Collection, storage and analysis of online teenage talk: assets and challenges

REINHILD VANDEKERCKHOVE

In collaboration with:

LISA HILTE

WALTER DAELEMANS

CLiPS

Universiteit Antwerpen

1. CORPUS – collection
2. CORPUS – storage & ethical issues
    - ➔ Obstacles / Special focus: social class
3. CORPUS – size & composition: data processing
4. OPERATIONALIZATION CMC-features
5. RESEARCH FRAME – quantitative + qualitative

1. CORPUS – collection

Informal CMC	**CORPUS 2007-2013**	**CORPUS 2015-2016**
Size	2 066 521	2 885 084
Media	MSN Netlog Facebook	Facebook WhatsApp
Variables	Age Gender Medium	Age Gender Level of education *Profession parents* *Home language*

1. CORPUS – collection

- Flemish adolescents - 13-20 years old
- personal approach: activating respondents (e.g. via schools)

Asset:

- Control over data and metadata

Challenge:

- Time management
- Consent needed from a lot of partners
- Reliability/interpretation provided information, e.g. profession parents

2. CORPUS – storage & ethical issues

- Funding projects = dependent on ethical clearance by Ethical advisory committee Social and Human Sciences

- Conditions for ethical clearance:
  - -consent adolescent
  - -consent parent
  - -anonymization
  - -secure storage ➔ no dessimination
  - -destruction data in 20 years

➔ practical obstacles

e.g.: 2015-2016 corpus:

willingness - consent of 4 ‘partners’:

- school management
- teachers
- parents
- adolescents

➔ practical obstacles

Computational skills of pupils with a low level of education:

- In spite of high ‘smartphone dexterity’
- Troubles with simple operations like ‘copy&paste’

➔ some send screenshots ➔ transcription

➔ practical obstacles

Interpretability social metadata,

especially with respect to profession parents:

e.g.:

“Self-employed” (?)

“Harbour” (?)

- ➔ Ambiguity for profession of one of the parents: other parent = reference point for classification
- ➔ Ambiguity for both parents: no classification for this variable
- ➔ Clear ‘labels’ for both parents: profession with highest ranking = reference point

Social class background:
level of education + profession parents

Level of education – three categories:

- ASO: general secondary education: theoretical ➔ higher education
- TSO: technical secondary education: theoretical + practical ➔ hybrid
- BSO: vocational secondary education: practical ➔ manual profession

Profession parents – three categories (based on Erikson & Goldthorpe):

- I: higher-grade professionals with (most probably) university degree
- II: hybrid category, administrators, non-manual workers
- III: manual workers

Poles of the continuum:

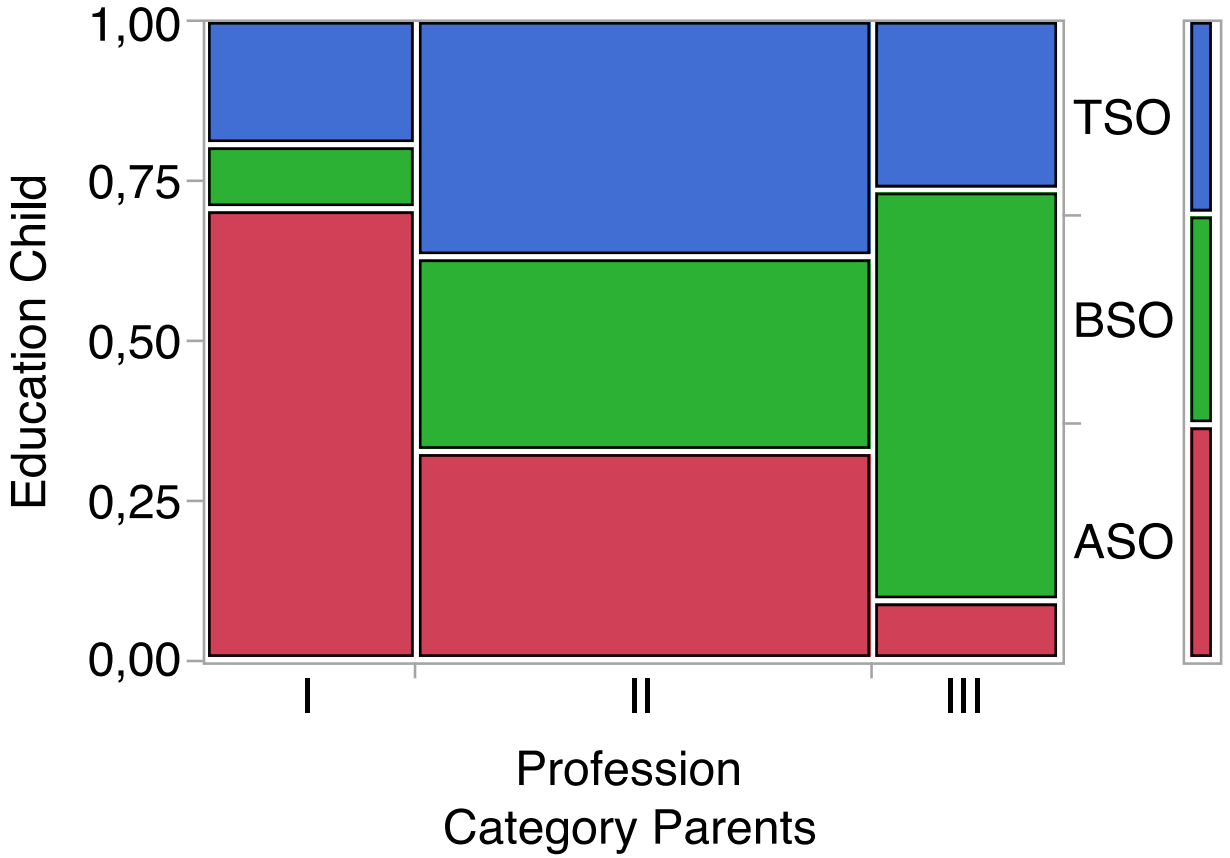
Level of education – three categories:

- ASO: general secondary education: theoretical ➔ higher education
- TSO: technical secondary education: theoretical + practical ➔ hybrid
- BSO: vocational secondary education: practical ➔ manual profession

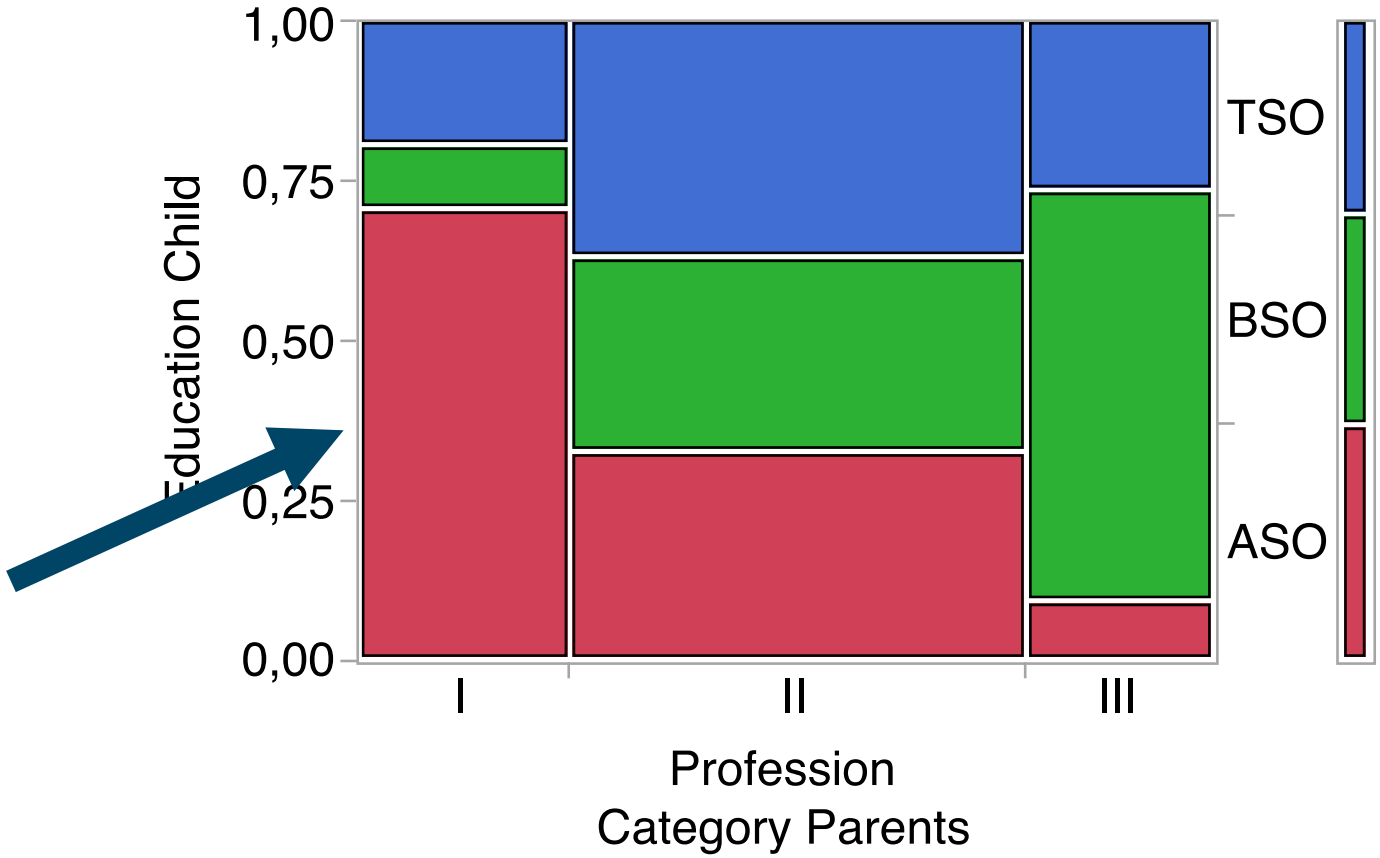
Profession parents – three categories (based on Erikson & Goldthorpe):

- I: higher-grade professionals with (most of them) university degree
- II: hybrid category, administrators, non-manual workers
- III: manual workers

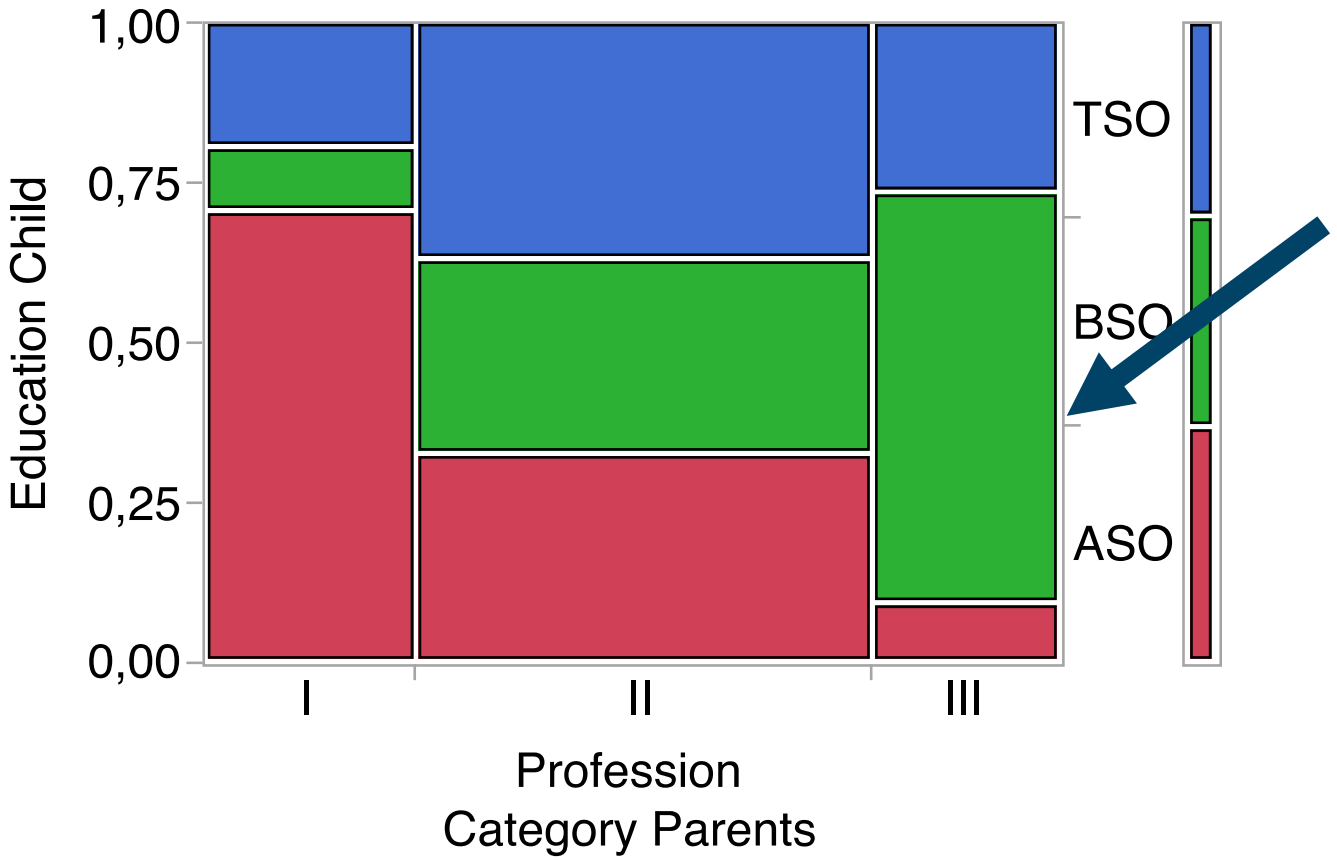
Limited social mobility:
strong correlation profession parent – educational level



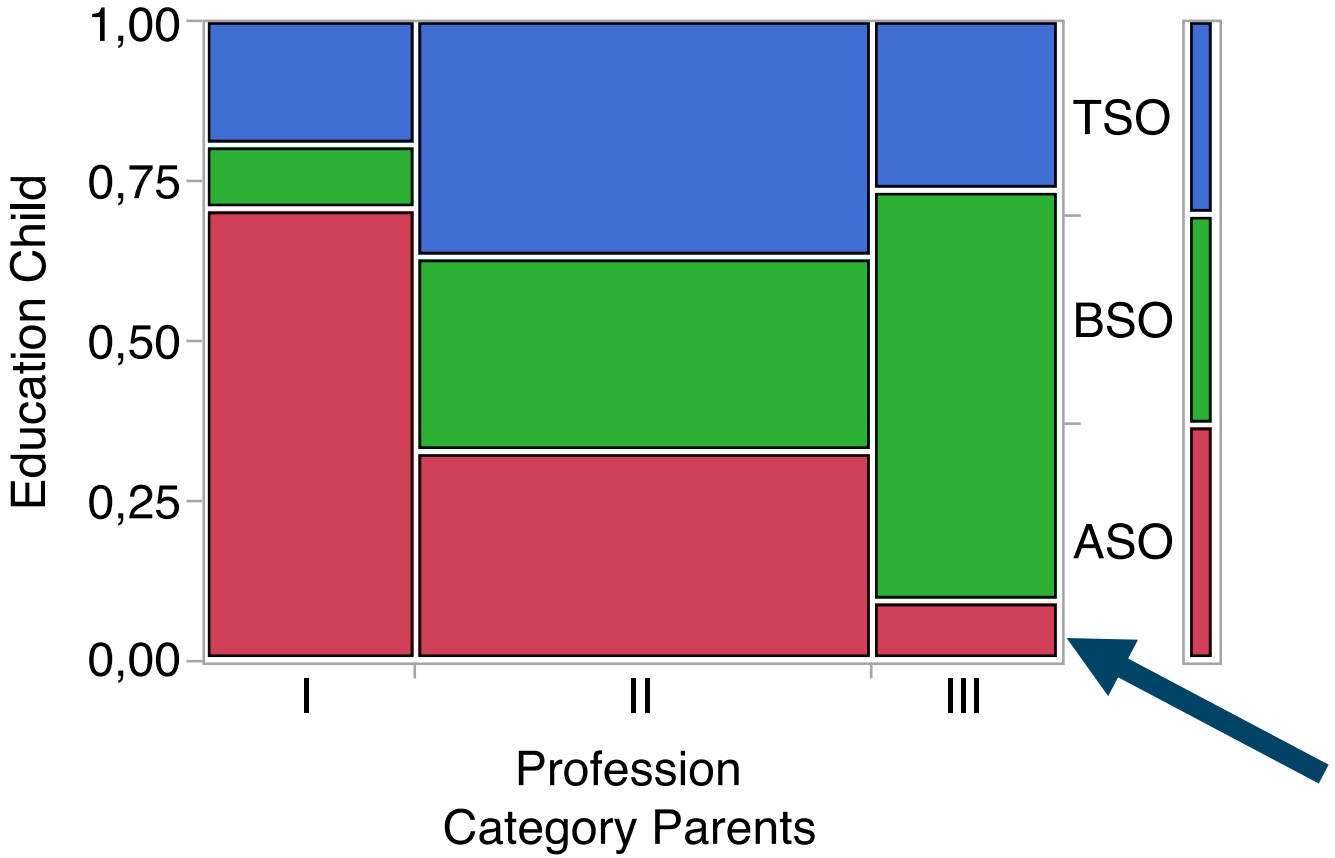
Limited social mobility:
stagnation for adolescents with ‘upper class’ parents



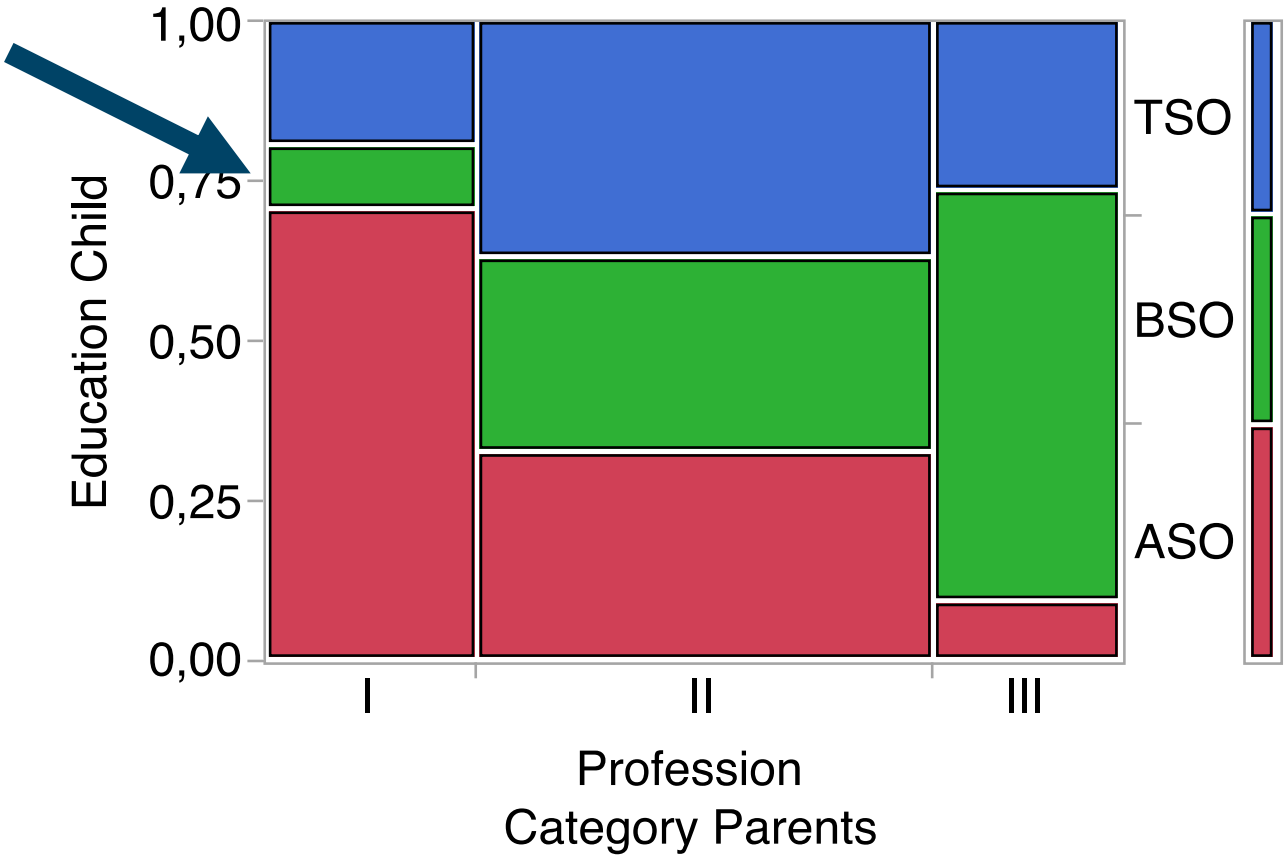
Limited social mobility:
stagnation for adolescents with ‘lower class’ parents



Limited social mobility:
upward social mobility for adolescents with lower class parents



Limited social mobility:
downward social mobility for adolescents with 'upper class' parents



Implication for online practices?

- Several CMC-cultures, CMC-lects?
- Connection with international chat culture?

3. CORPUS – size & composition: data processing

SIZE database: large database = statistical challenge:

Correlational analyses: extremely small differences in proportions render significant results

➔ Incorporate effect sizes (e.g. Odds ratio, Cramer's V)

➔ Alternative techniques: e.g. Bootstrapping:

--- 10 000 arbitrary samples of 100 000 tokens with replacement

--- For each of these samples:

X^2-value, p-value, odds ratio ➔ distribution?

e.g.

	male	female
A	55	60
B	60	55

X^2= 0.43, p = 0.51, odds ratio = 1.19

e.g.

	male	female
A	55	60
B	60	55

X^2= 0.43, p = 0.51, odds ratio = 1.19

	male	female
A	5500	6000
B	6000	5500

X^2= 43.48, p < 0.0001, odds ratio = 1.19

e.g.

	male	female
A	55	60
B	60	55

$X^2 = 0.43$, $p = 0.51$, odds ratio = 1.19

	male	female
A	5500	6000
B	6000	5500

$X^2 = 43.48$, $p < 0.0001$, odds ratio = 1.19

YET BIG DATA SETS REMAIN AN ASSET!

3. CORPUS – size & composition: data processing

COMPOSITION database:

spontaneous data from natural settings

> **unbalanced corpora**

➔ Assigning weights to specific groups in function of their over- or underrepresentation

➔ Linear mixed models: in order to deal with unbalanced contribution individual chatters, with missing data

4. OPERATIONALIZATION CMC-features

frame: development general index of non-standard/CMC writing

Analysis on token level - **binary** approach:

1: token with CMC feature

0: token without CMC features

However, e.g.:

Niiiiiiiice☺ = 1 token, 2 CMC features

Or:

☺ = 1 token, 1 CMC feature

☺☺☺☺☺☺☺ = 1 token, sequence of CMC features

➔ Information gets lost:

piling up CMC-features is relevant from a discourse-pragmatic perspective

➔ Proportions hide a more complex reality:

number of CMC features / total number of tokens

≠ percentage of tokens with CMC features

Solution: operationalisation **ordinal** variables

e.g., research **Hilte et al.**:

0 = token contains no CMC feature: nice

1 = token contains 1 CMC feature: niiiiice

2 = token contains more than 1 CMC feature: niiiiice☺

Disadvantages:

- many statistical packages do not support combination ordinal model – mixed approach (i.e. approach with random effects)
- results in '**heavy model**' since the response variable contains more options

➔ Combination ordinal variable + several independent variables *(e.g.: age, gender, level of education, home language, profession parents)* >> complex model >> decoding the output becomes a challenge!

Increase number of variables ➔ decrease output transparency

5. RESEARCH FRAME:
Integrating quantitative and qualitative approaches

Quantitative:

correlating micro-linguistic variation with ‘fixed’ social variables

Qualitative:

“The qualitative approach reveals **how** participants in CMC draw on various linguistic resources in **shaping their online personae and in accomplishing various interactional tasks**.” (Androutsopoulos & Ziegler 2004, see also Vandekerckhove & Nobels 2010)

➔ Exclusive focus on quantifying may obscure CMC-pragmatics

➔ Social indexicality of CMC features to chatters?

e.g. (1): research Hilte:

youngsters with low level of education: higher frequency of CMC features ➔ more attracted to CMC features? / Pivotal role in online identity construction?

Youngsters with high level of education: lower frequency ➔ disconnect themselves from particular ‘silly’ features (as they grow older)??

e.g. (2): discourse-pragmatic function of individual features: interpretation of smileys

- Expression happiness, humor, irony…

OR

- Developing into ‘standard’ means of closing a sentence?

Qualitative scope:

➔ discourse analysis:

Content and context analysis:

e.g.: - Do chatters poke fun at particular features?

- Metacomments?

➔ (Ethnographic) interviews/surveys

e.g. (example from corpus 2015-2016 translated into English):

V16: Shall I call you later on? You sounded really upset :/ :S

V15: no, is okay ☺ ☺

= actual post

COMPARE:

V15: no, is okay

➔ Difference in interpretation/tone?

VERSUS:

V15: no, is okay.

➔ Smileys

= expressive markers

function: establishing emotional connection

= general discourse markers

function: determining the general tone of the conversation

➔ Interviews/surveys with adolescents on:

- Interactional-pragmatic meaning of CMC-features
- Attitudes towards features

e.g. with Likert scale

V16: Shall I call you later on? You sounded really upset :/ :S

V15: No, is okay.

V15 sounds:

Disturbed neutral friendly

➔ Discourse analysis:

e.g. 18 year old - male:

M: *kga is wa minder emoticons gebruiken*
als ge da zo ziet ziet da er echt belachelijk uit
'I'm going to use less emoticons from now on
if you see that, it looks really ridiculous'

➔ Discourse analysis:

e.g. Conversation between two 19 year old boys:

M1: irriteren die afkortingen u eig?

‘do these abbreviations irritate you?’

M2: ja :p

afkortingen buiten brb en wtf zen stoem

‘yes :p, abbreviations apart from brb and wtf are stupid’

M1: aight

en ty/thnx?

‘aight, and ty, thnx’?

M2: zegt gewoon merci

gelak elke normale mens :P

‘just say ‘thanks’, like every normal human being :P’

➔ Qualitative research = complementary to quantitative research

THX☺!

References

Corpus 2007-2013:

De Decker, Benny & Reinhild Vandekerckhove (2017): Global features of online communication in Flemish: Social and medium-related determinants. In: *Folia Linguistica* 51/1: 253–281

Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans (2016). Expressiveness in Flemish online teenage talk: A corpus-based analysis of social and medium-related linguistic variation. In Darja Fiser & Michael Beisswenger (eds.), *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities, Ljubljana, Slovenia, 27-28 September 2016,* pp. 30-33.

Corpus 2015-2016:

Hilte, Lisa, Reinhild Vandekerckhove & Walter Daelemans (in preparation): Adolescents' social background and non-standard writing in online communication

Androutsopoulos, Jannis & Evelyn Ziegler (2004): Exploring language variation on the Internet: Regional speech in a chat community. In: Gunnarsson et al. (eds.): *Papers from the Second International Conference on Language Variation in Europe*. Uppsala, Sweden: Uppsala University. 99–111.

Vandekerckhove, Reinhild & Judith Nobels (2010): Code eclecticism: linguistic variation and code alternation in the chat language of Flemish teenagers. In: *Journal of Sociolinguistics* 14/5, 657-677.