Article

Written Communication
2020, Vol. 37(3) 365–400

DOI: 10.1177/0741088320917921
journals.sagepub.com/home/wcx
SAGE

Lexical Patterns in Adolescents' Online Writing: The Impact of Age, Gender, and Education

Lisa Hilte[1], Walter Daelemans[1], and Reinhild Vandekerckhove[1]

Abstract

This article examines the impact of the sociodemographic profile (including age, gender, and educational track) of Flemish adolescents (aged 13–20) on lexical aspects of their informal online discourse. The focus on lexical and more "traditional," print-based aspects of literacy is meant to complement previous research on sociolinguistic variation with respect to the use of prototypical features of social media writing. Drawing on a corpus of 434,537 social media posts written by 1,384 teenagers, a variety of lexical features and related parameters is examined, including lexical richness, top favorite words, and word length. The analyses reveal a strong common ground among the adolescents with respect to some features but divergent writing practices by different groups of teenagers with regard to other parameters. Furthermore, this study analyzes both standardized versions of social media messages and the original utterances (including nonstandard markers of online writing). Strikingly, different results emerge with respect to adolescents' exploitation of more traditional versus digital literacy skills in relation to their sociodemographic profile, especially with respect to sentiment expression (verbal versus typographic/pictorial). The study suggests that the inclusion of nonverbal communicative strategies, for instance in language teaching,

[1]Department of Linguistics, University of Antwerp, Antwerp, Belgium

Corresponding Author:
Lisa Hilte, Department of Linguistics, University of Antwerp, Antwerp, Belgium.
Email: lisa.hilte@uantwerpen.be

might be a pedagogical asset, since these strategies are eagerly adopted by teenagers who show proof of less developed traditional writing skills.

Keywords
social media writing, adolescent literacy, sentence length, word length, lexical richness, sentiment analysis, topic analysis

Informal online writing tends to deviate from more formal, school-based writing in various ways, for example through nonstandard spelling, grammar, and lexicon, and through the use of typographic markers such as emoji. While previous studies have examined these prototypical markers of online discourse (e.g., Hilte et al., 2018b, in press; Varnhagen et al., 2010; Verheijen, 2015), more "traditional" linguistic variables and patterns—for example, patterns regarding average sentence length—are less prominent in research on computer-mediated communication (CMC). The present study aims to fill that gap through analysis of lexical variables and related patterns in adolescents' online writing. This focus is motivated by the fact that typographic or pictorial CMC markers can take over the function of lexical items to a certain extent: for example, while social and emotional involvement can be expressed lexically, emoticons and emoji (i.e., typographic versus pictorial markers) may serve the same purpose.[1] So teenagers tend to have both a "traditional" (verbal/lexical) and "digital media" (e.g., typographic/pictorial) repertoire at their disposal for informal online communication. However, it has hardly been investigated to what extent they use both repertoires and whether their preferences in this respect are influenced by social variables. Previous research has indicated that teenagers' production of CMC markers is significantly impacted by multiple aspects of their sociodemographic profile (Hilte et al., 2018b, in press; Varnhagen et al., 2010; Verheijen, 2015). Building on this growing body of scholarship, we investigate whether five more traditional linguistic properties of their social media texts are impacted by these social variables too, and whether divergent writing patterns emerge for adolescents with different profiles (in terms of age, gender, and educational track).[2]

Findings from this study suggest ways to improve not only our insight into youths' online writing practices but also our understanding of their traditional versus digital literacies. While *digital* literacy tends to concern the use of new media-related communicative markers (which may be nonverbal, such as emoji), other linguistic properties, such as the production of long or lexically rich sentences, may be indicative of a strong *traditional* literacy.[3] We note that traditional literacy—that is, a "traditional notion of literacy," consisting in

"reading and writing print-based texts" (Verheijen, 2018, p. 21)—is also referred to as *conventional*, *classical*, *print-based*, or *verbal* literacy (see Verheijen, 2018, pp. 21–32 for definitions and approaches of "new" and "old" literacies). For today's youths, who are sometimes referred to as "digital natives" (Frey & Glaznieks, 2018), both types of literacy have become an integral part of their reading and writing experience (Tomita, 2009, pp. 189–190; Verheijen, 2018, p. 302). Therefore, the analysis of these two repertoires may reveal complementary patterns, with youths favoring one repertoire over the other, or, conversely, a correlation between (strong) digital and traditional literacy skills.

The article is structured as follows: It first reviews related research, and then presents the corpus and participants. Next, we introduce the linguistic variables along with the methodology for feature extraction and the linguistic and statistical analyses, and we discuss the normalization strategy that was required for the present study. Finally, we present and discuss our findings.

Related Research

The linguistic characteristics of informal CMC have been widely investigated. A range of sociolinguistic studies demonstrate how people with distinct sociodemographic profiles (e.g., in terms of age or gender) appear to favor certain prototypical CMC markers to different extents (De Decker & Vandekerckhove, 2017; Hilte et al., 2018b, in press; Varnhagen et al., 2010; Verheijen, 2015). Some pioneering work suggests that social media writing has distinct properties for more traditional linguistic features as well: For example, Verheijen (2016) compared Dutch youths' social media writing and school writing with respect to several lexical and syntactic measures and concluded that CMC writing is syntactically less complex (e.g., including shorter sentences) than school writing (p. 68), but lexically more dense (i.e., containing more content words, which may be caused by the "frequent omission of function words in CMC," p. 67). As for social patterns regarding the traditional language markers in CMC that are included in this article, no consensus emerges. We discuss a number of studies below, all of which concern English CMC unless specified otherwise.

First of all, with respect to sentence length, Lin reports that adult males produce longer sentences in chat room conversations than adult females (2007, pp. 20–21). However, for *adolescent* authors, she observes the opposite tendency, that is females producing longer sentences (2007, pp. 20–21)—a tendency that is confirmed for both writing and speaking (Finlay, 2014, p. 25; Newman et al., 2008, p. 213). We note that in these studies, gender is operationalized as a binary variable. The same holds for the articles discussed below

(but see the Data section for a discussion of alternative operationalizations). With respect to age effects, Finlay (2014, pp. 24, 26) observes an increase in post length by age for online comments, although it should be noted that the groups of participants aged 12–16 and 17–18—which are closest to the target groups in the present article—produce posts of similar lengths. As for average word length, consistent gender findings are reported, with males producing longer words in both spoken and written (chat) conversations (Lin, 2007, pp. 21, 25; Mehl & Pennebaker, 2003, p. 865; Newman et al., 2008, pp. 213–214, 223).

Lexical richness measures capture people's active vocabulary size and the degree of (non)repetition in their language use (Malvern & Richards, 2012, p. 1).[4] With respect to this linguistic variable, conflicting gender patterns are attested. A larger vocabulary range is reported in male adolescents' chat room conversations (Lin, 2007, pp. 21, 25). Some research confirms this pattern for spoken conversations (Singh, 2001, p. 260), but other studies reveal no significant gender differences, for example, in English speaking and writing tasks for nonnative speakers (Yu, 2009, p. 253). In formal nonconversational writing tasks, both more theoretically educated and older youths have been found to produce lexically richer (Dutch) texts than their less theoretically educated and younger peers, respectively (Verheijen & Spooren, 2017, p. 9). Concerning the diverging results suggested by these studies, there are notable differences with respect to the quantification of lexical richness. This complicates comparison, as "different measures may well produce very different, sometimes even conflicting results" (Yu, 2009, p. 241). We discuss these measures in the section on Linguistic Variables below.

Two major points of reference for the analysis of authors' top favorite words (and the associated topics) in CMC are the studies on English Facebook messages and blog posts conducted by Schwartz et al. (2013) and Argamon et al. (2009). With respect to gender, their results reveal that many of the female authors' most prominent words relate to personal life and relationships (e.g., *boyfriend*, *mom*, *bestie*) (Argamon et al., 2009, p. 121; Schwartz et al., 2013, p. 8). In addition, typically "female" words or word combinations often express enthusiasm (e.g., *yay*, *soooo excited*) or a positive evaluation or sentiment (e.g., *wonderful*, *amazing*) (Argamon et al., 2009, p. 121; Schwartz et al., 2013, p. 8). Finally, some prominent lexemes used by women reveal more stereotypical female topics (e.g., *chocolate*, *shopping*, *my hair*) (Schwartz et al., 2013, p. 8). Many of the male authors' most prominent words concern politics (e.g., *government*, *democracy*) and fighting (e.g., *fight*, *battle*) (Schwartz et al., 2013, p. 8). Swear words frequently occur among the top "male" lexemes as well (e.g., *fuck*, *shit*) (Schwartz et al., 2013, p. 8). Finally, some prominent lexemes used by men reveal more

stereotypical male topics, such as technology (e.g., *system*, *software*), gaming (e.g., *xbox*, *ps3*), and football (e.g., *football*, *team*) (Argamon et al., 2009, p. 121; Schwartz et al., 2013, p. 8). With respect to age-related lexical variation, previous research indicates that among teenagers, school-related words (e.g., *homework*, *math*) and words expressing a mood (e.g., *bored*) are prominent, whereas the online discourse of slightly older groups contains more words about social life and partying (e.g., *drunk*, *bar*) as well as lexemes referring to studying (e.g., *professor*, *campus*)—for college students—or to work (e.g., *office*, *job*)—for young adults in their 20s (Argamon et al., 2009, pp. 121–122; Schwartz et al., 2013, p. 10).

Finally, with respect to sentiment expression in CMC, consistent age and gender patterns are reported in previous work. Girls/women appear to use more emotionally expressive language than boys/men, and expressive language tends to decrease with age—this age difference already manifests itself during adolescence, with younger teenagers writing in more expressive ways than older teenagers. These tendencies hold for both the use of emotion words, that is, *lexical* expressiveness, and for *typographic* expressiveness (see the section on Normalized versus Nonnormalized Texts below), in offline as well as online communication (Baron, 2008, p. 51; Hilte et al., 2018b for Flemish texts; Newman et al., 2008, pp. 223, 229; Schwartz et al., 2013, p. 9). As for educational variation, youths in practice-oriented tracks (e.g., vocational tracks, focused on the acquisition of technical/manual skills—see the Data section) appear to use more typographic emotional markers in their informal CMC than their peers in more theoretical tracks, too (see Hilte et al., 2018a, 2018c, in press for Flemish CMC).

While related research reveals interesting tendencies concerning the lexical patterns and related parameters included in this article, these variables are quite seldom systematically included in research on informal CMC (with some notable exceptions, e.g., Lin, 2007). Furthermore, the use of written lexical/verbal expression is seldom set off against the exploitation of typographic means of expression that often mark online communication. The present study aims to fill that gap by conducting a corpus analysis of Flemish youths' CMC. We specifically compare online messages produced by different sociodemographic groups of teenagers (in terms of age, gender, and educational track), and investigate whether different patterns can be observed for these groups with respect to average post and word length, lexical richness, sentiment expression, and top favorite words. One might expect older and more theoretically educated teenagers to display more developed traditional literacy skills compared to their younger and more practice-oriented peers, due to a longer familiarity with or a stronger educational focus on these skills. However, informal interactional CMC is characterized by a distinctive set of

principles or maxims (see Androutsopoulos, 2011, p. 149). We believe it is therefore worth investigating whether these expected age- and education-related patterns concerning traditional literacy actually emerge in social media discourse, and whether gender is a shaping factor too. But finally and most of all, we want to investigate whether there is an interaction between the exploitation of traditional versus digital literacy skills and whether that interaction is different for students with distinct sociodemographic profiles.

Data

The corpus analyzed in this study contains 434,537 social media posts (over 2.5 million tokens) written by 1,384 teenagers. The teenagers all live in Flanders (i.e., Dutch-speaking Belgium) and are secondary school students between 13 and 20 years old.[5] The posts are private instant messages produced on Facebook Messenger and WhatsApp in Dutch, which is most of the participants' first language, and the official language (of education) in Flanders. The vast majority of the tokens (87%) was produced between 2015 and 2016. Dialect region is a quasi-constant, as 96% of the teenagers live in the central Flemish province of Antwerp. We do not have information about the device (e.g., computer or smartphone) on which the messages were produced.

The dataset was collected in a school context: We visited several secondary schools in the province of Antwerp and invited students to voluntarily donate private social media messages that were written outside of the school context and prior to our visits. The latter condition was meant to exclude the *observer's paradox*, which Labov (1972, p. 209) famously described as follows: "The aim of linguistic research in the community must be to find out how people talk when they are not being systematically observed; yet we can only obtain these data by systematic observation." We extensively informed the students about the project by giving a class on online writing, and we guaranteed anonymization of the data, which helped to engage them. Furthermore, we asked the students who donated conversations (and if they were minors, their parents/guardians too) to fill in a consent form to grant us permission to store and linguistically analyze their anonymized texts.

Three aspects of the teenagers' sociodemographic profile are included in the research design as independent variables: their age, gender, and educational track (see Table 1 for an overview of the distributions in the corpus). This meta-information was provided by the participants themselves. For age, we distinguish between younger teenagers (13–16 years old) and older teenagers or young adults (17–20 years old). Age is treated as a categorical rather than as a continuous variable, since related research suggests that

Table 1. Distributions in the Corpus.

Variable	Variable levels	Tokens	Participants
Educational track	General Secondary Education	739,831 (29%)	596 (43%)
	Technical Secondary Education	1,151,684 (46%)	393 (28%)
	Vocational Secondary Education	639,839 (25%)	395 (29%)
Gender	Female	1,696,517 (67%)	717 (52%)
	Male	834,837 (33%)	667 (48%)
Age	Younger teenagers (13–16)	1,360,898 (54%)	1,234[a]
	Older teenagers / young adults (17–20)	1,170,456 (46%)	897
Total		2,531,354	1,384

a. The number of younger and older participants adds up not to the total number of participants but to a higher number (therefore, we did not add percentages for age). Participants can occur in the corpus at different age points if they submitted recent chat conversations as well as older ones. We control for these repeated observations in the data by adding participant as a random effect in the statistical models (see the Method section below).

adolescents' nonstandard language use does not evolve linearly, but "peaks" mid-puberty (around the age of 16)—a phenomenon that is referred to as the *adolescent peak* (e.g., Holmes, 1992, p. 184). Nonstandard language use, as well as linguistic creativity and innovation, is said to be highest during adolescence (see, e.g., Eckert, 1997, p. 163; Holmes, 1992; Tagliamonte, 2016), due to peer "group pressure to not conform to established societal conventions" (Nguyen et al., 2016, p. 17). But adolescence is no homogeneous linguistic period: as teenagers age, their language use typically converges to the adult standard. Since for adults "social advancement matters [. . .], they use [linguistic varieties closer to the] standard language to be taken seriously" (Nguyen et al., 2016, p. 17). Finally, we note that a distinction between two similar age groups is often made in related work (see, e.g., Verheijen, 2018).

In the data collection phase of the project, students were asked to mark whether they were male or female; the study was thus limited to a binary approach to gender. We realize, however, that nonbinary approaches to gender identity might lead to more nuanced findings and interpretations, since "choosing [a] binary opposition as a starting point constrains the set of possible conclusions" (Bamman et al., 2014, p. 138). Bamman et al. (2014) warn for additional risks when operationalizing gender, such as the unintended integration of particular assumptions about gender in the research design (e.g., by focusing on people with specific gender identities during data collection) or the disregard of the role of gender in a larger personal identity and thus the potential interaction with other aspects of this identity (pp. 137–138).

In the present study, such interactions have been included in the research design (see below) in an effort to account for the complex relationship between gender and linguistic features of adolescents' informal online writing. For a more comprehensive discussion of gender and writing research, see Peterson and Parr (2012) and Lillis et al. (2018).

The final social variable in this study is educational track. All participants attend one of the three main types of Belgian secondary education. These range from the theory-oriented General Secondary Education, where students are prepared for higher education, to the practice-oriented Vocational Secondary Education, where students are taught a specific, often manual, profession. The Technical Secondary Education holds an intermediate position on this continuum, with a practical and theoretical orientation, and a focus on technical courses (Flemish Ministry of Education and Training, 2018, p. 10). An educational difference that may be of particular importance in the present study concerns the approach with respect to standard Dutch proficiency and formal writing. While the learning of correct and formal standard Dutch writing is an objective in all school systems, there are some important differences. One distinction concerns the educational setting or "infrastructure": Dutch, that is, the official language (of education) in Flanders, is instructed as a main course on its own in General and Technical Education, but is integrated in two applied courses in the Vocational track (Vlaams Verbond van het Katholiek Secundair Onderwijs [VVKSO], 2006, p. 5). While for Vocational students the main goal is to improve their practical language proficiency (VVKSO, 2006, pp. 7–8), General and Technical students are not only expected to improve their proficiency, but also their meta-linguistic skills (VVKSO, 2014, pp. 5–6). Furthermore, while the communicative function of language prevails in Vocational Education, a wider range of functions (including the cultural and expressive function of language) is focused on in the more theoretical tracks (VVKSO, 2014, p. 10). With respect to writing proficiency, functional writing is deemed crucial for teenagers in the Vocational track (VVKSO, 2006, p. 17), but unlike their more theory-oriented peers they hardly get any instruction on the writing process itself (e.g., in terms of structuring and revising their texts) (VVKSO, 2014, p. 46). From these distinctions, we hypothesize that a potentially more limited proficiency in formal standard writing might also influence practice-oriented students' social media writing.

Linguistic Variables and Methodology

In this section, the linguistic variables are presented. In addition, we discuss the challenges with regard to the "noisy" (nonstandard) nature of the social

media texts, and the applied normalization procedure. Finally, we present the methodology and statistical models used in this study.

Linguistic Variables

In order to obtain a nuanced view of lexical variation and related matters, a variety of features were examined. First of all, each author's (productive) *lexical richness* was measured, which "summarizes the range of vocabulary and the avoidance of repetition in the sample" (Malvern & Richards, 2012, p. 1). We operationalized lexical richness as the Guiraud correction of the type/token ratio. Type/token ratio (TTR), that is, the number of *different* words used by an author (types) divided by *all* the words he or she used (tokens), is the most widely applied implementation of this concept (Vermeer, 2000, p. 66). However, it is heavily criticized, as its outcome may be unreliable when samples of different lengths are compared (Van Hout & Vermeer, 2007, p. 121; Yu, 2009, p. 239). The measure is "notorious for being sensitive to sample size" (Yu, 2009, p. 239): Since an increase in sample size generally implies a stronger increase in number of tokens than types, the average TTR drops as samples grow larger (Malvern & Richards, 2012, p. 2; Vermeer, 2000, p. 68; Yu, 2009, p. 239). A simple transformation of the TTR that reduces the influence of sample size consists in dividing the number of types by the square root of the number of tokens in a sample. This *Guiraud TTR* (also *root TTR*) is considered a more adequate measure of lexical diversity, holding a more constant value for increasing sample sizes (Vermeer, 2000, p. 68). For an overview and evaluation of different approaches, see, for example, Malvern and Richards (2012), and Van Hout and Vermeer (2007). Finally, we note that in the present study, lexical diversity was calculated for the nonlemmatized tokens (e.g., *loop* "run" and *liep* "ran" are counted as two different types, and not as two occurrences of the same lemma—the canonical/dictionary form or "base form"—that is, the lemma *lopen* "to run").

The next variable concerns the authors' top favorite words. We automatically extracted the 500 most frequent words per subgroup of participants (e.g., girls versus boys) and manually inspected these (after the exclusion of function words) in order to discover the associated topics. However, as "direct association of word types with high-level dimensions remains problematic" (Bamman et al., 2014, p. 145), the topics that have been assigned to the lexemes should be interpreted as suggestive rather than absolute labels.

The third and fourth dependent variables are the authors' average token and post length, expressed in number of characters and number of tokens, respectively.[6] We note that after normalization of the corpus, a token always represents a word (and not, for example, an emoticon—see the section below

on Normalization of the Data). The production of longer tokens thus equals the production of longer words, and might be indicative of a stronger command of more complex words. The production of longer posts (i.e., instant messages) equals the production of utterances consisting of more words and may indicate a stronger lexical expression or orientation.

The final variable is, just like post length, an utterance-level rather than a (single-)token-level feature. For each post in the corpus, the lexical sentiment expression was measured, using the sentiment function in the Pattern package for Python (De Smedt & Daelemans, 2012a). The function uses a lexicon of Dutch adjectives annotated for polarity, subjectivity, and intensity (see De Smedt & Daelemans, 2012a, p. 2066, and 2012b, p. 3568). It assigns scorings to an utterance by calculating the average of the individual words' scores.[7] Modifiers such as negations (which inverse polarity) are taken into account as well (De Smedt & Daelemans, 2012b, p. 3571). The *polarity score* expresses how negative or positive an utterance is, ranging from −1 (very negative) to +1 (very positive). The *subjectivity score* expresses to what extent an utterance is subjective, ranging from 0 (objective/neutral) to 1 (very subjective). With the addition of this feature, the present study intends to complement research on prototypical expressive CMC markers (e.g., emoticons and emoji) by comparing adolescents' exploitation of a traditional (verbal) and a digital (typographic/pictorial) repertoire with respect to the expression of emotional or social involvement.

"Noisy" Text: Issues and Challenges

The feature extraction from the corpus and statistical analysis were complicated by the "noisy" nature of the social media texts: many messages contained various "deviations" from standard writing, mainly in terms of spelling or typography (e.g., deliberate letter repetition, emoticons). As illustrated below, this generated distorted results with regard to the measurement of lexical richness and the analysis of lexical sentiment expression. Therefore, a reliable analysis required normalization of the corpus, that is, a conversion of the original utterances into their standard Dutch equivalent.

Starting with lexical richness, example (1) is a standard Dutch sentence that contains a total of 8 words, and 7 *different* words (the pronoun *ik* ("I") occurs twice). Consequently, the Guiraud TTR would be 7 (types, i.e., different words) divided by the square root of 8 (tokens, i.e., all words), which equals 2.47.

(1) *nee denk ik, ik weet het niet goed* ("I don't think so, I'm not sure")

However, this approach may be problematic when applied to texts containing deviations from the formal writing standard. First of all, social media posts

often contain "nonword" elements, for example, emoji, like in example (2). While these elements to some extent might replace lexical expression, we do not wish to count them when measuring lexical richness.

(2) *dammn we look so hot* 😍🔥💋❤

In addition, instances of nonstandard orthography or morphology can distort the results with regard to lexical richness too. The two different spellings of the adverb *echt* ("really") in example (3) should, for instance, not be counted as two different words, as the actual variation is orthographic (with *egt* being a nonstandard spelling alternative) rather than lexical in nature. And in example (4), the verb *willen* ("to want") is conjugated in the second person singular either without (*wilt*) or with the enclitic pronoun *-de* (*wilde*) attached to it; this distinction illustrates morphological rather than lexical variation. Next, in example (5), the nonstandard contraction of *ik ga* ("I am going to") to the single token *kga* could lead to a misinterpretation of lexical richness too (as without normalization of the sentence, only one word would be counted, instead of two). Finally, in example (6), it is debatable whether the acronym *OMG* ("oh my god") should be considered as a token on its own, or whether it should be converted to its full form and counted as three words instead of one (in our own normalization procedure, which is described below, we opted for the conversion of acronyms and abbreviations to their full form, mainly because many automated language analysis tools are better at handling "full" words than shortened forms).

(3) ***egt*** *vervelend* / ***echt*** *vervelend* ("really annoying")

(4) ***wilt*** *gij komen* / ***wilde*** *gij komen* ("do you want to come")

(5) ***kga*** *eten* / ***ik ga*** *eten* ("I'm going to eat")

(6) ***OMG*** *geweldig* / ***oh my god*** *geweldig* ("oh my god, awesome")

Other issues emerged concerning the analysis of lexically expressed sentiment, since the automated tool used for this examination is based on a lexicon of standard Dutch words (see De Smedt & Daelemans, 2012a, 2012b). Table 2 illustrates how the results become less reliable when the tool is applied to nonstandard text.

The first three posts in Table 2 are standard Dutch sentences. For these examples, the sentiment function performs well: compared to message (7), the polarity score increases when the intensifying adverb *zeer* ("very") is added to the positive adjective *blij* ("happy") in message (8), and it increases

Table 2. Illustration of Sentiment Analysis.

No.	Utterance	Polarity [–1, 1]	Subjectivity [0, 1]
(7)	Ik ben blij ("I am happy")	0.55	0.95
(8)	Ik ben zeer blij ("I am very happy")	0.61	1.00
(9)	Ik ben zeer blij! ("I am very happy!")	0.76	1.00
(10)	**Ik ben zeer bly! ("I am very happy!")**	**–0.63**	**0.90**
(11)	**kben echt meeeeega bly!! :D ("I'm really suuuuuper happy!! :D")**	**0.66**	**0.70**

even more when an exclamation mark is added in message (9). Consequently, an increasingly positive sentiment appears to be expressed in messages (7) to (9). The subjectivity scores follow a similar pattern. However, when nonstandard elements are added to the utterances, the output of the sentiment function becomes less reliable. The sole deviation from standard Dutch in message (10) is the nonstandard spelling of the adjective *blij* ("happy") as *bly*. This small orthographic adaptation causes the polarity score to drop under zero, that is, the utterance is considered to express a negative rather than a positive sentiment. In addition, the subjectivity score slightly decreases too. Message (11), finally, can be considered as a more enthusiastic and also a very nonstandard version of the original message (7), containing multiple common CMC markers (contraction of *ik ben* "I am" to *kben*, expressive lengthening of the vowel in the intensifier *mega* ("super"), CMC spelling of *bly*, expressive repetition of the exclamation mark, and a smiley face emoticon). Even though intuitively message (11) seems to be the most positive and subjective one of all five utterances, its polarity and subjectivity scores are lower than those of other examples in Table 2. This demonstrates the unreliability of the tool's outcome when applied to social media data containing various nonstandard elements (elements that are, of course, of great relevance from a sociolinguistic and variationist point of view, but that cannot be dealt with accurately by many computational tools), and thus confirms the need for normalization of the corpus prior to automatic linguistic analysis.

Normalization of the Data

We first experimented with an existing tool for normalization. However, since it did not appear to perform optimally on our data, we developed our own normalization procedure in order to improve the results.[8] For alternative

Table 3. Normalization Procedure.

Step	Example: original post	Example: post after normalization step
1. Delete nonwords	*we look so hot*😍🔥💋❤	*we look so hot*
2. Normalize typography	*mooooooooooooooooiiiiiii*	*mooi* ("beautiful")
3. Replace common abbreviations and acronyms by full version	*Ja idd*	*Ja inderdaad* ("yes indeed")
4. Replace common nonstandard renderings of Dutch words and contractions of multiple words by their standard equivalents	*ni grappig* *kzie het*	*niet grappig* ("not funny") *ik zie het* ("I see [it]")

approaches to normalizing social media data, see De Clercq et al. (2013), and Han et al. (2013).

The applied normalization procedure was token-based and consisted of four steps (see Table 3). In Step 1, nonword tokens (e.g., emoji) were deleted. In Step 2, the remaining tokens' typography was normalized (e.g., expressive character repetition was reduced). These first two steps were carried out automatically using regular expressions. In Step 3, common nonstandard abbreviations and acronyms were replaced by their full versions, and in the 4th and final step, common nonstandard renderings of Dutch words or contractions of (multiple) words were replaced by their standard equivalents. For these last two steps, predefined handcrafted lists were used, containing nonstandard forms and their standard Dutch equivalent.

We note that neither the original nor the normalized corpus contains images, pictures, or stickers: these were automatically removed when the chat conversations were submitted by the participants as txt-files. Hashtags, indicating the topic of a conversation (e.g., *#levensles*, "#life lesson"), and mentions, used to address one specific person in a group conversation (e.g., *@robin*), appear in the original dataset but are very infrequent (resp. 0.012% and 0.005% of all tokens). In the normalized corpus, only the symbols # and @ were removed, which allowed us to keep the actual words for the lexical analyses. Finally, hyperlinks, (anonymized) email addresses, and filenames (of attached documents) were not adapted in the normalization procedure since we do not consider them to be "nonstandard." We argue that these tokens' influence on features such as average word length is minor, due to

Table 4. Error Analysis of the Normalization Procedure.

Scenario	Before	After	No. of tokens
1	**Standard**	**Standard (unchanged)**	**406 (69%)**
2	Standard	Incorrectly changed	0
3	Nonstandard	Nonstandard (unchanged)	48 (8%)
4	Nonstandard	Incorrectly changed	0
5	**Nonstandard**	**Standard (changed)**	**137 (23%)**

their low frequency in the corpus (only 0.09% of all tokens are hyperlinks, 0.006% are email addresses and 0.02% are filenames).

In order to evaluate the normalization accuracy, we performed an error analysis on a test set of 100 posts (591 tokens) that were randomly selected from the corpus. The quality of the normalizations was evaluated at token level: the (non)adaptation of each token in the test set was labeled as one of five possible scenarios (see Table 4). In the first two scenarios, the original token was rendered in a standard way already (i.e., not requiring normalization), which either remained unchanged (scenario 1), as desired, or was (unnecessarily so and thus incorrectly) adapted (2). In the final three scenarios, the original token deviated from formal standard Dutch in one or multiple ways. Undesired outcomes then consisted in leaving the token unchanged (3) or in adapting it incorrectly (4), whereas the desired outcome was an adequate adaptation of the token (5). In Table 4, the two desired scenarios, (1) and (5), are rendered in bold. Clearly, the other potential scenarios should be avoided. Only 8% of the tokens in the test set were dealt with incorrectly. All of these concerned nonstandard tokens that were not altered. Finally, it can be derived from Table 4 that the original test set contained 69% standard tokens, which rose to 92% after normalization. The results from the error analysis suggest that the output of the normalization procedure is reliable for further linguistic analysis.

Method

The analysis of the teenagers' top favorite words consisted of an automated and a manual component. First, each token's frequency of occurrence was counted automatically. Next, per subgroup of participants (e.g., boys versus girls), the 500 most frequent tokens were inspected manually.

All other analyses were carried out using linear mixed models (LMMs), as implemented in the lme4 package for R (Bates et al., 2017). Per linguistic feature a separate model was trained, with that particular feature serving as

response variable. The models enabled simultaneous inspection of the impact of the different social variables (serving as *fixed* effects or predictors) included in the research design, that is, the authors' age, gender, and educational track. Each predictor's main effect on the linguistic variable was examined as well as its impact in interaction with the other predictors. For each linguistic variable, the optimal model (and its optimal subset of predictors) was experimentally determined using a backward stepwise procedure in which fixed effects with a nonsignificant impact were removed. In addition to the fixed effects, a random effect for participant was added, which enabled the models to take into account the impact of individual chatters, and to deal with repeated observations (i.e., the teenagers could occur in the corpus at both a younger and an older age).

Finally, we note that all LMM analyses were carried out on the participant level (rather than a post or token level)—for example, average sentiment scores were calculated per participant, based on all his or her messages. Therefore, in terms of preprocessing and noise reduction, we deleted the material of participants who had donated fewer than 20 posts, as their text sample might be less representative of their online writing.

Results and Discussion

This section presents the results per linguistic variable. We recall that the analyses were carried out on the normalized version of the social media texts in the corpus.[9] Additional examinations of the original texts (including nonstandard elements) are discussed in the next section.

Average Post Length

Significant predictors with respect to the teenagers' average post length (expressed in number of tokens) are educational track and the interaction between age and gender (see Tables 5 and 6 for an overview of the fixed effects and the ANOVA). Authors' educational track significantly influences their average post length, with teenagers in the practice-oriented Vocational track producing significantly shorter messages than their peers in more theoretical tracks (see Figure 1, left panel, for the effect plot). Students in Technical and General Education do not significantly differ from one another in this respect. Furthermore, age and gender interact: While both girls and boys write longer social media posts as they grow older, this increase in post length is much stronger (and only significant) for girls (see Figure 1, right panel). Finally, a general gender effect can be found, with girls producing significantly longer messages than boys at any age.

Table 5. Average Post Length: Fixed Effects (Reference Category: Younger Girls in General Education).

	Estimate	Std. error	*t* value
(Intercept)	5.9556	0.1499	39.72
ageOlder	0.9401	0.1736	5.41
genderMale	−0.7418	0.1848	−4.01
educationTechnical	0.2083	0.1719	1.21
educationVocational	−0.5986	0.1829	−3.27
ageOlder:genderMale	−0.5910	0.2512	−2.35

Table 6. Average Post Length: ANOVA.

	χ^2	*df*	Pr(>χ^2)	Sig.
Age	27.3511	1	1.697e-07	***
Gender	47.1608	1	6.540e-12	***
Education	18.4015	2	0.000101	***
Age: gender	5.5363	1	0.018626	*

Sig. codes: *$p < .05$. **$p < .01$. ***$p < .001$.

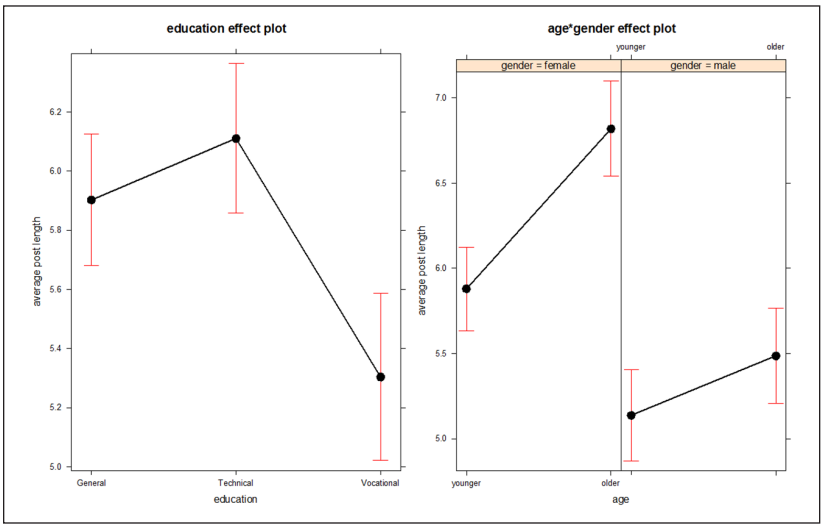

Figure 1. Average post length: effect plot.

The production of longer utterances might be considered an indication of a stronger (traditional) linguistic proficiency. In addition, as the texts are normalized and thus no longer contain nonword elements such as emoji, a longer post length implies more *lexical* expression. The observation that older teenagers and theory-oriented students produce longer posts might suggest that these youths are more proficient in writing or simply more verbally oriented. While the education-related findings correspond to the stronger focus on writing in more theoretical school systems, the results with respect to age suggest that teenagers in *each* educational track become more proficient in writing as they grow older. Or maybe they simply take more pleasure in writing or become more confident in it. The observation with regard to educational variation to some extent corresponds to findings of Verheijen and Spooren, who found that higher educated youths tend to produce longer texts than youths with lower levels of education (2017, p. 9). However, they experimented with formal writing tasks. The observed gender difference—that is, girls producing longer posts than boys—finally, is harder to explain than the age- and education-related variation, but does correspond to previous findings on average sentence length (Lin, 2007, pp. 20–21; Newman et al., 2008, p. 213) and average text length (Verheijen & Spooren, 2017, p. 9).

Average Token Length

For average token length (expressed in number of characters), gender is the only relevant predictor, with boys producing significantly longer words than girls (see Figure 2). This finding corresponds to previous results (Lin, 2007, p. 21; Mehl & Pennebaker, 2003, p. 865; Newman et al., 2008, pp. 213–214, 223). The production of longer words might be interpreted as the result of a stronger command of more complex words and thus potentially a stronger traditional literacy. Obviously, word choice (and word length) may also be related to the conversation topic: We will therefore compare the top favorite words and related topics for boys and girls (see below). Interestingly, the combination of this result and the findings on post length suggest that boys' and girls' online writing is fairly "balanced" in terms of complexity and traditional proficiency or literacy, with girls producing posts that contain more but shorter words, and boys producing posts that contain fewer but longer words. The fixed effects and the ANOVA test are presented in Tables 7 and 8.

Lexical Richness

With respect to lexical richness, expressed as Guiraud TTR, age and educational track are significant predictors (see Tables 9 and 10 for the fixed

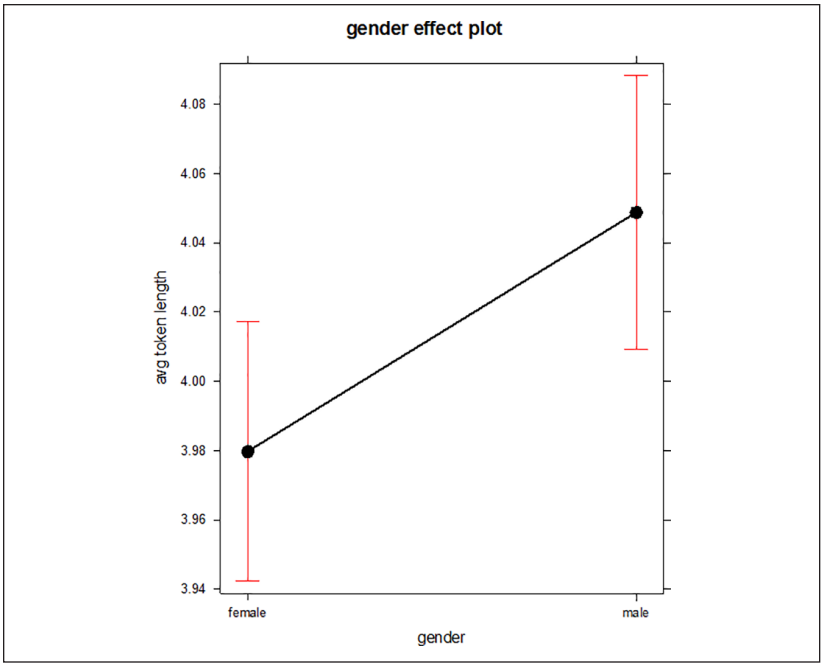

Figure 2. Average token length: effect plot.

Table 7. Average Token Length: Fixed Effects (Reference Category: Girls).

	Estimate	Std. error	*t* value
(Intercept)	3.97977	0.01908	208.55
genderMale	0.06907	0.02773	2.49

Table 8. Average Token Length: ANOVA.

	χ^2	*df*	Pr(>χ^2)	Sig.
Gender	6.2038	1	0.01275	*

Sig. codes: *$p < .05$. **$p < .01$. ***$p < .001$.

effects and the ANOVA). Older teenagers produce lexically richer texts than younger adolescents (sig.), as Figure 3 (left panel) shows, which confirms the assumption (and previously attested pattern) that people's

Table 9. Lexical Richness: Fixed Effects (Reference Category: Younger Teenagers in General Education).

	Estimate	Std. error	*t* value
(Intercept)	10.8161	0.1978	54.68
ageOlder	0.6752	0.2066	3.27
educationTechnical	0.6937	0.2672	2.60
educationVocational	0.4227	0.2865	1.48

Table 10. Lexical Richness: ANOVA.

	χ^2	*df*	Pr(>χ^2)	Sig.
Age	10.685	1	0.00108	**
Education	6.953	2	0.03092	*

Sig. codes: *$p < .05$. **$p < .01$. ***$p < .001$.

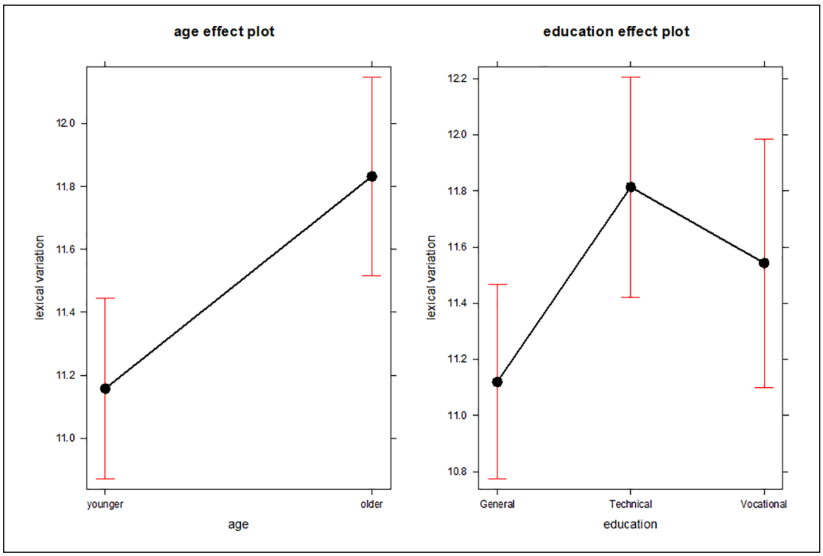

Figure 3. Lexical richness: effect plot.

vocabulary expands with age (see Sankoff & Lessard, 1975, p. 689). This age pattern with respect to active/productive vocabulary size appears to hold in the informal context of social media and CMC as well as in formal

writing tasks (for the latter, see Verheijen & Spooren, 2017, p. 9). Sankoff and Lessard (1975) conducted a similar linear regression analysis with lexical richness (TTR) as response variable for informal speech. Although their operationalization is somewhat different from ours, the results are worth comparing. The authors report a significant impact of the product of age and education, resulting in an enrichment of productive vocabulary by speaker age, which can be magnified through extensive education (Sankoff & Lessard, 1975, p. 689). However, most of their participants are adults. The effect of educational background, which in Sankoff and Lessard's study also includes tertiary education, may be stronger after completion of the complete educational cycle than in the midst of secondary education, as in our current study.

Our results reveal a somewhat surprising educational pattern, with students in General Education producing less lexical variation than students in Technical Education—see Figure 3, right panel.[10] As the focus on language teaching is more emphasized in more theory-oriented tracks, General Education students might have a larger *formal* vocabulary size—however, this does not appear to imply a greater lexical diversity in the informal setting of social media. In addition, our results do not correspond to previous findings on lexical richness in other genres. In related work, level of education and lexical richness are positively correlated (see, e.g., Sankoff & Lessard, 1975, p. 689; and Verheijen & Spooren, 2017, p. 9, for findings on informal speech and on formal writing tasks, respectively).

With respect to gender patterns and lexical richness, while our data reveal no significant correlation, previous studies report conflicting results (see the Related Research section). The discrepancy between some of our findings and related research suggests that tendencies for lexical richness based on the analysis of formal writing or traditional face-to-face conversations do not necessarily hold in the informal context of social media.

Lexical Expression of Sentiment

The final linguistic variable that is analyzed quantitatively is lexical sentiment expression in social media writing. Both the teenagers' polarity and subjectivity scores will be discussed. We calculated the average polarity score per participant using the absolute value of the original score for each utterance (i.e., the values regardless of their sign, so the *nonnegative* values of the scores); otherwise, negative and positive posts would level each other out, creating the false impression that the author did not produce polarized texts. The average polarity score (in absolute value) is significantly impacted by all three social variables, that is, the teenagers' age, gender, and

Table 11. Average Polarity (Abs. Value): Fixed Effects (Reference Category: Younger Girls in General Education).

	Estimate	Std. error	*t* value
(Intercept)	0.104377	0.002735	38.17
ageOlder	0.010299	0.002761	3.73
genderMale	−0.017084	0.002742	−6.23
educationTechnical	−0.006698	0.003174	−2.11
educationVocational	−0.018767	0.003484	−5.39

Table 12. Average Polarity (Abs. Value): ANOVA.

	χ^2	*df*	Pr($>\chi^2$)	Sig.
Age	13.918	1	0.000191	***
Gender	38.833	1	4.617e-10	***
Education	29.022	2	4.988e-07	***

Sig. codes: *$p < .05$. **$p < .01$. ***$p < .001$.

educational track (see Tables 11 and 12 for the fixed effects and the ANOVA). Significantly more polarized messages are written by female, older, and theoretically educated students, with a gradual increase from Vocational to Technical to General Education (all levels significantly differing from one another). This is shown in Figure 4.

Strongly related to the variable of polarity is the (lexically expressed) subjectivity of a text.[11] Again and in a similar way, all three social predictors significantly influence this linguistic variable (see Tables 13 and 14 for the fixed effects and the ANOVA): older teenagers, girls, and theoretically educated students produce more lexically subjective messages, once again with a gradual and significant increase from Vocational to Technical to General Education. This is shown in Figure 5.

With respect to gender and age, similar patterns have been reported in related research: Girls/women and younger people tend to be more committed to emotional expressiveness than their male and older peers, both in offline and online communication (Baron, 2008, p. 51; Newman et al., 2008, pp. 223, 229; Schwartz et al., 2013, p. 9). As for the observed educational patterns, we recall that students in the two more theory-oriented tracks are taught about other functions of language apart from the strictly communicative one, such as its expressive function: They learn how to express themselves and their feelings *verbally* (VVKSO, 2014, p. 10).

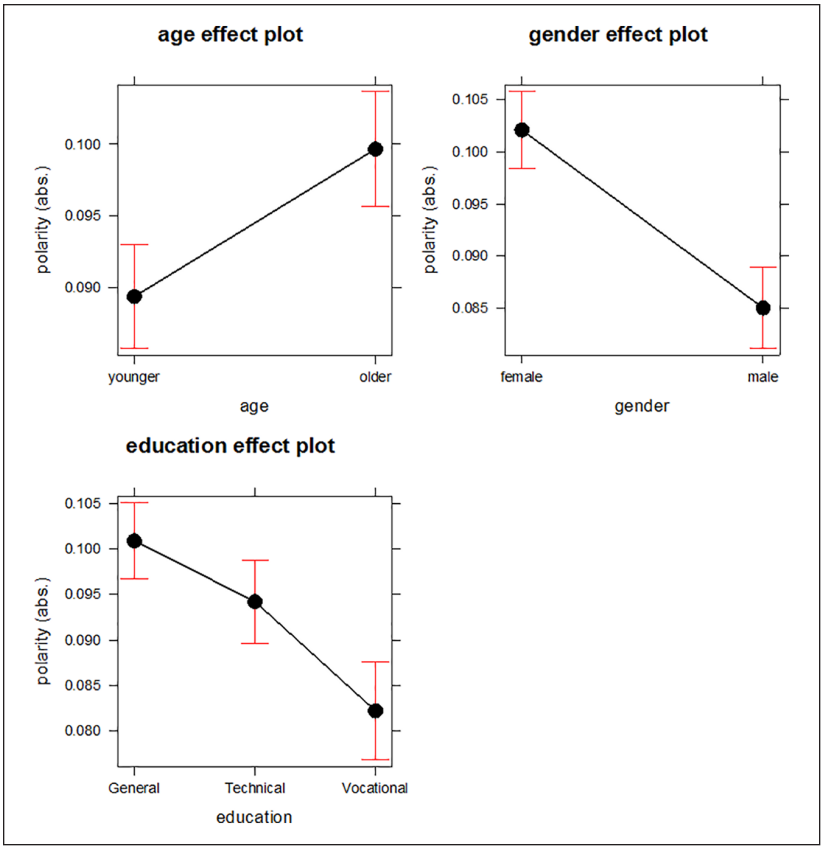

Figure 4. Average polarity (absolute value): effect plot.

Table 13. Subjectivity: Fixed Effects (Reference Category: Younger Girls in General Education).

	Estimate	Std. error	*t* value
(Intercept)	0.211047	0.004858	43.44
ageOlder	0.024774	0.004797	5.16
genderMale	−0.030562	0.004905	−6.23
educationTechnical	−0.012354	0.005696	−2.17
educationVocational	−0.033836	0.006205	−5.45

Table 14. Subjectivity: ANOVA.

	χ^2	*df*	Pr(>χ^2)	Sig.
Age	26.672	1	2.411e-07	***
Gender	38.818	1	4.652e-10	***
Education	29.742	2	3.479e-07	***

Sig. codes: *$p < .05$. **$p < .01$. ***$p < .001$.

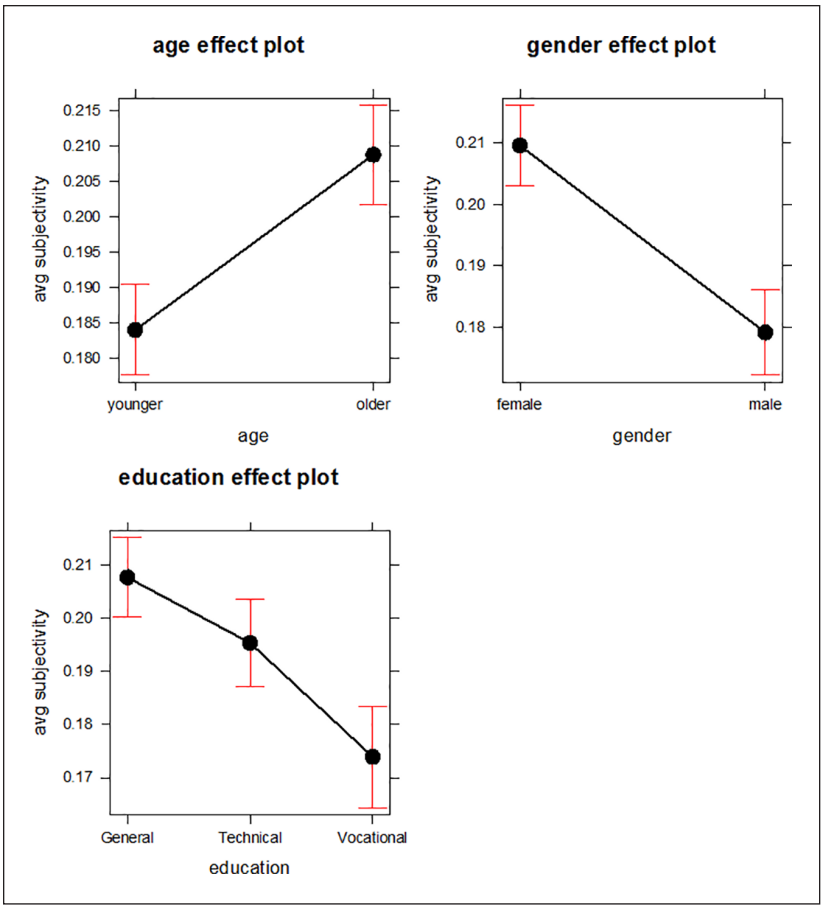

Figure 5. Average subjectivity: effect plot.

Top Favorite Words

For each subgroup of participants, the 500 most frequent tokens were extracted from the normalized corpus. Consequently, only actual words are taken into account, and not, for example, emoji. Below, we discuss the words and associated topics that appear to be popular among all groups of teenagers. For this analysis, we excluded function words. Next, we present more detailed findings per social group.

Manual inspection of different groups of youths' top-500 content words reveals that the most prominent topics discussed on social media are nearly identical for all teenagers irrespective of their sociodemographic profiles. Consequently, there appears to be a strong common ground among adolescents with respect to the contents of their social media messages. Many of the most popular words relate to family and friends (e.g., *mama* "mom," *zus* "sister"). Another popular topic is school (e.g., *school, studeren* "to study," *wiskunde* "mathematics") (for similar observations, see Argamon et al., 2009, pp. 121–122; Schwartz et al., 2013, p. 10). A final prominent category consists of words related to social media or communication (e.g., *gsm* "cellphone," *Facebook, doorsturen* "to forward").

The top-500 content words for younger and for older teenagers are nearly identical, which implies that these two groups of youths communicate about very similar topics on social media (i.e., the topics mentioned above). With respect to gender-related patterns, manual inspection of the 500 most frequent content words for boys and girls reveals very similar tendencies too, with largely the same topic preferences (see above). Still some subtle differences can be found. While all authors tend to use school-related terms, the word *stress* holds a prominent position in the girls' texts only. This might indicate a different school experience for teenage girls versus teenage boys. Another discrepancy concerns the presence of words relating to social interaction or conflict, which are prominent for girls only. For instance, *ruzie* ("quarrel"), *praten* ("to talk"), *wenen* ("to cry"), *mis* ("miss"), and *lachen* ("to laugh"). Taboo words (e.g., *fucking, shit*) and words with a "tougher" or "cooler" connotation (e.g., *gast* "dude"), on the other hand, are only favored by boys. Finally, words relating to sports and games appear to be typically male too (e.g., *trainen* "to train," *spel* "game," *online*). Some of these tendencies have been reported before. This holds, for instance, for the male preference for swear words (Mehl & Pennebaker, 2003, p. 866; Newman et al., 2008, pp. 213–214, 223; Schwartz et al., 2013, p. 8) and for lexemes related to football and gaming (Argamon et al., 2009, p. 121; Schwartz et al., 2013, p. 8) versus the female preference for words referring to social or psychological processes (Newman et al., 2008, p. 223). However, related research suggests that "family and friends" is a prominent topic in female discourse only (Argamon et al., 2009, p. 121;

Schwartz et al., 2013, p. 8), whereas our data do not reveal a distinction between both genders in this respect.

As for the educational tracks, finally, the analysis once again reveals more similarities than divergence. Obviously, the same general topics prevail (see above), but once again, some subtle differences emerge. While school-related words are popular among all teenagers, a larger proportion of these lexemes can be found in the more theory-oriented (General and Technical) students' top words only, potentially revealing a slightly stronger preoccupation with school issues (e.g., the following lexemes do not occur among the Vocational students' top-500 words: *examen(s)* "exam(s)," *tekst* "text," *wiskunde* "mathematics"). Another difference concerns the use of "tougher"/"cooler" words (e.g., *fuck, shit*); while some of these lexemes figure among the top words for all three groups of adolescents, a wider diversity of these words is present in the top-500 lexemes for the two more practice-oriented tracks. This might indicate an attitudinal difference, that is, this particular vocabulary could hold a higher prestige in the eyes of these students compared to their peers in General Education. Strikingly, in addition to these "tougher"/"cooler" words, some love-related lexemes appear to be more favored by students in practice-oriented tracks too (e.g., *schat* "honey," *love*). In fact, these students, and especially the students in the Vocational track, seem to use more words that relate to social or emotional processes in general (e.g., *samen* "together," *praten* "to talk," *mis* "miss," *helpen* "to help," *ruzie* "quarrel," *voel* "feel," *pijn* "pain," *kwaad* "angry"). This higher rate of social words might be indicative of an attitudinal difference that mirrors a finding from our previous work (Hilte et al., 2019). We asked Flemish teenagers to evaluate anonymous social media messages and to guess the authors' educational track. Moreover, they had to list the (stylistic and content-related) cues used in their decision making. On a content level, students in more practice-oriented tracks (Technical and Vocational) were considered to be more "sociable," and, according to their peers, this characteristic was apparent in their online communication too (Hilte et al., 2019). A final minor difference concerns words relating to communication and social media. While this is a popular topic among all students, some additional terms relating to calling each other on the phone only appeared in the Vocational students' top-500 (e.g., *bel, bellen*, "call, to call"). This finding potentially suggests a difference in communicative style and medium preference.

Normalized versus Nonnormalized Texts: A Comparison

For the analyses described in the previous section, the corpus was normalized first: consequently, only the strictly *lexical* realizations of the variables were

examined. This research design provides more insight in adolescents' traditional (verbal) literacy in the informal setting of social media writing. However, digital literacy may play a key role in online writing too—that is, familiarity with the characteristics and conventions of informal online communication, and the inclusion of nonlexical realizations of the above mentioned phenomena. In this section, we will compare both types of literacy in the adolescents' instant messages.

Example (12), for instance, is a social media message consisting solely of emoji. It clearly is a highly expressive utterance: it is subjective rather than neutral, and it appears to convey a positive message of love and happiness. However, it does not contain any *lexical* expression of sentiment or emotion. Consequently, analyses with a strictly lexical focus would not deal with the emotion expressed in this utterance (and similar ones), which would clearly be an underestimation of the expression of sentiment on social media. Moreover, the author of (12) inserts a whole range of emoji, which seems to indicate a strong reliance on typographic expression rather than verbal expression. Therefore, we compared the adolescents' reliance on their digital repertoire to their reliance on more traditional literacy in a social media setting.

(12)😍😊☺️😘💋❤️

For each of the five variables discussed above, we compared the normalized social media posts to the original ones, that is, the authentic texts that include nonstandard features and CMC phenomena. Largely the same patterns were attested for the raw data as for the normalized texts with respect to average post length, except that the interaction between age and gender lost significance for the raw texts: so girls and older teenagers still produce longer posts than boys and younger teenagers, respectively, but the increase in post length by age is no longer (significantly) more outspoken for girls than it is for boys. With respect to lexical richness, the patterns found in the normalized texts appeared to hold in the raw texts, with an additional gender effect emerging, that is, girls producing more diverse messages than boys. This could suggest that girls either use more alternative (nonstandard) spellings for the same word (e.g., spelling errors, but also deliberate, expressive, typographic manipulations, such as vowel repetition), or that they use more nonword elements (e.g., emoticons). Both of these assumptions are confirmed in previous research, see for example, Hilte et al. (2018b, in press). For average token or word length, the additional analyses on the raw data yield a truly different result: for the raw texts, education (and not gender) is the only significant predictor for token length, with longer tokens being produced by more

theoretically educated teenagers. A potential explanation is that students in the practice-oriented Vocational track use more emoticons and emoji (see, e.g., Hilte et al., 2018a, in press). Emoji were always counted as tokens consisting of a single character, and the manually composed emoticons typically consist of only a couple of characters (e.g., <3), so they may decrease the average token length of teenagers in practice-oriented educational tracks.

The most striking differences, however, concern sentiment expression. The teenagers' *lexical* expression was compared to their *typographic* expression. Some illustrations of typographic CMC markers that express emotion can be found below. In example (13), the use of capital letters (which mimics shouting) and the repetition of the exclamation mark both intensify the expression of anger. In example (14), the lengthening of the vowel (which mimics oral stress) increases the expression of enthusiasm.

(13) *IK BEN ECHT BOOS!!!!* "I AM REALLY ANGRY!!!!"

(14) *suuuuuuper leuk!* "suuuuuuper nice!"

Previous research on the present corpus (Hilte et al., in press) revealed that the use of these typographic expressive markers is significantly impacted by the teenagers' educational track and the interaction between their age and gender. The results appear to be complementary to the findings of the present article: the lexical analyses discussed in the previous section revealed that posts produced by older teenagers are more subjective and polarized than those produced by their younger peers. However, the analyses discussed in Hilte et al. (in press) show that younger teenagers (of both genders) use significantly more *typographic* expressive markers (with the decrease by age being much stronger for girls). In other words, to some extent these age groups express emotion and social engagement in different ways, with a stronger preference for typographic expression amongst the younger teenagers and a stronger preference for lexical expression amongst the older ones.

With respect to gender, the typographic and lexical analyses reinforce each other. Girls produce significantly more *lexically* subjective and polarized instant messages (see the Results and Discussion section). In addition, they also use significantly more *typographic* expressive markers than boys, at both a younger and an older age (with the discrepancy being largest at a younger age) (Hilte et al., in press). So girls appear to invest more in the expression of emotional engagement and they appear to do so by all means, that is, both through lexical expression and through typographic new media features.

Finally, with respect to the teenagers' educational profile, the lexical and typographic analyses once again yield complementary results. The lexical

analyses (see the Results and Discussion section) showed that students in more theory-oriented tracks produce more subjective and polarized social media posts. However, Vocational students (i.e., students in the most practice-oriented track) use significantly more *typographic* expressive markers than their theoretically educated peers (Hilte et al., in press). In other words, the lexical expression of sentiment is more favored by theory-oriented students, whereas typographic and pictorial new media markers like emoticons and emoji are more popular among their peers in the Vocational track.

These results may shed more light on the complex matter of whether and how informal online writing practices interact with literacy skills. While some studies report no significant short-term effect of CMC on youths' school writing (e.g., Verheijen & Spooren, 2017, p. 6 for Dutch youths) or suggest the existence of register sensitivity among teenagers with respect to CMC- versus school writing (Hilte et al., 2019; Vandekerckhove & Sandra, 2016, both for Flemish youths), other analyses suggest (weak yet consistent) correlations between the overall use of CMC features and lower literacy levels (Drouin & Driver, 2014, pp. 264–265 for American undergraduates). Verheijen's results suggest a different impact on youths' school writing depending on the nature of the youths' (self-reported) online communication practices: "passive engagement with CMC, by [. . .] consumption of others' social media messages, was negatively related to the quality of school writings, whereas active and creative language production in CMC [. . .] was positively related to school writing performance" (2018, p. 244, for Dutch youths). Other findings suggest that *some* types of online language features (e.g., emoticons and emoji) may have a positive rather than a negative effect on literacy (Drouin & Driver, 2014, p. 265). Finally, Herring (in press) concludes that "creative uses of typography and orthography [. . .] have been found to be positively associated with literacy skills," whereas "features motivated by ease of production [. . .] are negative predictors of literacy." She hypothesizes that "[r]ather than e-grammar causing low literacy, however, it could be that people with lower literacy skills gravitate toward more casual [. . .] uses, and people with higher literacy gravitate toward more creative uses." However, our results show that Vocational students and young teenagers use more emoticons and emoji (i.e., "positive-effect" CMC phenomena) than their peers in more theory-oriented educational tracks and than older teenagers, respectively, while at the same time showing proof of less developed *traditional* literacy skills.

Discussion and Conclusion

The present study of teenagers' online writing practices focused on general linguistic variables that have received minor attention in research on online

language use compared to the prototypical (e.g., typographic) CMC markers. As the informal setting of social media writing allows authors to express themselves in traditional (e.g., verbal/lexical) as well as alternative (e.g., typographic or pictorial) ways, the present study analyzed both of these available repertoires and the way adolescents exploit them in their instant messages. In addition, we were particularly interested in potential sociolinguistic variation, as teenagers with distinct sociodemographic profiles might use these repertoires to a different extent.

We specifically examined a set of five linguistic variables, including lexical patterns and related parameters. The analyses revealed a strong common ground among the teenagers for some features (i.e., top favorite words and associated topics) and divergent writing practices for different groups of youths for other features (i.e., average word and post length, lexical richness, lexical expression of sentiment).

While some subtle nuances could be noted depending on the authors' profiles, prominent topics in all adolescents' instant messages were school, family, friends, communication, and social media. This significant overlap in top favorite words suggests that the teenagers in the study, regardless of their specific age, gender, or educational background, largely share the same interests and preoccupations.

The authors' profile did appear to significantly impact average word and post length and the lexical richness of social media posts (analyzed in the normalized version of the corpus). Higher scores for these three variables—that is, the production of longer words, longer utterances, and more lexically diverse utterances—may be indicative of a more developed traditional literacy or a stronger verbal orientation. While such traditional literacy might be expected to increase with age, and potentially be more developed for students in more theory-oriented tracks, our findings indicate that the latter expectation is not fully met by the informal social media data.

Older teenagers appear to produce longer and lexically richer posts than younger teenagers, which suggests an increase in (productive) vocabulary range with age, and potentially a stronger verbal expression or orientation and thus more developed traditional literacy skills for older teenagers or young adults. As for gender, girls appear to produce longer social media posts (in number of words), whereas boys use longer words (in number of characters). Girls' production of longer posts might reveal a more outspoken verbal expression, but boys' use of longer words might indicate a stronger command of more complex words. Consequently, these combined findings suggest that boys' and girls' online writing is rather balanced in terms of complexity, at least for these particular traditional literacy skills.

Finally, divergent patterns of educational variation could be attested. While theory-oriented students produce longer social media posts, which indicates a more outspoken verbal orientation, Technical students (i.e., students in the middle of the educational continuum from practice to theory) outperform their more theory-oriented peers in lexical richness. If the production of more lexically diverse utterances indeed is related to a stronger focus on language education, one would expect the General students to obtain the highest scores for this variable—but then this hypothesis does not seem to hold in the context of social media. Another potential explanation is that the normalized utterances are still noisy, and that, for example, a higher rate of alternative spellings (including genuine mistakes) still remain in more practice-oriented students' texts. In addition, the more practice-oriented students could have a larger dialect or regiolect vocabulary—we recall that apart from some common nonstandard Flemish renderings of general Dutch words, actual dialect lexemes (for which standard Dutch equivalents exist) were not systematically replaced in the normalization procedure, as this would imply an unwanted reduction of lexical diversity. In previous case studies, we found that practice-oriented students indeed use more nonstandard words, spelling, and typography in their online writing (Hilte et al., 2018a, in press). So to some extent, heavier reliance on a nonstandard lexical repertoire might be an explanatory factor.

This article combined two perspectives on online writing practices: a more traditional, strictly lexical, focus (examining normalized versions of the social media texts) and a digital media focus (examining the original texts, including nonstandard CMC markers). The analyses revealed a clear interaction between traditional verbal expression and the exploitation of the new media repertoire, particularly with respect to the expression of sentiment. While traditional lexical expressions of sentiment appeared to be favored by older teenagers and students in more theory-oriented educational tracks, typographic or pictorial expressions were more popular among younger teenagers and students in practice-oriented tracks. This finding suggests that the former groups are more verbally oriented, whereas the latter ones are more inclined to express themselves using a digital media-specific repertoire, including for instance emoji. We would argue that this discrepancy reveals a notable difference with respect to the expression of emotional and social involvement in online writing between specific teenage groups. As for gender, finally, teenage girls appear to exploit both traditional and digital repertoires to a greater extent than their male peers in order to convey a sentiment or emotion. We note that the subtle gender differences concerning the top words and topics might be relevant in this respect too. While boys and girls

share a set of popular conversation topics, some additional lexemes only occurred among the female top words (i.e., words related to stress and to social interaction or conflict). These particular lexemes might be indicative of a higher social sensitivity, or of a more emotional focus.

An interesting path for further research concerns the fine tuning of certain lexical features. It could be relevant to examine alternative operationalizations of lexical richness, as they may help us understand the linguistic phenomenon from different—and complementary—perspectives. A recommended strategy consists in taking frequency distributions into account, for example, by adding a distinction between more common and more difficult/sophisticated vocabulary, between function and content words, or by adding a general frequency measure (Malvern & Richards, 2012, p. 1; Read, 2000; Van Hout & Vermeer, 2007; Vermeer, 2000, p. 79). On a content level, it is advised to control for conversation topic: lexemes belonging to the semantic domain of the topic are more likely to be selected from the lexicon, and several properties of the topic (e.g., whether it is personal in nature) may impact lexical diversity (Van Hout & Vermeer, 2007, p. 129; Vermeer, 2000, p. 77; Yu, 2009, p. 254). However, we hypothesize that the influence of topic on our current dataset will be minor, since our findings reveal that teenagers in the study largely discussed the same topics. Furthermore, it is wise to control for semantic relations between words, as using alternative terms for a single concept (synonymy) essentially differs from the use of multiple words to describe distinct concepts (conceptual variation) (Ruette et al., 2014, p. 95).

With respect to the expression of sentiment on social media, it would be highly relevant to expand existing automatic sentiment analysis tools by incorporating typographic and pictorial CMC markers that serve an expressive purpose (e.g., emoticons and emoji, character repetition, allcaps), as is also argued by Hogenboom et al. (2015). This way, such tools could be applied on social media data too, as they would take both traditional and digital media-specific expressions of sentiment into account.

A final suggestion concerns the potential pedagogical relevance of this study. Since practice-oriented students maximally exploit the typographic and pictorial tools in social media settings and in doing so seem to favor writing associated with digital literacy over traditional writing skills, it seems appropriate to include these nonverbal strategies in language classes and as learning tools for other courses. This might appeal to these and other students, and stimulate the development of their communicative skills. Moreover, it might trigger reflection on the referential and pragmatic function of certain visual elements in social media writing and other informal online discourse.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the FWO (Research Foundation Flanders) under grant G041115N.

Notes

1. While emoticons are manually composed "sequences of keyboard characters that prototypically imitate facial expressions" (e.g., *:)* representing a smiling face), emoji are "small, colorful graphical icons that represent facial expressions, objects, actions and symbols," selected from a keyboard interface (e.g., the fire/flame icon 🔥) (Herring, in press—our examples).
2. The term *educational track* designates the type of secondary education that the teenagers attend. We distinguish between the three main types of Belgian secondary education, ranging from highly theory-oriented to highly practice-oriented (see the Data section).
3. For other interpretations of this term, we refer to Verheijen (2018, pp. 299–302). *Digital literacy* is often used as an umbrella term, including not only "mastering technical skills with digital tools" (p. 299) but also other (cognitive, sociological, . . .) skills, such as critical reflection on digital content (pp. 299–302).
4. This variable is referred to by a variety of names, for example, lexical diversity, lexical density, lexical variation, vocabulary richness, and vocabulary size.
5. The age of secondary school pupils in Belgium generally ranges between 12 and 18 years old. We used 13 as a lower limit to exclude primary school children and 20 as an upper limit (as long as the students were still in secondary school) so as not to exclude teenagers with a study delay.
6. A token is a visual unit separated by whitespace from the preceding visual unit.
7. See https://www.clips.uantwerpen.be/pages/pattern-en#sentiment and https://github.com/clips/pattern
8. The suboptimal performance of the tool (see Van der Goot & Van Noord, 2017) on our Flemish Dutch texts may—at least partially—be due to the fact that it was trained on Netherlandic Dutch data.
9. As mentioned in the section on "Noisy" Text, the substitution of acronyms and abbreviations by their written-out equivalent (e.g., *omg* > "oh my god") is debatable since teenagers may perceive and use certain shortened forms as words on their own rather than as shortened versions of the "actual" words. Therefore, we reran the analyses on a second normalized version of the dataset, in which the acronyms/abbreviations were kept as such. For average post length, token length, and lexical richness, the same (significant) patterns were found as for the initial normalization procedure (reported in the Results and Discussion section).

No analyses were reconducted with respect to sentiment since the sentiment function cannot handle (certain) shortened lexemes well, returning counterintuitive and unreliable scores (as illustrated in the "Noisy" Text section).

10. The Vocational students hold a middle position in this respect, but their lexical richness score does not differ significantly from their peers in Technical or General Education.
11. Although the polarity and subjectivity of a text are strongly related, they are not entirely the same. While many words with a negative or positive connotation are also subjective, subtle differences exist. For instance, De Smedt and Daelemans (2012b, p. 3569) make a distinction between *sick* meaning "sadistic" and *sick* meaning "ill." While the former is both negative and subjective, the latter is fairly objective but is still connected to a negative sentiment.

References

Androutsopoulos, J. (2011). Language change and digital media: A review of conceptions and evidence. In T. Kristiansen & N. Coupland (Eds.), *Standard languages and language standards in a changing Europe* (pp. 145–161). Novus.

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, *52*(2), 119–123. https://doi.org/10.1145/1461928.1461959

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, *18*(2), 135–160. https://doi.org/10.1111/josl.12080

Baron, N. S. (2008). Are instant messages speech? The world of IM. In N. S. Baron (Ed.), *Always on: Language in an online mobile world* (pp. 45–70). Oxford University Press.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2017). *Package "lme4."* https://cran.r-project.org/web/packages/lme4/lme4.pdf

De Clercq, O., Schulz, S., Desmet, B., Lefever, E., & Hoste, V. (2013). Normalization of Dutch user-generated content. In G. Angelova, K. Bontcheva, & R. Mitkov (Eds.), *Proceedings of the international conference recent advances in natural language processing RANLP 2013* (pp. 179–188). Incoma.

De Decker, B., & Vandekerckhove, R. (2017). Global features of online communication in local Flemish: Social and medium-related determinants. *Folia Linguistica*, *51*(1), 253–281. https://doi.org/10.1515/flin-2017-0007

De Smedt, T., & Daelemans, W. (2012a). Pattern for Python. *Journal of Machine Learning Research*, *13*, 2031–2035.

De Smedt, T., & Daelemans, W. (2012b). "Vreselijk mooi!" (terribly beautiful): A subjectivity lexicon for Dutch adjectives. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 3568–3572). ELRA.

Drouin, M., & Driver, B. (2014). Texting, textese and literacy abilities: A naturalistic study. *Journal of Research in Reading*, *37*(3), 250–267. https://doi.org/10.1111/j.1467-9817.2012.01532.x

Eckert, P. (1997). Age as a sociolinguistic variable. In F. Coulmas (Ed.), *The handbook of sociolinguistics* (pp. 151–167). Blackwell.

Finlay, S. C. (2014). Age and gender in Reddit commenting and success. *Journal of Information Science Theory and Practice*, *2*(3), 18–28. https://doi.org/10.1633/JISTaP.2014.2.3.2

Flemish Ministry of Education and Training. (2018). *Statistisch jaarboek van het Vlaams onderwijs. Schooljaar 2016 -2017* [Statistical yearbook of Flemish education. School year 2016-2017]. Department of Education and Training.

Frey, J.-C., & Glaznieks, A. (2018). The myth of the digital native? Analysing language use of different generations on Facebook. In R. Vandekerckhove, D. Fišer, & L. Hilte (Eds.), *Proceedings of the 6th conference on computer-mediated communication (CMC) and social media corpora* (pp. 41–44). University of Antwerp.

Han, B., Cook, P., & Baldwin, T. (2013). Lexical normalization for social media text. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *4*(1), 1–27. https://doi.org/10.1145/2414425.2414430

Herring, S. C. (in press). Grammar and electronic communication. In C. Chapelle (Ed.), *The concise encyclopedia of applied linguistics*. Wiley-Blackwell. http://ella.ils.indiana.edu/~herring/CEAL.pdf

Hilte, L., Vandekerckhove, R., & Daelemans, W. (2018a). Adolescents' social background and non-standard writing in online communication. *Dutch Journal of Applied Linguistics*, *7*(1), 2–25. https://doi.org/10.1075/dujal.17018.hil

Hilte, L., Vandekerckhove, R., & Daelemans, W. (2018b). Expressive markers in online teenage talk: A correlational analysis. *Nederlandse Taalkunde*, *23*(3), 293–323. https://doi.org/10.5117/NEDTAA2018.3.003.HILT

Hilte, L., Vandekerckhove, R., & Daelemans, W. (2018c). Social media writing and social class: A correlational analysis of adolescent CMC and social background. *International Journal of Society, Culture & Language*, *6*(2), 73–89.

Hilte, L., Vandekerckhove, R., & Daelemans, W. (2019). Adolescents' perceptions of social media writing: Has non-standard become the new standard? *European Journal of Applied Linguistics*, *7*(2), 189–224. https://doi.org/10.1515/eujal-2019-0005

Hilte, L., Vandekerckhove, R., & Daelemans, W. (in press). Modeling adolescents' online writing practices. The sociolectometry of non-standard writing on social media. *Zeitschrift für Dialektologie und Linguistik*.

Hogenboom, A., Bal, D., Frasincar, F., Bal, M., De Jong, F., & Kaymak, U. (2015). Exploiting emoticons in polarity classification of text. *Journal of Web Engineering*, *14*(1–2), 22–40.

Holmes, J. (1992). *An introduction to sociolinguistics*. Longman.

Labov, W. (1972). *Sociolinguistic patterns*. University of Pennsylvania Press.

Lillis, T., McMullan, J., & Tuck, J. (2018). Gender and academic writing. *Journal of English for Academic Purposes*, *32*, 1–8. https://doi.org/10.1016/j.jeap.2018.03.003

Lin, J. (2007). *Automatic author profiling of online chat logs* (Master's thesis). Naval Postgraduate School. http://calhoun.nps.edu/public/bitstream/handle/10945/3559/07Mar_Lin.pdf?sequence=1

Malvern, D., & Richards, B. (2012). Measures of lexical richness. In *The Encyclopedia of Applied Linguistics*. https://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal0755

Mehl, M. R., & Pennebaker, J. W. (2003). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality and Social Psychology*, *84*(4), 857–870. https://doi.org/10.1037/0022-3514.84.4.857

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, *45*(3), 211–236. https://doi.org/10.1080/01638530802073712

Nguyen, D. A., Doğruöz, S., Rosé, C. P., & de Jong, F. (2016). Computational sociolinguistics: A survey. *Computational Linguistics*, *42*(3), 537–593. https://doi.org/10.1162/COLI_a_00258

Peterson, S. S., & Parr, J. M. (2012). Gender and literacy issues and research: Placing the spotlight on writing. *Journal of Writing Research*, *3*(3), 151–161. https://doi.org/10.17239/jowr-2012.03.03.1

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press.

Ruette, T., Speelman, D., & Geeraerts, D. (2014). Lexical variation in aggregate perspective. In A. Soares da Silva (Ed.), *Pluricentricity: Language variation and sociocognitive dimensions* (pp. 103–126). Walter de Gruyter.

Sankoff, D., & Lessard, R. (1975). Vocabulary richness: A sociolinguistic analysis. *Science*, *190*(4215), 689–690. https://doi.org/10.1126/science.190.4215.689

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS ONE*, *8*(9), e73791. https://doi.org/10.1371/journal.pone.0073791

Singh, S. (2001). A pilot study on gender differences in conversational speech on lexical richness measures. *Literary and Linguistic Computing*, *16*(3), 251–264. https://doi.org/10.1093/llc/16.3.251

Tagliamonte, S. (2016). So sick or so cool? The language of the youth on the internet. *Language in Society*, *45*, 1–32. https://doi.org/10.1017/S0047404515000780

Tomita, D. K. (2009). Text messaging and implications for its use in education. In C. P. Ho (Ed.), *Proceedings of technology, colleges and community (TCC) 2009* (pp. 184–193). TCC Worldwide Online Conference.

Vandekerckhove, R., & Sandra, D. (2016). De potentiële impact van informele online communicatie op de spellingpraktijk van Vlaamse tieners in schoolcontext [The potential impact of informal online communication on Flemish teenagers' spelling practices in a school context]. *Tijdschrift voor Taalbeheersing*, *38*(3), 201–234. https://doi.org/10.5117/TVT2016.3.VAND

Van der Goot, R., & Van Noord, G. (2017). MoNoise: Modeling noise using a modular normalization system. *ArXiv preprint*. https://arxiv.org/pdf/1710.03476.pdf

Van Hout, R., & Vermeer, A. (2007). Comparing measures of lexical richness. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 93–115). Cambridge University Press. https://www.researchgate.net/publication/254801850_Comparing_measures_of_lexical_richness

Varnhagen, C. K., McFall, G. P., Pugh, N., Routledge, L., Sumida-MacDonald, H., & Kwong, T. E. (2010). Lol: New language and spelling in instant messaging. *Reading and Writing*, *23*(6), 719–733. https://doi.org/10.1007/s11145-009-9181-y

Verheijen, L. (2015). Out-of-the-ordinary orthography: The use of textisms in Dutch youngsters' written computer-mediated communication. In *Proceedings of the second postgraduate and academic researchers in linguistics at York (PARLAY 2014)* (pp. 127–142). University of York.

Verheijen, L. (2016). Linguistic characteristics of Dutch computer-mediated communication: CMC and school writing compared. In D. Fišer & M. Beißwenger (Eds.), *Proceedings of the 4th conference on CMC and social media corpora for the humanities* (pp. 66–69). University of Ljubljana.

Verheijen, L. (2018). *Is textese a threat to traditional literacy? Dutch youths' language use in written computer-mediated communication and relations with their school writing* (Doctoral thesis). Radboud University.

Verheijen, L., & Spooren, W. (2017, October). The impact of WhatsApp on Dutch youths' school writing. In E. W. Stemle & C. R. Wigham (Eds.), *Proceedings of the 5th conference on CMC and social media corpora for the humanities (cmc-corpora17)* (pp. 6–10). EURAC Research.

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, *17*(1), 65–83. https://doi.org/10.1177/026553220001700103

Vlaams Verbond van het Katholiek Secundair Onderwijs. (2006). *Dutch third cycle Vocational Secondary Education. Curriculum secondary education*. Author.

Vlaams Verbond van het Katholiek Secundair Onderwijs. (2014). *Dutch third degree aso-kso-tso. Secondary education curriculum*. Author.

Yu, G. (2009). Lexical diversity in writing and speaking task performances. *Applied Linguistics*, *31*(2), 236–259. https://doi.org/10.1093/applin/amp024

Author Biographies

Lisa Hilte is a postdoctoral researcher in Computational Sociolinguistics at the University of Antwerp, where she is a member of the CLiPS research group. Her fields of interest include the correlation between youths' online writing style and their sociodemographic profiles, and the way in which people adapt their language use to their conversation partner in an online setting.

Walter Daelemans is professor of Computational Linguistics at the University of Antwerp. His research interests are in machine learning of natural language, computational psycholinguistics, computational stylometry, and language technology applications, especially biomedical information extraction, conversational agents, and cybersecurity systems for social networks.

Reinhild Vandekerckhove is senior lecturer in Sociolinguistics and Dutch Linguistics at the University of Antwerp, where she is the head of the Department of Linguistics and a member of the research group CLiPS. Her research focus is on computer-mediated communication, the dynamics of adolescent peer group language and geolinguistic versus sociolinguistic variation patterns in online discourse.