- Create optimal conditions by booking computer classes for their students where data could be donated on the spot

HOW TO ACTIVATE AND REASSURE THE ADOLESCENTS?

Challenges:

Too much administration = barrier:

paper consent forms end up in dust bin

➔ 1 website for all of the operations (www.chatproject.be):

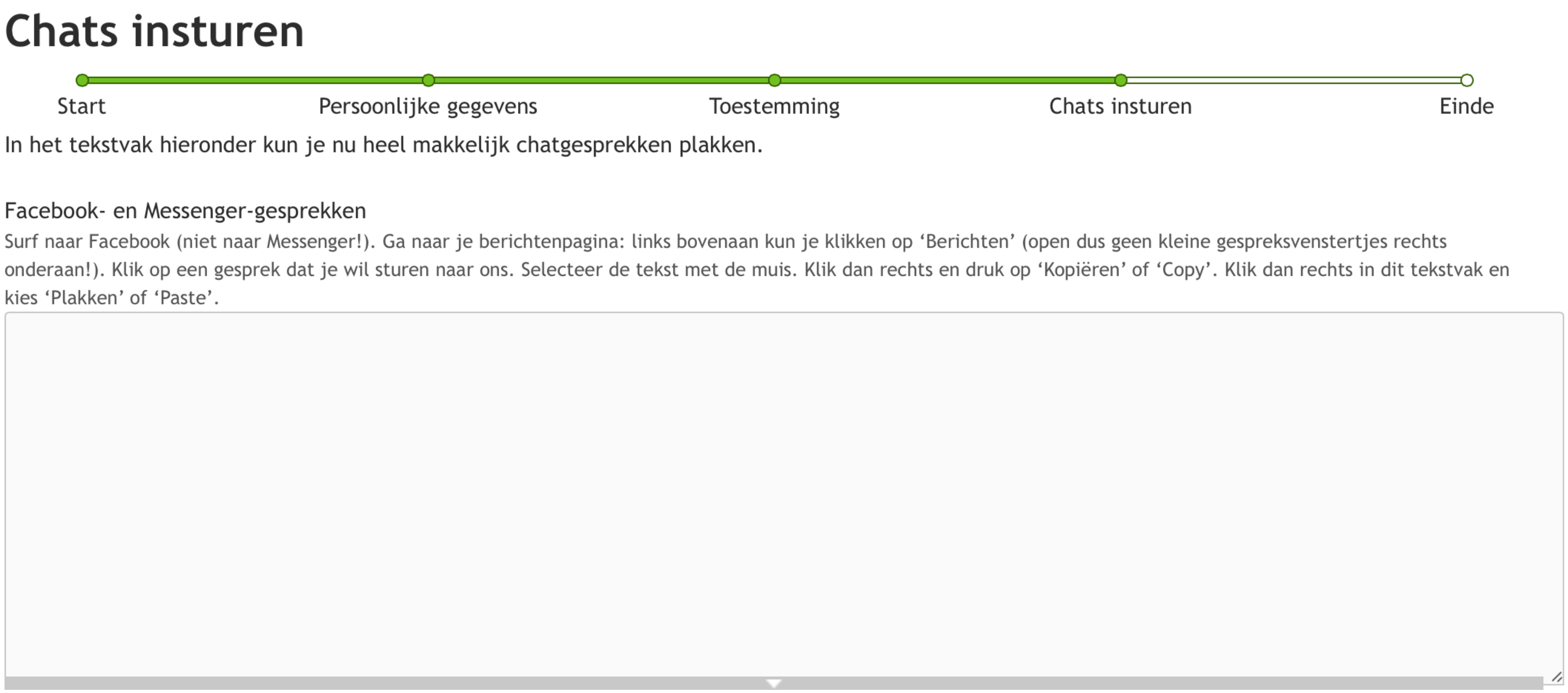
-giving consent

-entering profile data: place of residence, age, gender, educational track, home language, profession parents,

-copying chat conversations

-getting extra information on the project, data storage…

+ 1 paper form for the parents



Chats insturen

In het tekstvak hieronder kun je nu heel makkelijk chatgesprekken plakken.

Facebook- en Messenger-gesprekken

Surf naar Facebook (niet naar Messenger!). Ga naar je berichtenpagina: links bovenaan kun je klikken op ‘Berichten’ (open dus geen kleine gespreksvenstertjes rechts onderaan!). Klik op een gesprek dat je wil sturen naar ons. Selecteer de tekst met de muis. Klik dan rechts en druk op ‘Kopiëren’ of ‘Copy’. Klik dan rechts in dit tekstvak en kies ‘Plakken’ of ‘Paste’.

Challenges:

- Consent of parents = complicating factor
- Computer skills of the students! (especially in practice and vocational oriented educational tracks)

Reassuring elements / positive incentives:

- Guaranteed anonymity
- Automatic deletion of pictures when chat conversations are copied on the site
- Students can make their own selection and can send as many or as few conversations as they want
- Information and data could be entered on the spot >> no homework!
- The most ‘generous donors’ of each school won duo-film tickets

2.3. Data issues

Format of the data:

- Data collection via website ➔ uniform format

- Format choice :

(1) Excel:

-- loss of particular emoji

-- cut off for number of words per cell

2.3. Data issues

Format of the data:

- Data collection via website ➔ uniform format

- Format choice :

(1) Excel ✗
-- loss of particular emoji
-- cut off for number of words per cell

(2) txt with tsv format:
Extremely basic but no loss of data/special features

(note: csv format leads to unintended utterance splits due to smileys containing semicolons)

Interpretability of the social metadata:

Especially with respect to parental profession:

- Reluctance or ignorance
- Ambiguous or vague labels impede classification:

  ‘self-employed’, ‘harbour’, ‘bank’

Reliability of the data:

- Did students enter correct profile data?

educational level = unproblematic, in view of context of the context of the data collection

- Potential bias in data that are donated:

people may avoid sending their most intimate conversations = bias for the topics that are discussed (content level) rather than bias for linguistic features that are used

In spite of all the challenges and barriers…

we ended up with two corpora:

- Documenting 10 years of informal CMC and offering a unique view on private CMC produced by the trendsetting generation

e.g.: from ;-)

:-) to

:-(

- Big enough for reliable quantitative data processing, both with a sociolinguistic and computational linguistic approach

  for both descriptive and predictive statistics

e.g.: --Analysing linguistic gender patterns in the data

--Training software on the basis of the data for gender prediction on new data

- Enabling qualitative discourse-pragmatic analyses

- In some cases enabling age grading research since some students copied their entire chat history

Boy 18 years old:

kga is wa minder emoticons gebruiken

als ge da zo ziet ziet da er echt belachelijk uit

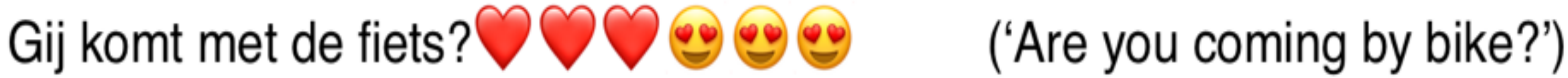
‘I’m going to use fewer emoticons

it looks so ridiculous’

Finally: feedback to students and teachers?

- Some schools were ‘revisited’
- Combination of survey on evaluation of CMC practices and presentation results

E.g.:



Gij komt met de fiets?❤️❤️❤️😍😍😍 (‘Are you coming by bike?’)

- Questionaire: produced by boy/girl - younger teenagers/older teenager…
- Afterwards: presentation of some of our findings with respect to gender and age correlations…

thx☺!