4th Conference on CMC and Social Media Corpora for the Humanities

28-09-2016

Linguistic Characteristics of Dutch Computer-Mediated Communication:

CMC and School Writing Compared

Lieke Verheijen

Netherlands Organisation for Scientific Research

CLS | Centre for Language Studies
Radboud University


LIKE
Follow!
@
ONLINE
SOCIAL MEDIA
WWW.
Check-In
Friends
Chat
SHARE
Enter

Introduction

CMC: communication via modern technologies

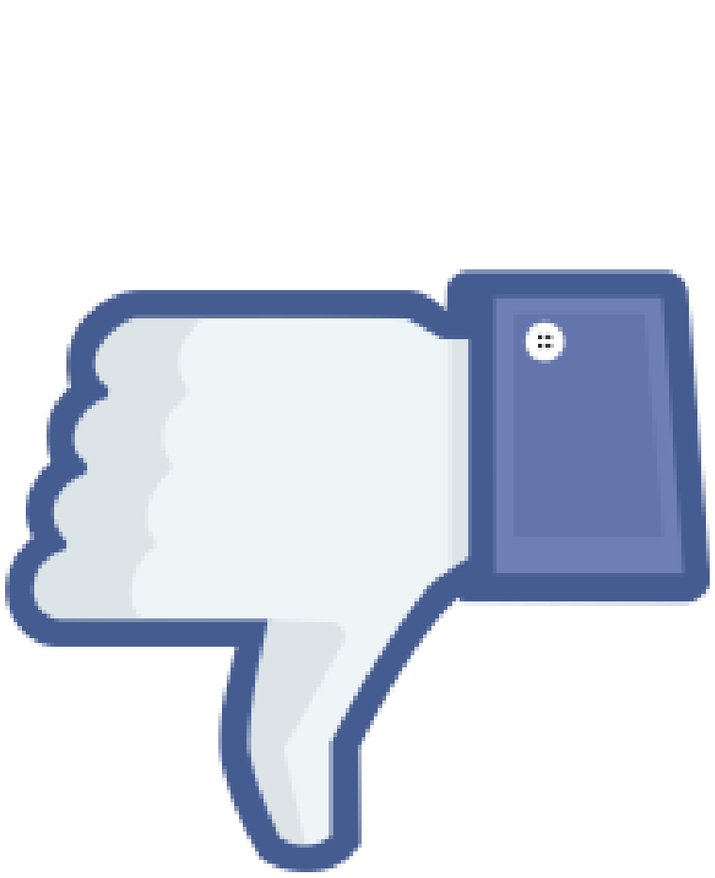
Previous research on written CMC:
English, German, French, Italian,
Spanish, Portuguese, Finnish, Swedish, …

Differences from standard language conventions
(e.g. Thurlow & Brown 2003, Crystal 2008, Frehner 2008, Cougnon & Fairon 2014)

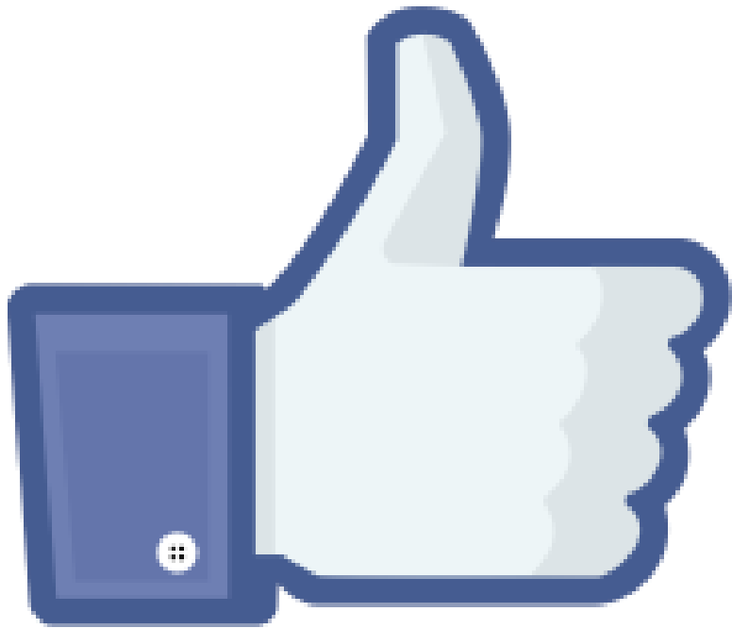
- **nonstandard orthography**: *fyi i'll B @home l8er 2night, r u OK with that? :-)*
- **syntactic omissions**: *car broken down, mailed garage yesterday, haven't responded yet*


YOU GUYS SERIOUSLY DON'T CHECK E-MAIL??
CAN YOU TWEET THE HOMEWORK @ US?
OR SNAP CHAT IT?
BTW, CAN YOU-
WAIT.
WAIT!
DID YOU JUST SAY "B-T-W" OUT LOUD?
INSTEAD OF "BY THE WAY"?
UH, YEAH.
JORGE CHAM ©2014
IT ACTUALLY HAS MORE SYLLABLES!!
RLY?
OMG!
OBVS.

Gr8 Db8



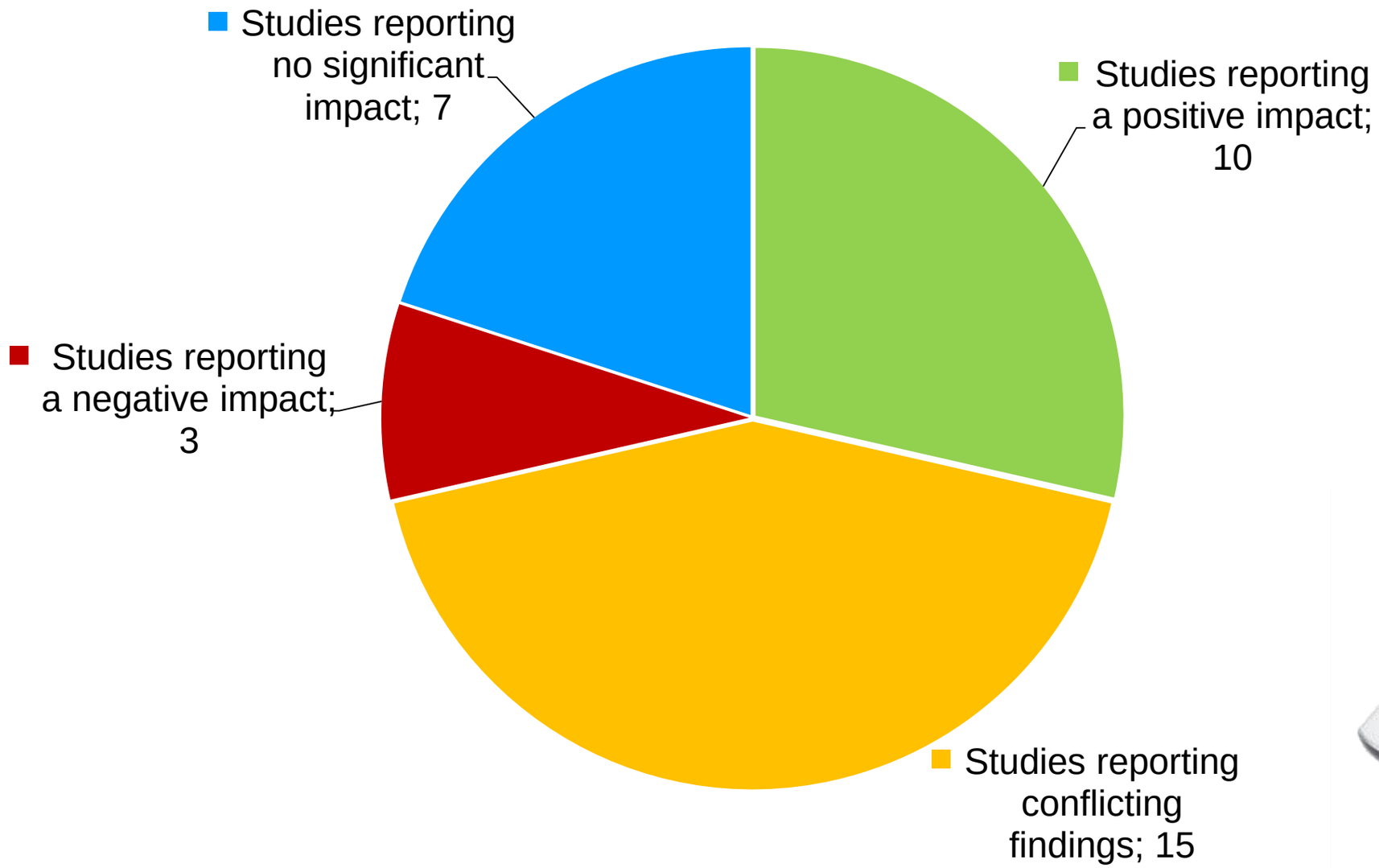
or



Research questions: What does the language used by Dutch youths in CMC actually look like? How does it differ from Standard Dutch?

Research goal: explore how Dutch youths' informal written CMC linguistically differs from their more formal school writings

Previous studies on how CMC affects literacy



Mixed results...

Explanation: many differences in methodology & participants

(Verheijen 2013)

Materials

CMC writings

Instant messaging with Internet application: **MSN**

Selected from existing corpus (SoNaR: *STEVIN Nederlandstalig Referentiecorpus*, Oostdijk et al. 2013)

Text messaging (**SMS**)

Microblogging: **Twitter**

Instant messaging with mobile phone app: **WhatsApp**

Collected via cls.ru.nl/whatsapptaal

Social networking site: **Facebook**

Collected via cls.ru.nl/facebooktaal

- ages: 12-23

School writings

- lower & higher educational levels
- adolescents & young adults

Materials: CMC writings

Corpus of CMC texts so far

Genre	Year(s) of collection	Age group	Mean age	# words	# chats or contributors[i]
MSN	2009-2010	12-17	16.2	45,051	106
		18-23	19.5	4,056	21
			total:	**49,107**	**127**
SMS	2011	12-17	15.4	1,009	7
		18-23	20.4	23,790	42
			total:	**24,799**	**49**
Twitter	2011	12-17	15.9	22,968	25
		18-23	20.6	99,296	83
			total:	**122,264**	**108**
WhatsApp	2015	12-17	14.4	55,865	11 / 84
		18-23	20.1	140,134	23 / 132
			total:	**195,999**	**34 / 216**
			grand total:	**392,169**	

*No. of chats: MSN, WhatsApp; no. of contributors: SMS, tweets, WhatsApp

Materials: school writings

Corpus of school essays

Educational level	Year(s) of production	Age group	# words	# texts
lower secondary education (‘vmbo’)	2013–2014	± 14-15, 3rd grade	50,143	128
higher secondary education (‘vwo’)	2013–2014	± 14-15, 3rd grade	50,070	153
lower tertiary education (‘mbo’)	2012–2014	± 17-18, 2nd grade	39,793	137
higher tertiary education (‘uni’)	2012–2014	± 18-19, 1st grade	50,175	169
		total:	**190,181**	

Method

Register analysis: quantitative study of linguistic features

1) manual analysis: CMC writings

Orthographic features
- textisms
- misspellings
- typos
- emoticons
- symbols

Syntactic feature
- omissions

Lexical features
- borrowings
- interjections

2) automatic analysis: CMC writings vs. school writings, **T-Scan** (Pander Maat et al., 2014)

Lexical measures
- lexical diversity (MTLD)
- density of ‘special’ words
- lexical density
- density of ellipses

Syntactic complexity measures
- average of all dependency lengths
- average no. of subordinate clauses
- average sentence length
- D-level

Statistics

independent t-tests

(one-tailed probability values reported)

Results + discussion:

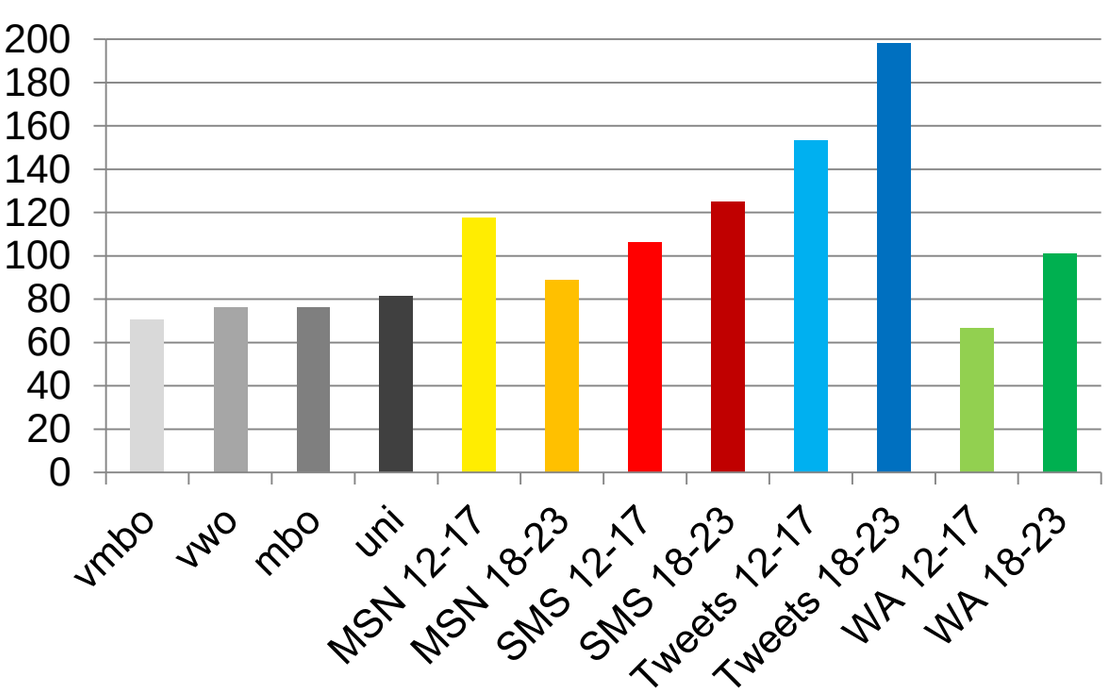
automatic analysis with T-Scan, comparing CMC texts to school essays

CLS | Centre for Language Studies
Radboud University

Lexical measures

MTLD: measure of textual lexical diversity
= avg. length of sequential word strings in a text that maintain a TTR above a specified threshold (McCarthy & Jarvis 2010)
TTR = **type-token ratio** = no. of types [different words] / no. of tokens [total number of words]

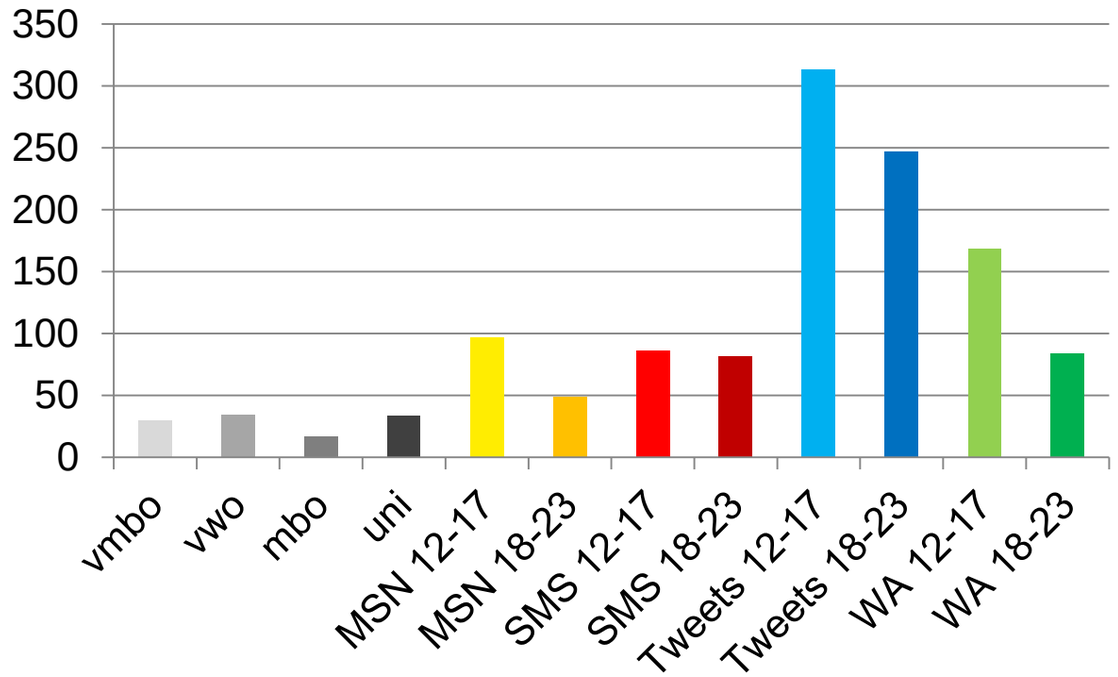
assumption: higher MTLD value → more lexical diversity → more different(ly spelled) words



result: CMC texts (*M* = 119.62, *SE* = 14.39) > school essays (*M* = 76.10, *SE* = 2.23)
t(10) = -2.08, ***p* < 0.05**
explanation: textisms, typos, misspellings in CMC
hypothesis: confirmed

Density of 'special words'
= no. of 'special words' (character strings T-Scan cannot recognize as words) per 1,000 words

assumption: higher density of 'special words' → more unrecognizable words → more differently spelled words + non-words

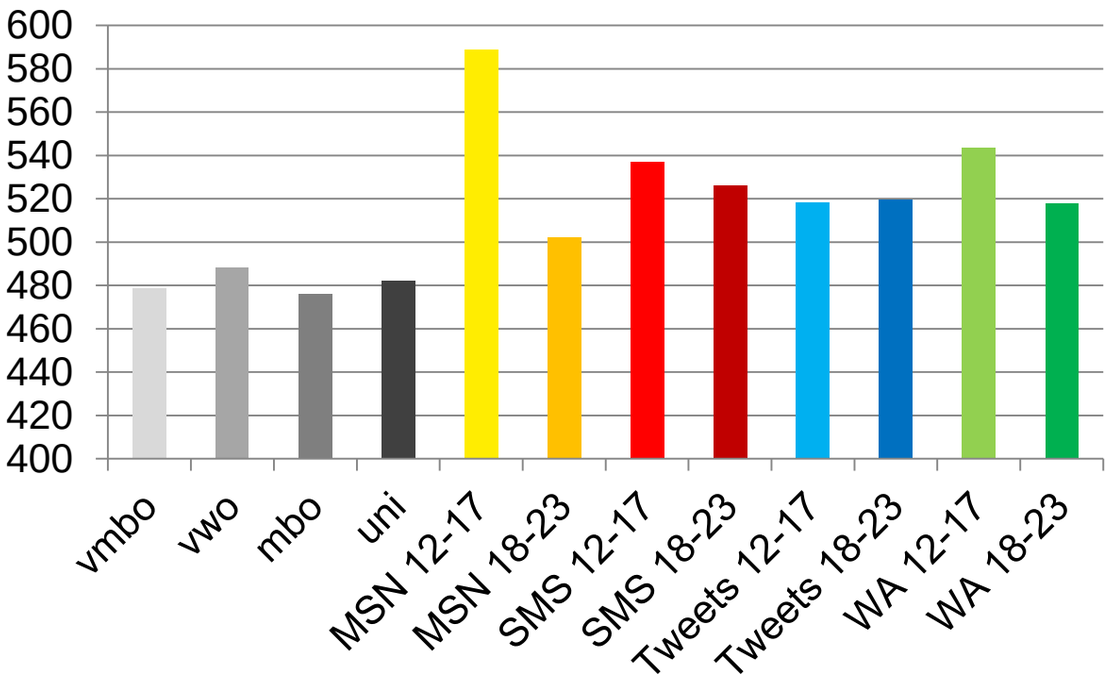


result: CMC texts (*M* = 140.77, *SE* = 33.20) > school essays (*M* = 28.58, *SE* = 4.02)
t(10) = -3.35, ***p* < .01**
explanation: textisms, typos, misspellings, URLs in CMC
hypothesis: confirmed

Lexical measures cont'd

Lexical density
= no. of content words (nouns, verbs, adjectives, adverbs) per 1,000 words (e.g. Johansson 2008)

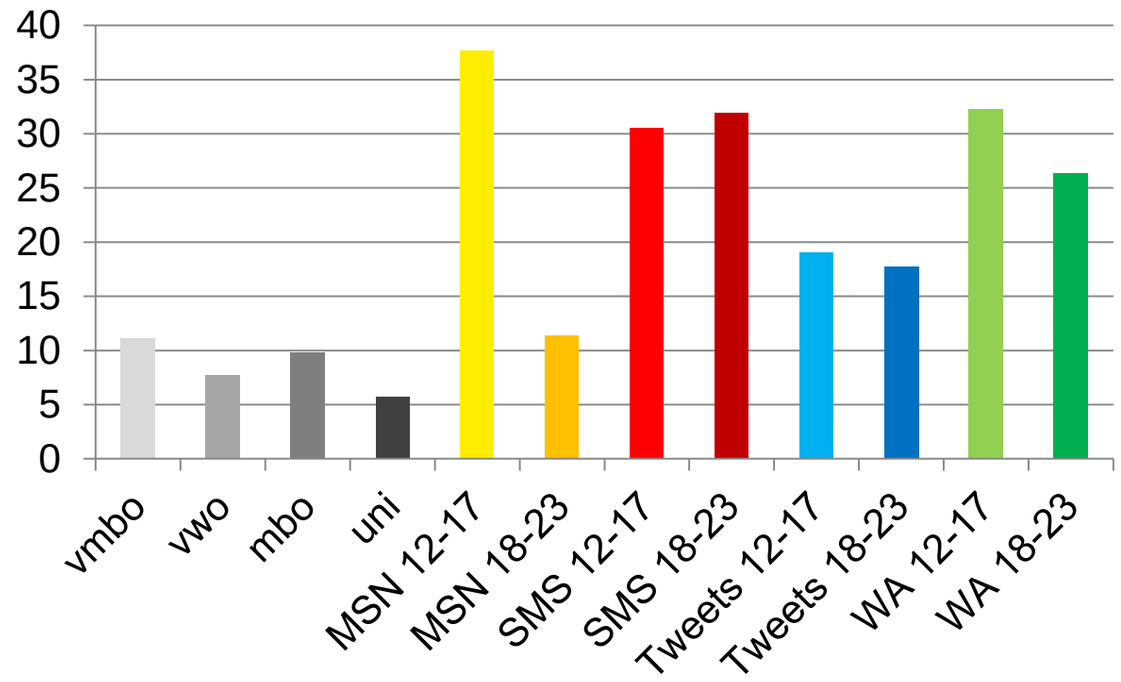
assumption: higher lexical density → more content words → fewer function words



result: CMC texts (*M* = 531.70, *SE* = 9.28) > school essays (*M* = 481.31, *SE* = 2.68)
t(10) = -3.71, ***p* < .01**
explanation: omission of function words in CMC
hypothesis: confirmed

Density of ellipses
= no. of finite verbs without a subject per 1,000 words

assumption: higher density of ellipses → fewer grammatical subjects

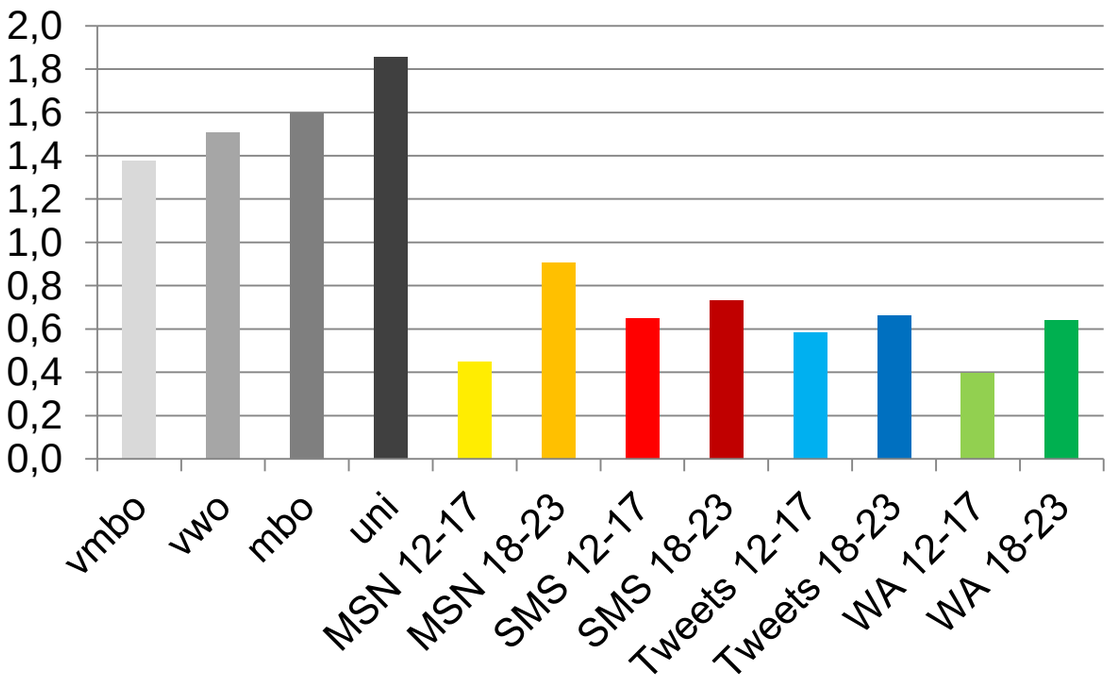


result: CMC texts (*M* = 25.86, *SE* = 3.17) > school essays (*M* = 8.60, *SE* = 1.18)
t(10) = -5.10, ***p* < .001**
explanation: omission of subjects in CMC
hypothesis: confirmed

Syntactic measures

Average of all dependency lengths
= avg. no. of words that need to be skipped from head to dependent per sentence
dependency length = distance between head (of sentence/phrase) and its dependent

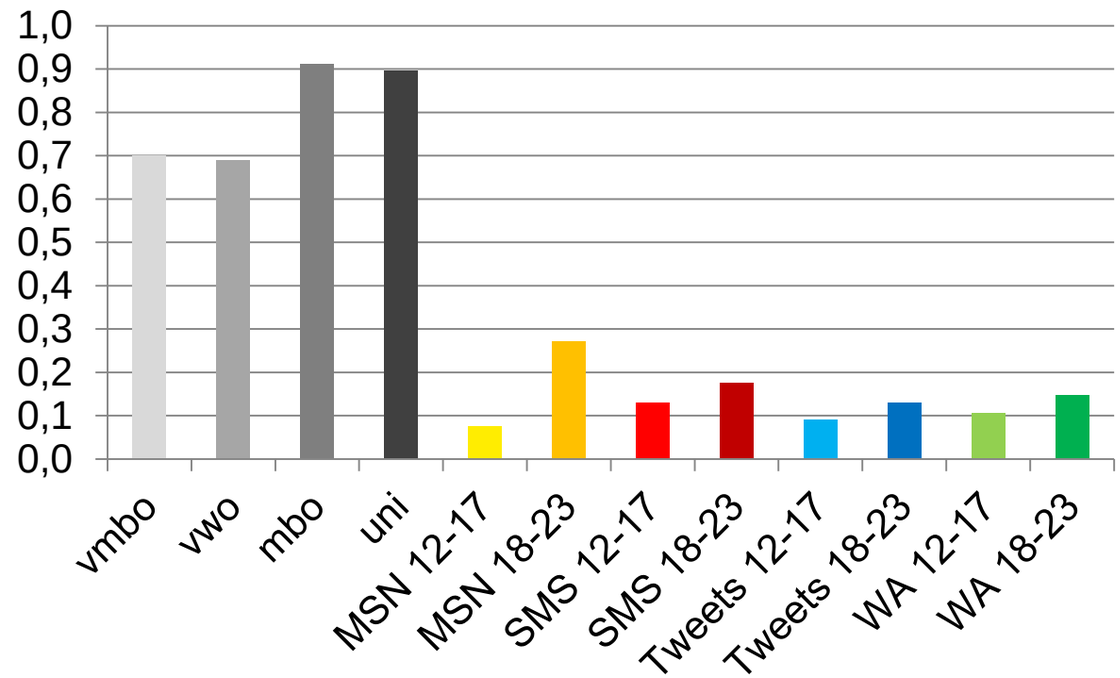
assumption: lower avg. of all dependency lengths → fewer discontinuous structures → less syntactic complexity (Gibson, 2000)



result: CMC texts (*M* = 0.63, *SE* = 0.06) < school essays (*M* = 1.59, *SE* = 0.10)
t(10) = 9.04, ***p* < .001**
hypothesis: confirmed

Average number of subordinate clauses
= avg. no. of subclauses (relative clauses, adverbial clauses, complement clauses, infinitival subclauses) per sentence

assumption: lower avg. no. of subclauses → less syntactic complexity

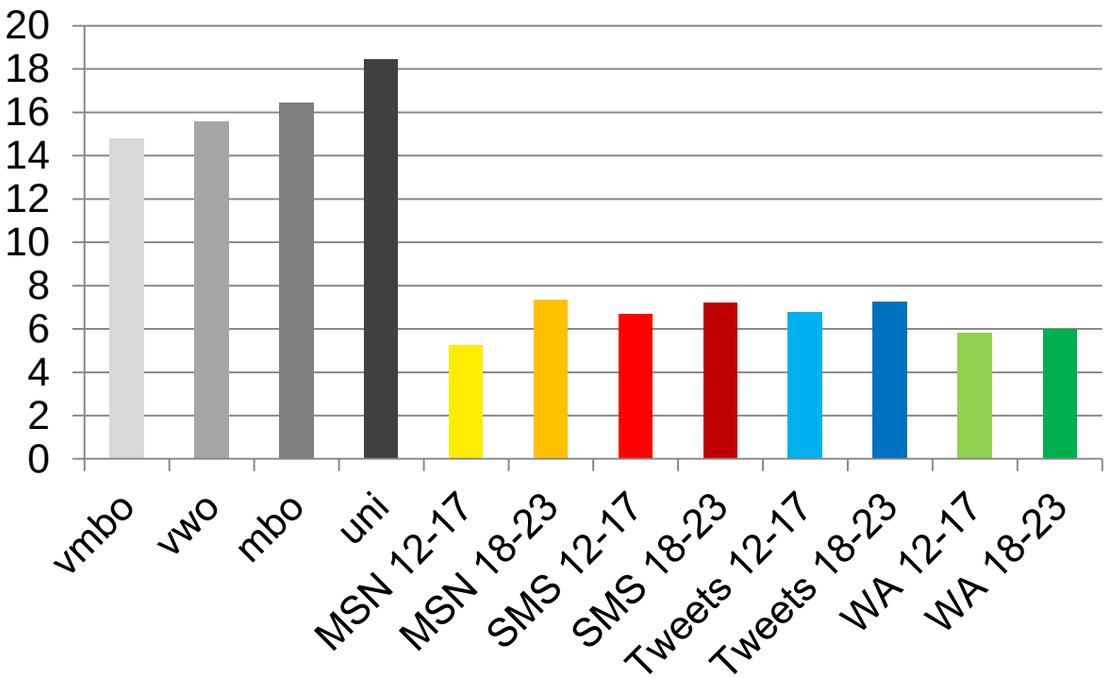


result: CMC texts (*M* = 0.14, *SE* = 0.02) < school essays (*M* = 0.80, *SE* = 0.06)
t(10) = 10.21, ***p* < .001**
hypothesis: confirmed

Syntactic measures cont'd

Average sentence length
= avg. no. of words per sentence

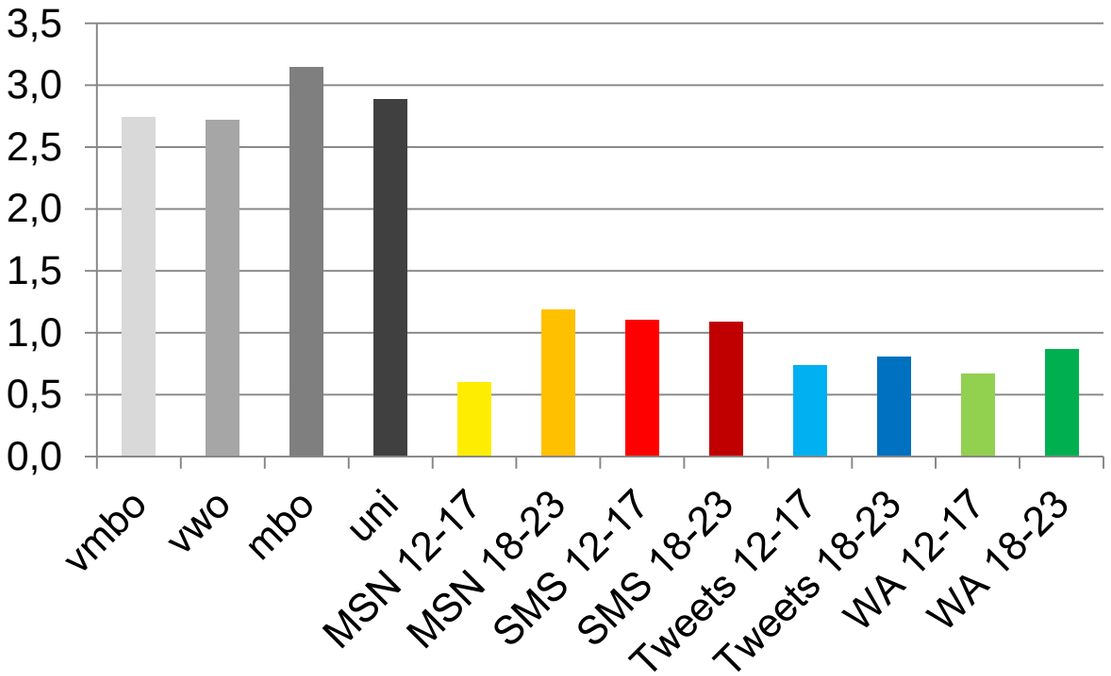
assumption: lower avg. sentence length → less syntactic complexity



result: CMC texts (*M* = 6.55, *SE* = 0.28) < school essays (*M* = 16.33, *SE* = 0.79)
t(10) = 14.76, ***p* < .001**
hypothesis: confirmed

D-level: developmental level
= based on classification and rank order of sentence types in eight increasingly complex developmental levels
(Rosenberg & Abbeduto 1987, Covington 2006)

assumption: lower D-level → less syntactic complexity



result: CMC texts (*M* = 0.88, *SE* = 0.08) < school essays (*M* = 2.87, *SE* = 0.10)
t(10) = 15.51, ***p* < .001**
hypothesis: confirmed

Conclusion

- Compared to school essays, written CMC:

  lexis > is more diverse, different, dense
  syntax > contains more omissions; is less complex

- Different **registers** ➜ informal CMC vs. more formal school writing

- **Hopeful results**: no great interference of CMC with youths' traditional writing skills after all...?

Future work

Corpus analysis

Next steps:

- Analyzing Facebook posts

Correlational study

RQ: Does youths' CMC use (intensity/manner) correlate with writing proficiency?

✓ Conducting surveys on CMC use at secondary & tertiary schools

✓ Collecting school writings of the same students

Next steps:

- Analysing school writings qualitatively/quantitatively
- Computing correlations between answers on surveys & school writings

Experimental study

RQ: Is there a causal connection between CMC use and literacy?

References

Cougnon, L.-A., & Fairon, C., Eds. (2014). *SMS Communication: A Linguistic Approach*. Amsterdam: John Benjamins.

Covington, M.A., He, C., Brown, C., Naçi, L., & Brown, J. (2006). *How Complex is That Sentence? A Proposed Revision of the Rosenberg and Abbeduto D-Level Scale*. CASPR Research Report 2006-01. University of Georgia: Artificial Intelligence Center.

Crystal, D. (2008). *Txtng: The Gr8 Db8*. Oxford: Oxford University Press.

Frehner, C. (2008). *Email - SMS - MMS: The Linguistic Creativity of Asynchronous Discourse in the New Media Age*. Bern: Peter Lang.

Gibson, E. (2000). The dependency locality theory: a distance-based theory of linguistic complexity. In Y. Miyashita, A.P. Marantz & W. O'Neil (Eds.), *Image, Language, Brain*. Cambridge: MIT Press, pp. 95–126.

Humphrys, J. (2007, September 24). I h8 txt msgs: How texting is wrecking our language. *Daily Mail*. http://www.dailymail.co.uk/news/article-483511/I-h8-txt-msgs-How-texting-wrecking-language.html.

Jacobs, G.E. (2008). People, purposes, and practices: Insights from cross-disciplinary research into instant messaging. In J. Coiro, M. Knobel, C. Lankshear, & D.J. Leu (Eds.), *Handbook of Research on New Literacies*. New York, NY: Routledge, pp. 469–490.

Johansson, V. (2009). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working Papers in Linguistics*, 53, 61–79.

McCarthy, P., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2), 381 – 392.

Oostdijk, N., Reynaert, M., Hoste, V., & Schuurman, I. (2013). The construction of a 500-million-word reference corpus of contemporary written Dutch. In P. Spyns & J. Odijk (Eds.), *Essential Speech and Language Technology for Dutch: Results by the STEVIN Programme*. Heidelberg: Springer, pp. 219–247.

Pander Maat, H., Kraf, R., van den Bosch, A., Dekker, N., van Gompel, M., Kleijn, S., Sanders, T., & van der Sloot, K. (2014). T-Scan: A new tool for analyzing Dutch text. *Computational Linguistics in the Netherlands Journal*, 4, 53–74.

Rosenberg, S., & Abbeduto, L. (1987). Indicators of linguistic competence in the peer group conversational behavior of mildly retarded adults. *Applied Psycholinguistics*, 8(1), 19–32.

Thurlow. C., & Brown, A. (2003). Generation txt? The sociolinguistics of young people's text-messaging. *Discourse Analysis Online*, 1.

Treurniet, M., & Sanders, E. (2012). Chats, tweets and SMS in the SoNaR corpus: Social media collection. In D. Newman (Ed.), *Proceedings of the First Annual International Conference on Language, Literature & Linguistics*. Singapore: Global Science and Technology Forum, pp. 268–271.

Verheijen, L. (2013). The effects of text messaging and instant messaging on literacy. *English Studies*, *94*(5), 582–602.

Thanks for your attention

Questions or comments?

lieke.verheijen@let.ru.nl

CLS | Centre for Language Studies
Radboud University