Linguistic Characteristics of Dutch Computer-Mediated Communication: CMC and School Writing Compared

Lieke Verheijen
Radboud University (Nijmegen, the Netherlands)
E-mail: lieke.verheijen@let.ru.nl

Abstract

Computer-mediated communication has become essential in many youths' lives. Because language in CMC frequently deviates from standard language norms, it is feared to harm youngsters' traditional literacy skills. To determine if and, if so, how social media affect their writing skills, we first need to establish how CMC actually differs from the standard language. This paper presents findings of a study comparing CMC texts and school essays by youths from the Netherlands. Linguistic analyses were done with T-Scan, software specifically designed for Dutch texts. A range of lexical measures (lexical diversity, 'special' words, lexical density, ellipses) and syntactic measures (dependency lengths, subordinate clauses, sentence length, D-level) were studied. Results reveal that in comparison to their school writings, Dutch youths' computer-mediated communication is syntactically less complex, contains more omissions, and is lexically more diverse, different, and dense. These youths thus employ different registers in the writing contexts of CMC and school.

Keywords: computer-mediated communication, social media, writing, register, literacy

1. Introduction

Most youths' daily lives are nowadays filled with computer-mediated communication. Instant messaging, texting, and other social media are essential for them to keep in touch with friends and family. In computer-mediated messages, it is key to communicate effectively, expressively, and informally. As a result, CMC writings frequently differ from standard language conventions (e.g. Thurlow & Brown, 2003; Crystal, 2008; Frehner, 2008; Cougnon & Fairon, 2014). Notable differences are nonstandard orthography and syntax, as in '*fyi i'll B @home l8er 2night, u OK with that? car broke down ☹*'. This sentence contains abbreviations, omissions, an emoticon, and lacks capitalisation and punctuation at the appropriate places. Such deviations in CMC from the 'official' language norms are a source of worry for many parents and language teachers: they fear it damages youths' traditional literacy skills.

2. Research Goals

This paper presents a study that is part of my PhD project into the impact of CMC on literacy. In order to determine whether and, if so, how youths' social media use affects their writings at school, it is imperative to first investigate what youths' CMC actually looks like and how it differs from the standard language. The main goal of this study is to explore in what ways the informal language used by Dutch youths in CMC differs from their more formal school writings. These questions were analysed by means of a manual analysis, as well as an automatic analysis; the present paper focuses on the latter.

3. Methodology

3.1 Materials

For my study into Dutch written CMC, I used a corpus of CMC texts by youths between 12 and 23 years old, with MSN chats, SMS, tweets, and WhatsApp chats. These social media represent four CMC genres: instant messaging with an internet application, text messaging, microblogging, and instant messaging with a mobile phone app. The first three genres were selected from SoNaR ('STEVIN Nederlandstalig Referentiecorpus'), a reference corpus of written Dutch (Treurniet & Sanders, 2012; Oostdijk et al., 2013). WhatsApp chats were gathered especially for the purposes of my project, via a website where youths could voluntarily donate their messages, http://cls.ru.nl/whatsapptaal/. Table 1 shows specifics of the CMC corpus. For comparison, I also collected school writings. These were written by youths of similar ages as the CMC texts, of different educational levels. Table 2 shows more details on the school essays.

Genre	Years of collection	Age group	# words	# chats or contributors
MSN	2009-2010	12-17 18-23	45,051 4,056	106 21
SMS	2011	12-17 18-23	1,009 23,790	7 42
Twitter	2011	12-17 18-23	22,968 99,296	25 83
WhatsApp	2015	12-17 18-23	55,865 140,134	11 / 84 23 / 132
total	2009-2015	12-23	392,169	

chats: MSN, WhatsApp; # contributors: SMS, Twitter, WhatsApp

Table 1: CMC texts.

Educational level	Years of production	Age group	# words	# texts
lower secondary (*vmbo*)	2013-2014	± 14-15, 3rd grade	50,143	128
higher secondary (*vwo*)	2013-2014	± 14-15, 3rd grade	50,070	153
lower tertiary (*mbo*)	2012-2014	± 17-18, 2nd grade	39,793	137
higher tertiary (*uni*)	2012-2014	± 18-19, 1st grade	50,175	169
total	2012-2014	± 14-19	190,181	587

Table 2: School essays.

3.2 Method

A quantitative corpus study was conducted. For the first part of the analysis, frequencies of several linguistic features were counted manually in the CMC texts. Yet this paper focuses on the second/automatic part of the analysis, comparing the CMC texts to school writings with T-Scan – software specifically designed for Dutch texts (Pander Maat et al., 2014). On the basis of theoretical considerations, a range of relevant lexical and syntactic measures were selected. It was hypothesized that CMC texts, compared to school essays, are lexically more diverse, different, and dense; contain more omissions; and are syntactically less complex. Independent *t*-tests were conducted to compute whether differences were significant; one-tailed probability values are reported here.

4. Results and Discussion

4.1 Lexical Analysis

The measure of textual lexical diversity (MTLD) is the average length of sequential word strings in a text that maintain a type-token ratio (TTR) above a specified threshold (McCarthy & Jarvis, 2010). The MTLD depends on the TTR, which is calculated by dividing the number of types (different words) by the number of tokens (total number of words). Although the TTR is a classic measure, the MTLD is more reliable, because it is insensitive to text length. A higher MTLD value indicates more lexical diversity: more different words or *differently spelled* words. On average, the CMC writings had a higher lexical diversity (M = 119.62, SE = 14.39) than the school writings (M = 76.10, SE = 2.23), $t(10) = -2.08$, $p < 0.05$. Figure 1 shows that the MTLD was higher in the CMC texts, with the exception of WhatsApp chats by 12-17-year-olds.[1] The higher lexical diversity depends on the orthographic variation in written CMC, due to textisms (unconventional spellings, deviating from the standard language norms), misspellings ('errors', as judged by linguistic prescriptivists), and typos (incorrect key presses or false predictions by predictive software). This confirms the hypothesis that CMC is lexically more diverse.

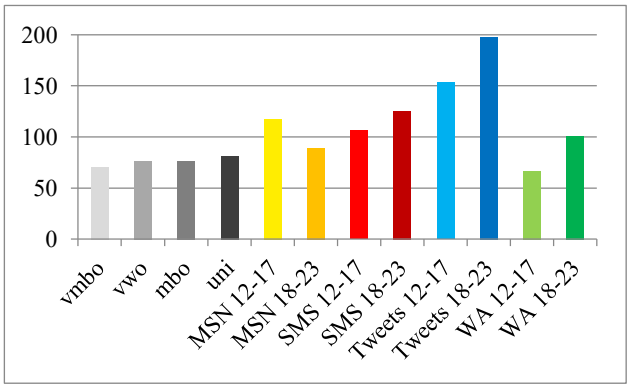

Figure 1: Measure of textual lexical diversity (MTLD).

[1] This apparent exception can be attributed to the frequent repetition of chain messages and certain words in a spam-like manner by one contributor; excluding this outlier, the MTLD would be 92.70 – higher than the school essays, as hypothesized.

T-Scan computes the density of 'special words', measured per one thousand words. This includes names, loanwords, numbers, Roman numerals, and times. On average, the CMC writings had a higher density of 'special words' (M = 140.77, SE = 33.20) than the school writings (M = 28.58, SE = 4.02), $t(10) = -3.35$, $p < .01$. Figure 2 illustrates this and shows that there is much variation between CMC genres. The greater frequency of 'special words' is because of textisms, misspellings, typos, and URLs in CMC – character strings that T-Scan cannot recognize as words, since they deviate orthographically from Standard Dutch and are not listed in any standard dictionaries. Tweets in particular include many URLs and 'words' of the format *@username*, within messages in response to another user's tweet (replies) or messages directed at another user (mentions). This higher density endorses the hypothesis that CMC is lexically more different from the standard language.

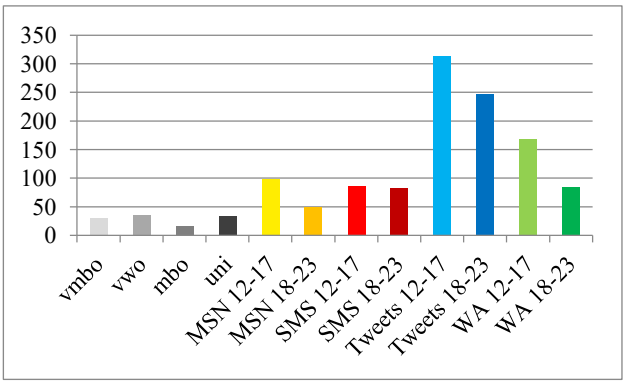

Figure 2: Density of 'special words'.

The third lexical measure that was selected is lexical density. This is the number of content words (nouns, verbs, adjectives, and adverbs) per one thousand words (e.g. Johansson, 2008). When a text has a high lexical density, it contains many content words and few function words. On average, the CMC writings had a higher lexical density (M = 531.70, SE = 9.28) than the school writings (M = 481.31, SE = 2.68), $t(10) = -3.71$, $p < .01$, as shown in Figure 3. This is due to the frequent omission of function words in CMC, which is known for its concise writing style, somewhat similar to that of telegrams or newspaper headlines. The findings from T-Scan thus support the hypothesis that CMC is lexically denser.

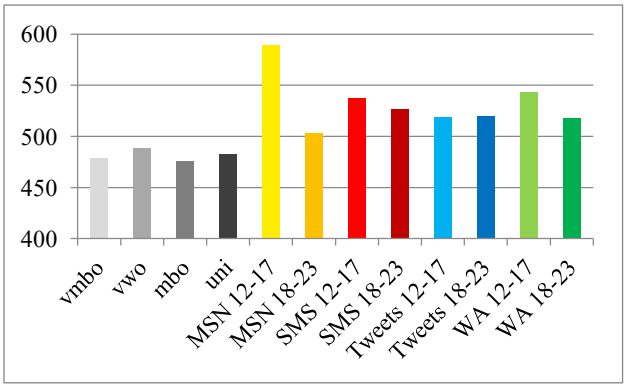

Figure 3: Lexical density.

Another interesting measure is the density of elliptical constructions, quantified as the number of finite verbs without a subject per one thousand words. On average, the CMC writings had a higher density of ellipses (M = 25.86, SE = 3.17) than the school writings (M = 8.60, SE = 1.18), $t(10)$ = -5.10, p < .001. Figure 4 shows that the CMC writings of all genres contained more elided subjects (though just barely for MSN chats by 18-23 year olds). This backs up the abovementioned results on lexical density: informal written CMC contains fewer function words than formal school essays, at least partly due to the frequent omission of grammatical subjects.

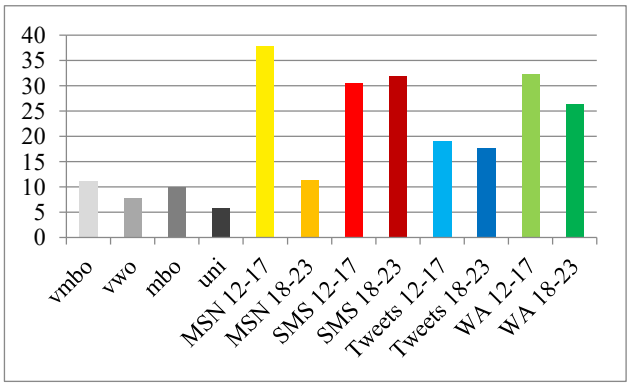

Figure 4: Density of ellipses.

4.2 Syntactic Analysis

One measure of syntactic complexity is the average of all dependency lengths per sentence. The dependency length is the distance between a head (of a sentence or phrase) and its dependent, such as a finite verb and the subject or an article and the corresponding noun. T-Scan expresses the distance in number of words that need to be skipped from head to dependent. Texts with a higher average dependency length contain more discontinuous structures, making them syntactically more complex and more difficult to process for readers (Gibson, 2000). On average, the CMC writings had a lower average of all dependency lengths per sentence (M = 0.63, SE = 0.06) than the school writings (M = 1.59, SE = 0.10), $t(10)$ = 9.04, p < .001. It is clear from Figure 5 that the CMC texts of all genres had lower average dependency lengths, no matter what the writer's age or educational level. This supports the idea that CMC is syntactically less complex.

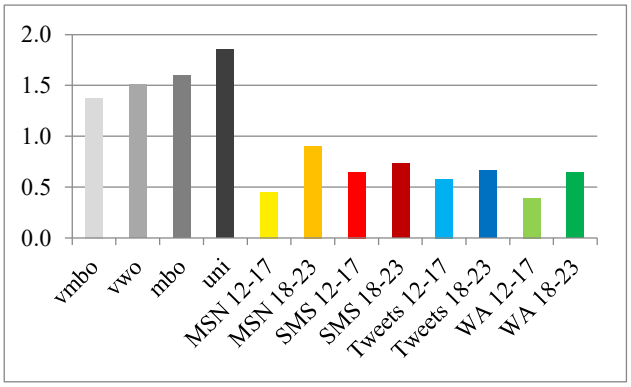

Figure 5: Average of all dependency lengths per sentence.

T-Scan also measures the average number of subordinate clauses per sentence. It includes both finite (relative, adverbial, and complement clauses) and infinitival subclauses. A higher density of subclauses is indicative of greater syntactic complexity. On average, the CMC writings had a lower average no. of subordinate clauses per sentence (M = 0.14, SE = 0.02) than the school writings (M = 0.80, SE = 0.06), $t(10)$ = 10.21, p < .001. Figure 6 clearly shows that the CMC texts overall contained fewer subordinate clauses. Again, the lower syntactic complexity of CMC is confirmed by T-Scan.

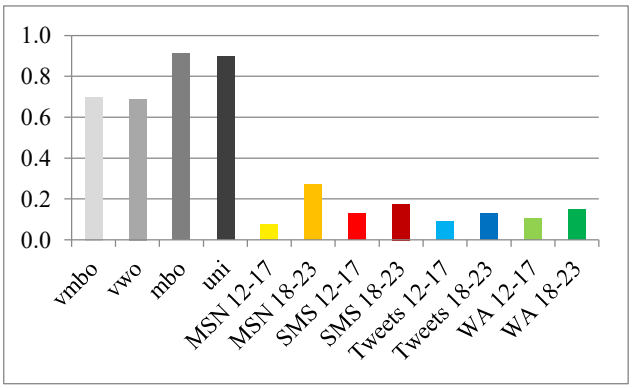

Figure 6: Average no. of subordinate clauses per sentence.

Another complexity measure provided by T-Scan is the average sentence length, which is measured in number of words. A higher average sentence length indicates more syntactic complexity. On average, the CMC writings had a lower average sentence length (M = 6.55, SE = 0.28) than the school writings (M = 16.33, SE = 0.79), $t(10)$ = 14.76, p < .001. Figure 7 shows that the texts of all four CMC genres contained much shorter sentences than the school essays, irrespective of the writer's educational level or age. Once more, the hypothesis is confirmed.

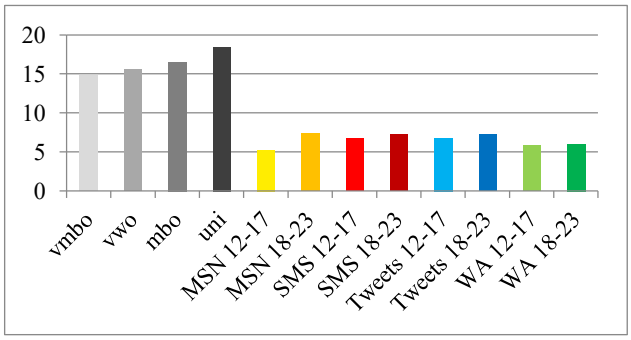

Figure 7: Average sentence length.

A final relevant syntactic measure is the so-called D-level. The D-level of a text is determined on the basis of a classification and rank order of sentence types in eight increasingly complex developmental levels, in the order in which children learn these constructions (Rosenberg & Abbeduto, 1987; Covington, 2006). The assumption is that a higher D-level value suggests more syntactic complexity. On average, the CMC writings had a lower D-level (M = 0.88, SE = 0.08) than the school writings (M = 2.87, SE = 0.10), $t(10)$ = 15.51, p < .001. The CMC texts of all four genres had lower D-levels, as can be seen

in Figure 8. This result is in line with the proposed hypothesis on syntactic complexity.

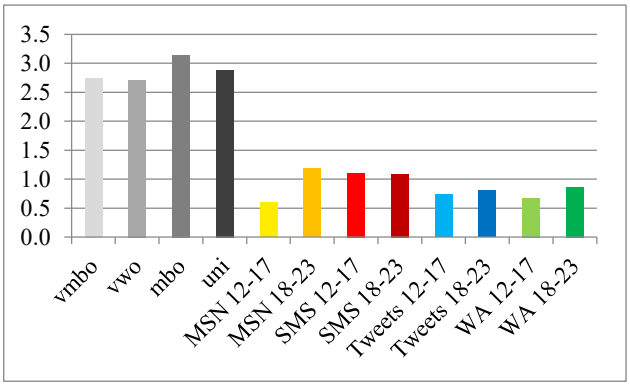

Figure 8: D-level.

5. Conclusion

To conclude, the lexical and syntactic analysis of CMC texts of four social media support my hypothesis: in comparison to school writing, CMC is lexically more diverse, different, and dense, while syntactically it contains more omissions and is less complex. This proves that Dutch youths in secondary and tertiary education employ a different register in informal computer-mediated communication than in texts written in more formal settings. These results are hopeful: perhaps deviations from the standard language in youngsters' CMC do not cause great interference with their traditional writing skills after all – they might be quite capable of keeping the registers separate, as societal norms expect them to do.

6. Future Work

A limitation of the present study is that the materials compared here, i.e. CMC discourse and texts written at school, were not produced by the same writers. In addition, they have been collected over a relatively long time span, of six years. For a more precise answer to the question if and, if so, how CMC use affects school writing, I plan to conduct research in which (a) social media data and school texts of the same students are collected and analysed and (b) additional information about writers' use of CMC and social media (in terms of frequency/intensity) are gathered through surveys. Future work will include one more genre, namely posts from the social networking site Facebook. Furthermore, it unfortunately exceeded the scope of this paper to closely examine variation between texts of different genres, educational levels, ages; this may also be explored further. Still, this study can serve as a fruitful basis for analyses on the impact of written computer-mediated communication on young people's literacy skills.

7. Acknowledgements

This study is part of a research project funded by the Dutch Organisation for Scientific Research (NWO), project number 322-70-006. I would like to thank Wilbert Spooren and the anonymous reviewers for their useful comments on previous versions of this paper.

8. 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.

Johansson, V. (2008). 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.