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

REINHILD VANDEKERCKHOVE

In collaboration with:
LISA HILTE
WALTER DAELEMANS

CLiPS
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

<table>
<thead>
<tr>
<th>Informal CMC</th>
<th>CORPUS 2007-2013</th>
<th>CORPUS 2015-2016</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Size</strong></td>
<td>2,066,521</td>
<td>2,885,084</td>
</tr>
<tr>
<td><strong>Media</strong></td>
<td>MSN, Netlog, Facebook</td>
<td>Facebook, WhatsApp</td>
</tr>
<tr>
<td><strong>Variables</strong></td>
<td>Age, Gender, Medium</td>
<td>Age, Gender, Level of education, Profession, parents, Home language</td>
</tr>
</tbody>
</table>


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 dissemination
  - 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 $\Rightarrow$ distribution?
**e.g.**

<table>
<thead>
<tr>
<th></th>
<th>male</th>
<th>female</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>55</td>
<td>60</td>
</tr>
<tr>
<td>B</td>
<td>60</td>
<td>55</td>
</tr>
</tbody>
</table>

\[ \chi^2 = 0.43, \ p = 0.51, \ \text{odds ratio} = 1.19 \]
### e.g.

<table>
<thead>
<tr>
<th></th>
<th>male</th>
<th>female</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>55</td>
<td>60</td>
</tr>
<tr>
<td>B</td>
<td>60</td>
<td>55</td>
</tr>
</tbody>
</table>

\(X^2 = 0.43, p = 0.51, \text{ odds ratio } = 1.19\)

<table>
<thead>
<tr>
<th></th>
<th>male</th>
<th>female</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5500</td>
<td>6000</td>
</tr>
<tr>
<td>B</td>
<td>6000</td>
<td>5500</td>
</tr>
</tbody>
</table>

\(X^2 = 43.48, p < 0.0001, \text{ odds ratio } = 1.19\)
### Example

<table>
<thead>
<tr>
<th></th>
<th>male</th>
<th>female</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>55</td>
<td>60</td>
</tr>
<tr>
<td>B</td>
<td>60</td>
<td>55</td>
</tr>
</tbody>
</table>

\[ X^2 = 0.43, \ p = 0.51, \ \text{odds ratio} = 1.19 \]

<table>
<thead>
<tr>
<th></th>
<th>male</th>
<th>female</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5500</td>
<td>6000</td>
</tr>
<tr>
<td>B</td>
<td>6000</td>
<td>5500</td>
</tr>
</tbody>
</table>

\[ X^2 = 43.48, \ p < 0.0001, \ \text{odds ratio} = 1.19 \]

**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.:
Niiiiiiiiice😊 = 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: niiiiiice

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?*
- *Metacommments?*

- *(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

X
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:


Corpus 2015-2016:
Hilte, Lisa, Reinhild Vandekerckhove & Walter Daelemans (in preparation): Adolescents’ social background and non-standard writing in online communication