ZEITSCHRIFT FÜR DIALEKTOLOGIE UND LINGUISTIK 87, 2020/2, 173–201
DOI 10.25162/ZDL-2020-0007

LISA HILTE / REINHILD VANDEKERCKHOVE / WALTER DAELEMANS

Modeling Adolescents' Online Writing Practices: the Sociolectometry of Non-Standard Writing on Social Media

Modellierung des Online-Sprachgebrauchs von Jugendlichen: die Soziolektometrie atypischer Schreibstile in sozialen Medien

ABSTRACT: The paper discusses four generalized linear mixed models fitted to capture distinct patterns of non-standard writing practices in Flemish adolescents' social media messages. Apart from a general model that predicts the count of all "deviations" from the Dutch formal writing standard, additional models were fitted for specific types of non-standard features. These types relate to the so-called chatspeak "maxims" of orality, brevity and expressive compensation. While the general non-standardness model reveals interesting correlations between the teenagers' online writing style and their socio-demographic profile, the more specific models allow for a better and more nuanced sociolinguistic understanding: for different types of non-standard writing practices, they reveal distinct dynamics between the social predictors gender, age and educational track. Strikingly different gender patterns are found for the oral features, representing traditional non-standard writing, compared to the expressive features, representing new kinds of non-standard writing, bound to digital media. Furthermore, gender does not appear to be a predicting factor for the brevity-related features, except for the most theory-oriented educational track. Consequently, we argue that non-standard writing on social media platforms should not be operationalized as one comprehensive cluster of deviations from the formal writing standard, but rather as different subsets of non-standard features that, by serving different purposes, appeal to a different extent to different groups of youngsters and consequently display distinct sociolinguistic patterns. In other words, although Flemish adolescents may have access to the same pool of non-standard markers, they do not share one and the same social "digilect".
Keywords: computer-mediated communication, language modeling, adolescents, social correlates, computational sociolinguistics

KURZFASSUNG: Im vorliegenden Beitrag werden vier statistische Modelle diskutiert, die charakteristische Muster in Nichtstandard-Schreibstilen flämischer Jugendlicher in sozialen Medien identifizieren. Es wird einerseits ein generelles Modell besprochen, das die Anzahl aller Abweichungen vom formalen Standardniederländisch vorhersagt, und andererseits weitere Modelle, die anhand spezifischer Abweichungen angepasst wurden. Diese spezifischen Abweichungen beziehen sich auf die sogenannten „chatspeak Maximen" der Mündlichkeit, der prägnanten Ausdrucksweise und der expressiven Kompensation. Das generelle statistische

Modell enthüllt interessante Korrelationen zwischen dem online Sprachgebrauch der Jugendlichen und ihren soziodemografischen Profilen, während die spezifischen Modelle eine differenziertere soziolinguistische Betrachtung ermöglichen: Für verschiedene Nichtstandard-Schreibstile ergeben sich nämlich eigene Dynamiken zwischen den sozialen Prädiktoren Gender, Alter sowie Ausbildungsmodell. Gender-spezifische Verhaltensmuster in der mündlichen Kommunikation, sozusagen traditionelle Nichtstandard-Stile, unterscheiden sich auffallend stark von Verhaltensmustern, die mit neueren Nichtstandard-Schreibstilen innerhalb digitaler Medien einhergehen. Außer im theoretischsten Ausbildungsmodell scheint Gender zudem kein geeigneter Prädiktor für prägnante Ausdrucksweise zu sein. Wir schlagen deshalb vor, dass Nichtstandard-Schreibstile auf sozialen Medien nicht durch die Gesamtheit ihrer Abweichungen vom formalen Standard-Schreibstil operationalisiert werden sollten. Stattdessen sollten Nichtstandard-Schreibstile auf Grundlage spezifischer Abweichungen definiert werden, die sich an verschiedene Gruppierungen von Jugendlichen richten und darum unterschiedliche soziolinguistische Charakteristiken aufweisen. Anders ausgedrückt: Obwohl flämische Jugendliche wahrscheinlich ein ähnliches Repertoire an Nichtstandard-Markern besitzen, benutzen sie nicht alle ein und denselben „Digilekt".
Schlagworte: computervermittelte Kommunikation, Modellierung von Sprache, Jugendliche, soziale Korrelationen, computergestützte Soziolinguistik

1. Introduction

Informal online writing on social media platforms tends to diverge from formal writing practices in several respects. Some of its non-formal or non-standard features result from the integration of substandard spoken language markers in informal computer-mediated communication (CMC), others are more typically related to digital media. Most of the prototypical features of informal written CMC can be described in terms of the three "maxims" of chatspeak or the implicit "rules" of informal online communication captured by for example ANDROUTSOPOULOS, i. e. the principles of expressive compensation, orality and brevity (ANDROUTSOPOULOS 2011: 149, DE DECKER/VANDEKERCKHOVE 2017: 255). First of all, the principle of brevity (also speed or economy) leads to a maximization of the typing speed and a minimization of the typing effort, for example through the use of acronyms and abbreviations. The orality maxim relates to the fact that the register in many forms of informal CMC is to a large extent "conceptually oral": style and register reflect oral communication and typical speech patterns rather than classical written communication. Symptomatic in this respect is for example the use of regional features and slang. Finally, the principle of expressive compensation entails the application of a large set of – mostly typographic – strategies to compensate for the absence of certain expressive cues in face-to-face communication (for example intonation, volume, facial expressions). Emoticons are a well-known example of such typographic expressive markers.

Another useful distinction that captures the different types of non-standard features and to a certain extent overlaps with the three maxims, is the dichotomy between "old"

and "new vernacular" (ANDROUTSOPOULOS 2011: 146). Old vernacular relates to "traditional" non-standardness, for example the use of regional linguistic variants. In other words, the principle of orality leads to the integration of old vernacular in CMC. "New" vernacular, however, consists of non-standard or non-formal features that are specifically bound to the online writing culture. Consequently, the linguistic features that are related to the principles of expressive compensation and brevity can generally be referred to as instances of "new vernacular". In informal computer-mediated communication, features of both old and new vernacular can be used as tools for self-profiling and identity construction. However, different social groups might favor different types of features.

The main aim of the present study is to identify correlations between teenagers' socio-demographic profile and their online writing practices, and to reveal potentially divergent social digilects for distinct groups of youths. Previous research on informal online communication indicates that distinct social groups tend to favor certain linguistic markers to a different extent. However, the distinction between old and new vernacular features has not yet been operationalized systematically in this context. For instance, while related studies suggest that some new vernacular markers such as emoticons generally appeal more strongly to girls and women (see for example HILTE/VANDEKERCKHOVE/DAELEMANS 2018c, PARKINS 2012, VARNHAGEN et al. 2010), the picture tends not to be completed or "balanced" with the analysis of social patterns for more traditional vernacular markers in online writing. Moreover, there has been almost too strong a focus on gender, to the detriment of other social variables. While this has led to very straightforward and clear findings, especially with respect to gender patterns, part of the social and linguistic reality of online communication tends to remain out of the picture.

GRONDELAERS/VAN HOUT/VAN GENT (2016) note that digitalization (including the emergence of online communication) has led to a "new social and linguistic reality" (GRONDELAERS/VAN HOUT/VAN GENT 2016: 143) in which language norms are pluralized (GRONDELAERS/VAN HOUT/VAN GENT 2016: 130) and new types of linguistic superiority criteria have become increasingly important, such as "dynamism", "media cool" or "modern media prestige" (GRONDELAERS/VAN HOUT/VAN GENT 2016: 132, see also KRISTIANSEN/GARRETT/COUPLAND 2005: 12). But obviously, different social groups might construct "media cool" in different ways. In order to capture this complex linguistic reality and social dynamics adequately, we need research on online writing in which a wide range of linguistic markers is combined with a wider range of social variables. The present paper meets this requirement by combining a range of both digital and oral vernacular markers and by including three socio-demographic variables. Since we assume that the appeal of the feature sets included in this paper might depend on the teenagers' profiles, as different types of linguistic prestige may correlate with different types of vernacular features, this should lead to a more nuanced picture of group bound preferences and in the end a better understanding of why youths prefer specific types of (standard or non-standard) features. In other words, we want to discover how teenagers construct media cool or dynamic prestige by analyzing how their socio-demographic profile influences the type of social capital they pursue in their online com-

munication, and what type of features from their linguistic repertoire are exploited to construct that social capital.

Methodological contributions of the paper concern the multidimensional conceptualization of the linguistic and social variables and the inclusion of interactions between the social variables in the research design. The latter enables us to build upon the findings of DE DECKER (2014), who actually operationalized a wide range of linguistic markers in his research on online communication by Flemish youngsters, but did not include educational track as a social variable and did not investigate the interactions between the social variables of gender and age.

The paper is structured as follows: in Section 2, the corpus and variables will be described. Next, in Section 3, we will explain the methodology, and finally, in Sections 4 and 5, we will report and evaluate our findings.

2. Corpus and variables

The present section describes the corpus and its participants (2.1) and the linguistic variables (2.2).

2.1 Corpus and participants

The corpus consists of 434 537 social media posts (more than 2.5 million tokens) written by 1 384 Flemish[1] high school students between 13 and 20 years old. The posts are private instant messages produced in Dutch on Facebook Messenger and WhatsApp. The vast majority of the tokens (87 %) was produced between 2015 and 2016. All participants' age, gender and educational track is known. An overview of the distributions in the corpus can be found in Table 1.

The participants' socio-demographic profile is operationalized as a combination of three factors, i. e. their age, gender and educational track. For age, we distinguish two groups of high school students: younger teenagers (13 to 16 years old) and older teenagers or young adults (17 to 20 years old). Age is treated as a categorical rather than a continuous variable, as previous sociolinguistic studies suggest that teenagers' non-standard language use does not evolve linearly as they age, but "peaks" during mid-puberty: it increases until the age of 15 or 16, and then decreases again. This phenomenon is often referred to as the "adolescent peak" (COATES 1993: 94, DE DECKER/VANDEKERCKHOVE 2017: 277, HOLMES 1992: 184).

Gender is operationalized as a binary variable too, since a non-binary approach (for example operationalizing gender as a continuum[2]) was infeasible with the profile infor-

1 I. e. living in Flanders, the Dutch-speaking part of Belgium.

2 See for example BAMMAN/EISENSTEIN/SCHNOEBELEN (2014), who (linguistically) approach gender as consisting of multiple gender-oriented (language) clusters, and KILLERMANN (2014) for a conceptual-

mation we had access to. As a consequence, we distinguish between teenage boys and girls.

The final social variable 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 (FMET 2017: 10).

Region is no variable in the present data set: 96.13 % of the teenagers live in the central province of Antwerp. 1.51 % of the data is produced by adolescents from the neighboring province of Flemish-Brabant. Both provinces belong to one and the same dialect area.

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	Girls	1 696 517 (67 %)	717 (52 %)
	Boys	834 837 (33 %)	667 (48 %)
Age	Younger teenagers (13–16)	1 360 898 (54 %)	1 234[3]
	Older teenagers / young adults (17–20)	1 170 456 (46 %)	897
Total		2 531 354	1 384

Table 1: Distributions in the corpus

The data were collected in a school context: we visited several secondary schools in the central province of Antwerp and invited students to voluntarily donate private social media messages that were written outside the school context and before our school visits. The latter conditions were meant to exclude the observer's paradox. We asked the students' (and for minors also their parents') consent to store and analyze their anonymized texts.

ization of gender identity as a combination of values on four continuums, relating to identity, attraction, expression and sex.

3 We note that the number of younger and older participants does not add up to the total number of participants (1 384), but to a higher number (which is why we did not add percentages for age). Participants can occur in the corpus at both a younger and older age if they submitted recent chat conversations as well as older ones. We will control for these repeated observations in the data by adding subject (participant) as a random effect in the statistical models (see Section 3.2).

2.2 Linguistic variables: Features of non-standard writing

We operationalize authors' non-standard writing as their use (in number of occurrences) of eleven "non-standard" features, i. e. not belonging to the Dutch formal writing standard or to general formal writing practices (with "general" implying non-language-specific formal writing practices; for example the insertion of emoji is generally considered to belong to informal rather than formal language). The selection of these linguistic variables is based on related research (for example PARKINS 2012, VARNHAGEN et al. 2010, VERHEIJEN 2015, and many more). Below, each of these features is presented and illustrated. The features are grouped into three sets, based on their relation to the so-called maxims of chatspeak, that were introduced above.

The largest set of features corresponds to the maxim of expressive compensation. Most of them are typographic expressive markers:

1. Emoticons and emoji:
   for example *zie u graag!* 😍😊☺️😘💋🖤 ('love you!')
2. Allcaps, i. e. entire words or utterances in capital letters:
   for example *DIT MAAKT MIJ KWAAD* ('THIS MAKES ME ANGRY')
3. Deliberate letter repetition (letter "flooding"):
   for example *W**ooooo**w goed gedaan* ('W**ooooo**w good job')
4. Deliberate repetition of punctuation marks (punctuation "flooding"):
   for example *Proficiat**!!!!!!*** ('Congratulations**!!!!!!**')
5. Combinations of question and exclamation marks:
   for example *Wat**?!*** ('What**?!**')
6. The onomatopoeic rendering of laughter:
   for example *Hahahahah*
7. The typographic rendering of kisses and/or hugs through combinations of the letters 'x' and 'xo':
   for example *Dankje **xxx*** ('Thanks **xxx**')
   for example *Veel beterschap **xoxo*** ('Get well soon **xoxo**')

The second set of non-standard features is related to the principle of orality, which entails the integration of features from substandard Dutch (for example regional varieties) or informal speech:

8. Non-standard Dutch lexemes (i. e. dialect, regiolect, colloquial or slang lexemes, or representations of non-standard pronunciation):
   for example *ik was **efkes** in de war* (std. Dutch 'ik was **even** in de war', 'I was confused **for a while**')
   for example ***gij** ook* (std. Dutch '**jij** ook', '**you** too')
   for example ***da** was mijn vraag* (std. Dutch '**dat** was mijn vraag' (t-deletion), '**that** was my question')
9. English lexemes that are not identified as (part of) Dutch:
   for example *echt **awesome*** ('really **awesome**')

We note that each token in the corpus is classified as either a non-word element (for example an emoticon), or as a standard Dutch, standard English, or non-standard Dutch word through a dictionary-based pipeline approach (i. e. the token's presence in multiple dictionaries is checked). This approach is discussed in Section 3.1 (and the specific dictionaries used are listed in footnotes 5 and 6). Concerning the integration of English lexemes, it should be noted that the base language of the selected chat messages is always Dutch, as entire chat conversations in a different language were excluded from the corpus. Furthermore, English loan words that are now officially part of the standard Dutch vocabulary and that consequently are codified in Dutch dictionaries (for example *computer*), are not counted as English lexemes in this analysis, but as Dutch. (For more detailed analyses on Dutch-speaking youths' integration of English loan words into their Dutch social media messages, see DE DECKER/VANDEKERCKHOVE 2012, 2013, and VERHEIJEN/DE WEGER/VAN HOUT 2018). For this language detection task, we used an automated pipeline approach: we only verified whether a word should be classified as English if it was not detected as Dutch. A word like *computer* was recognized as Dutch in the first step of the procedure and therefore not registered as an English lexeme. This pipeline approach will be explained in a more detailed way and evaluated in Section 3.1.

The third group of non-standard markers is related to the principle of brevity (also economy or speed) and covers all kinds of strategies to compress words or utterances:

10. typical chatspeak abbreviations and acronyms (none of them standard Dutch abbreviations)
    for example ***omg*** *hahaha* (full version: '**oh my god** hahaha')
    for example ***idd*** *man* (full version: '**inderdaad** man', '**indeed** man')

The final set of features included in the research design does not belong to any of the three main categories, but is nevertheless typical of online discourse[4]:

11. Discourse markers: # ("hashtag", to indicate a topic or express a feeling about it) and @ ("at", to address one specific person in a group conversation)
    for example *#crisis*
    for example *@nina*

As these discourse markers do not belong to any of the three subcategories, they will only be studied in the general model, where all eleven non-standard markers are combined as the response variable.

We note that the inclusion of English lexemes challenges the distinction between old and new vernacular presented above. First of all, the insertion of English words or

4 We note that these online discourse markers are especially relevant and popular on the microblogging platform Twitter. However, they are used in instant messaging too (though less frequently), as is described by ZAPPAVIGNA (2015: n. p.): "Hashtags emerged via microblogging [...] and have since spread to other forms of social media". A similar evolution can be noted for "ats" or "mentions" (@).

phrases in Flemish teenagers' informal Dutch communication can hardly be considered a traditional vernacular feature. On the contrary, most of these English lexemes appear to be trendy markers of adolescent slang (for example some popular examples from the corpus are the insertion of the adjectives *awesome* or *awkward* in a Dutch sentence, instead of their Dutch equivalent). Furthermore, while most of the English lexemes seem to reflect adolescents' oral practices, some of these features are bound to (international) internet culture and thus mark (online) writing practices rather than speech patterns. Yet, our observations suggest that the former type is dominant and therefore we decided to include the English features in the oral category (see below).

Furthermore, the present operationalization of non-standard writing covers a wide range of highly different features, both in form and function. While all of the features can be considered non-standard if formal writing practices serve as the overall reference point, it may seem incongruent that in the general model presented below the use of emoticons is considered to be non-standard just as much as the use of dialect words. Evidently, one could argue that for features such as emoticons, the comparison with formal writing makes no sense, since they are typical characteristics of the genre and have become "standard" in informal online writing. However, the latter also holds for the integration of many substandard speech features; so to some extent, this is a matter of labeling, with formal standard writing as a reference point. In order to address this tricky operationalization of non-standardness, the general model will be compared to more specific submodels, in which (mostly typographic) expressive markers and (traditional) oral features are analyzed separately.

We hypothesize that the distinct feature sets might appeal differently to different groups of youngsters, as they potentially hold distinct types of prestige. New vernacular (i. e. the typographic expressive features) might evoke "modern media prestige" (KRISTIANSEN/GARRETT/COUPLAND 2005: 15, GRONDELAERS/VAN HOUT/VAN GENT 2016: 132) and "dynamism" (GRONDELAERS/VAN HOUT/VAN GENT 2016: 133), and connotations of informality, casualty, and trendiness (GRONDELAERS/SPEELMAN 2013: 178), while many old vernacular markers, especially the dialect and regional features, might evoke localness and a certain amount of toughness. In our analyses, we will examine how these "competing standards" (GRONDELAERS/VAN HOUT/VAN GENT 2016: 133, KRISTIANSEN 2001) interact with each other in the online writing practices of Flemish adolescents and young adults.

3. Methodology

Section 3.1 discusses the data preprocessing and feature extraction. The statistical models are presented in Section 3.2.

3.1 Preprocessing and feature extraction

The dataset was ordered at a participant-level, so that each line contains information about one participant at either a younger or an older age. We recall that participants can occur in both age categories if they submitted recent as well as older chat conversations; each participant can thus be represented on maximum two lines in the dataset. Each line contains the participant's meta-information (a unique identifier as well as information on gender, age and educational track) along with the size of their submission (number of tokens) and the absolute counts for all non-standard features.

The feature occurrences in the corpus were counted automatically using Python scripts. For a test set of 200 randomly selected posts (1 257 tokens), the software's output was compared to manual annotations. The software reached a satisfying average F-score (for all eleven features) of 0.90 (90 %). Table 2 shows the evaluation metrics per feature: for all features, the metrics are sufficiently high, which indicates that the software is reliable. We note that discourse markers and combinations of question and exclamation marks are very infrequent features and did not occur in the test set. Therefore, no evaluation scores can be provided for these features. The precision score (ranging between 0 and 1) indicates the share of detected occurrences of a feature that are indeed valid occurrences of that feature. The recall score (also ranging between 0 and 1) shows the share of all occurrences in the corpus of a feature that are detected as such by the software. The F-score is the harmonic mean of precision and recall.

Feature	Precision	Recall	F-score
Emoticons and emoji	1.00	1.00	1.00
Allcaps	0.75	1.00	0.86
Letter flooding	1.00	1.00	1.00
Punctuation flooding	1.00	1.00	1.00
Combinations ? and !	undefined	undefined	undefined
Laughter	1.00	0.96	0.98
Kisses	1.00	0.89	0.94
Non-standard Dutch lexemes	0.95	0.70	0.81
English lexemes	0.60	0.47	0.53
Chatspeak abbreviations and acronyms	1.00	0.90	0.95
Discourse markers # and @	undefined	undefined	undefined
Average	**0.92**	**0.88**	**0.90**

Table 2: Evaluation metrics for the automated feature extraction

Table 2 shows that the software's performance is lowest for the features that are extracted with a dictionary-based approach, i. e. English lexemes and non-standard Dutch lexemes. Below, we provide an error analysis (performed on the test set) for these features (see also HILTE/VANDEKERCKHOVE/DAELEMANS 2018a) and discuss the extraction procedure in a more detailed way.

First, we will analyze the errors made with respect to the detection of English lexemes. The test set contains 19 English words, of which only 9 (47 %) were detected as such. The remaining 10 were not recognized: these are false negatives. In addition, 6 non-English words were labeled as English: these are false positives. The substantial size of the false negative category is mostly due to the noisy nature of the word lists used for language recognition. For the automatic count of the number of words per language or register in the corpus (standard Dutch/standard English/non-standard Dutch), a dictionary-based pipeline approach is used. The software first checks each token's presence in a large standard Dutch word list and in a list of named entities[5] (including names of people, events, etc.). If the token is in one of these lists, it is categorized as standard Dutch. If not, the software checks the token's presence in a standard English word list.[6] If it is in the list, it is labeled as English. If not, it is labeled as non-standard Dutch. A problem with this pipeline approach is that words that exist in both Dutch and English are automatically seen as Dutch in the first step. For example, in the first step of the pipeline, the English article *an* was recognized as the common Flemish/Dutch girl name *An*, and thus not detected as English. In addition, the standard Dutch word list appears to be quite noisy, containing some popular English words that are quite frequent in informal Dutch speech and writing, such as *not*, *yes*, and *geek*. This type of misclassification happened for 8 out of 10 false negatives. The false positive category is less homogeneous, and consists of different types of misclassifications, for example some misspellings in Dutch words accidentally ended up in the English category.

Since the software might be underestimating the actual presence of English words in the corpus, we must be careful when interpreting the results for this feature. In this study, however, the English category will never be analyzed on its own, but always in combination with other features (either with non-standard Dutch lexemes, for the orality model (see below), or with all 10 other non-standard markers). In follow-up research however, the extraction of this feature could be improved if less noisy word lists would be available.

5 We merged multiple existing word lists to create the final standard Dutch list: ANW, DPC, Roularta and SoNaR. Before merging them, we filtered these lists (for example English words were deleted as far as possible) and applied a frequency cutoff, in order to exclude very infrequent lexemes. For the named entities, we combined an existing list of named entities collected within our research group and lists of first and last names provided by the Belgian government. Both lists were filtered (for example a frequency cutoff was applied on the name lists) and updated (for example some specific Belgian locations were added to the named entities). For complete references of these corpora, please see Section 8.

6 The English word list was created as a combination of the existing COCA and Brown corpora. A frequency cutoff was applied, in order to exclude lexemes that were highly infrequent. For full references of these corpora, please see Section 8.

With respect to the detection of non-standard Dutch words, 97 errors were made, of which 89 % (86) were false negatives, i. e. non-standard lexemes that the software "missed". More than half of these false negatives concerned tokens that, without context, could actually be standard Dutch lexemes, and were thus classified as such by the (token-based) software in the first step of the pipeline described above. For example, the token *me* can simply be the standard Dutch pronoun *me* ('me'), but it can also represent the Flemish colloquial pronunciation of the preposition *met* ('with'). Similar errors can occur for spelling or typing errors when the incorrect form is identical to a standard Dutch word. A much lower proportion of the errors (11 out of 97, or 11 %) were false positives, i. e. the software incorrectly labeling a token as non-standard Dutch. Many of these misclassified lexemes were very specific named entities (for example the name of a local dance school) that did not occur in the standard Dutch word list (including some named entities, see above) nor in the list of English words, and were thus automatically classified as non-standard Dutch.

3.2 Model fitting

We modeled adolescents' non-standard writing practices or, more specifically, the degree of "non-standardness" using generalized linear mixed models (GLMMs) with a Poisson distribution, as implemented in the "lme4" package for R (BATES et al. 2017). These models enable simultaneous inspection of the impact of different predictors (i. e. the fixed effects) – both of their main effects and of their possible interactions with each other. The models can also take into account the impact of individual chatters and correct for repeated observations for one participant by adding a random effect for subject. Finally, they can deal with differences in sample size between participants by adding an "offset" for the logarithm of the number of tokens per chatter (see Section 4).

We chose to use GLMMs with a Poisson distribution, as these "Poisson models" are a classical (and often recommended) choice for the analysis of count data (HARRISON 2014: 2, ISMAIL/JEMAIN 2007: 105). ZEILEIS/KLEIBER/JACKMAN (2008: 5) explain that the Poisson distribution is the "simplest distribution for modeling count data" – for the mathematical background on why this distribution can adequately capture the properties of count data, we refer to COXE/WEST/AIKEN (2009: 123). However, a common problem with "naïve" Poisson models occurred in the initial experiments: there were indications of overdispersion[7], i. e. the variance of the response variable exceeding the mean (HARRISON 2014: 1–2, ISMAIL/JEMAIN 2007: 103). The equality of the mean and variance functions is a "key feature of the Poisson model" (HILBE 2011: 2), which, in reality, often does not hold for count data. However, the violation of this assumption can influence the results and validity of the trained models. First of all, overdispersion

7 For different causes of overdispersion, see e. g. HARRISON (2014: 2).

can result in a poor fit to the data (HARRISON 2014: 2). Through the underestimation of standard errors and the overestimation of parameter estimates and significance, it can lead to unreliable results, such as wrong or overestimated conclusions about the predictive power and significant influence of the predictors (HARRISON 2014: 1, 2, 17–18 and references therein; ISMAIL/JEMAIN 2007: 103). Moreover, while simple statistical models are generally preferred, "ignoring overdispersion during model selection can result in the retention of overly complex models" (HARRISON 2014: 17–18 and references therein).

In order to account for overdispersion, we added an observation-level random effect (OLRE),[8] i. e. a random effect for each observation in the data. We recall that in this study, one observation or line in the dataset contains information about one participant in one particular age group. This strategy has been described as a common, simple and robust way to deal with overdispersion in count data (HARRISON 2014: 1), as the OLRE "model[s] the extra-Poisson variation in the response variable", and does so "without making implicit, potentially erroneous, assumptions about the process that generated that overdispersion" (HARRISON 2014, resp. 2 and 17–18). The application of this strategy solved the overdispersion in our models and significantly increased their goodness of fit. We note that an alternative solution is the use of a negative binomial model (HARRISON 2014: 2, HILBE 2011: 2, ISMAIL/JEMAIN 2007: 103) or a quasi-Poisson model (HILBE 2011: 2): we also experimented with these approaches, and obtained very similar results as the ones reported in Section 4.

4. Results

The present section discusses the following four models:[9]

(4.1) A general model in which all non-standard features are analyzed jointly as one response variable
(4.2) A model for expressive features
(4.3) A model for oral features
(4.4) A model for brevity-related features

All models are generalized linear mixed models with a Poisson distribution, and a random effect for participant and observation (for a detailed description, see Section 3.2).

8 Poisson models with an observation-level random effect are also known as Poisson-lognormal models (HARRISON 2014: 2 and references therein).

9 We note that "reverse" models are possible too, i. e. models that predict authors' socio-demographic profiles based on their language use. For a pilot study on the prediction of teenagers' educational track based on their social media texts, see HILTE/DAELEMANS/VANDEKERCKHOVE (2018). Simultaneous inspection of the different dependent variables (i. e. expressive, oral, and brevity-related non-standard markers) and their potential correlations, for example through a multivariate analysis of variance (Manova), falls outside the scope of the present paper, but is an interesting path for future research, as it may complement our findings.

Each model predicts the participants' counts for certain linguistic features, while also taking the participants' sample size into account by adding the logarithm of the total number of tokens as an offset. The addition of an offset expands the Poisson model, allowing it to model rates instead of counts.[10] This is crucial in our experimental design, since the sample size (total number of tokens) differs among the participants, and the absolute feature counts may depend on sample size. For each model, different "formulas" were tested, i. e. different combinations of the predictors age, gender and educational track. Below, we will always report the model with the formula that resulted as best fit for the data. These optimal formulas were experimentally determined using a backward stepwise deletion of predictors with a non-significant impact (i. e. we systematically compared nested models with Anova tests and used the resulting p-values as selection criterion).

The sociolinguistic patterns emerging from the different models presented below (Sections 4.1 to 4.4) will be compared and discussed in the discussion section (Section 5).

4.1 General model: Non-standardness (all features)

We first modeled the occurrence (counts) of all non-standard features, without making a distinction between different types of features. For example, the total count of non-standard markers in the utterance below would be 8: 6 expressive markers (3 hearts and 3 infatuated faces), plus 1 oral feature (the Flemish colloquial pronoun *gij* instead of the standard Dutch *jij*, meaning 'you'), plus 1 non-standard abbreviation (*wrs* for *waarschijnlijk*, 'probably').

Gij komt wrs met de fiets? 🖤🖤🖤😍😍😍 ('You are probably coming by bike?')

The best results for this general model were obtained with the predictors education on the one hand and the interaction between age and gender on the other. A visualization can be found in Figure 1. The estimates and standard errors (compared to the reference category, here younger girls in the theoretical General Secondary Education track) are presented in Table 3 and the Anova for the overall effects per factor (all levels taken into account) can be found in Table 4.

10 COXE/WEST/AIKEN (2009: 134) describe such "time-varying models" as Poisson type models that, rather than "assum[ing] observation for all individuals occurs in the same length time period", are extended "to variable time periods". With regards to the "offset term", they note that "including the natural log of the measurement interval as a predictor with regression coefficient equal to 1 allows incorporation of variable time periods and maintains the Poisson error structure of the data" (ALLISON 1999, as paraphrased in COXE/WEST/AIKEN 2009: 134).

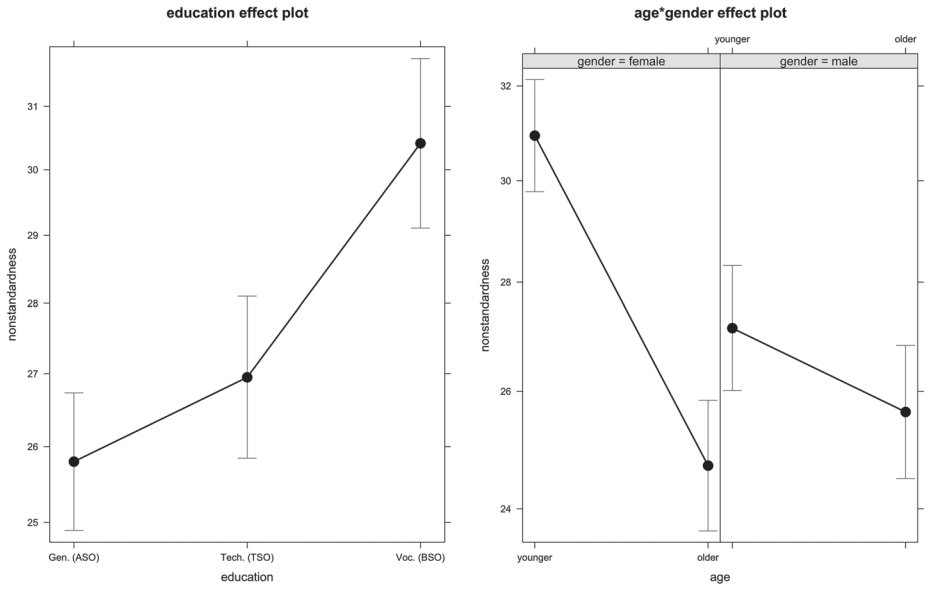

Fig. 1: Non-standardness model: effect plot (predicted counts of non-standard features per 100 tokens)

	Estimate	Std. Error	z value	Pr(>|z|)	Signif.
(Intercept)	-1.23043	0.02333	-52.73	< 2e-16	***
ageOlder	-0.22442	0.02701	-8.31	< 2e-16	***
genderMale	-0.13088	0.02913	-4.49	7.01e-06	***
educationTechnical	0.04363	0.02808	1.55	0.12	
educationVocational	0.16452	0.02877	5.72	1.08e-08	***
ageOlder:genderMale	0.16737	0.03934	4.25	2.10e-05	***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Table 3: Non-standardness model: fixed effects (reference category: younger girls in General Secondary Education)

	Chisq	Df	Pr(>Chisq)	Signif.
age	54.6779	1	1.420e-13	***
gender	6.0558	1	0.01386	*
education	33.4053	2	5.574e-08	***
age:gender	18.0960	1	2.100e-05	***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 4: Non-standardness model: Anova

Table 4 shows that all predictors, including the interaction term, have a significant impact on the adolescents' use of non-standard features on social media. Regarding educational track, Figure 1 shows that the highest number of non-standard features is predicted in the Vocational students' texts (significantly differing from the other two educational tracks, for all age/gender groups). There is no significant difference (for none of the age/gender groups) between the Technical and General students' use of non-standard markers.

The statistical significance of the interaction term indicates that the teenagers' gender and age influence each other and that their effects depend on each other: the impact of these two factors should therefore be interpreted simultaneously. A cross-interaction emerges from Figure 1: in both gender groups, older teenagers use fewer non-standard features than younger teenagers, but the decrease is much steeper for the girls. While the age difference is always significant (for all gender/education groups), the gender difference is only statistically significant for younger teenagers (in all three educational tracks), with girls using more non-standard features than boys. At an older age, girls use slightly fewer non-standard markers than boys, but not significantly so.

4.2 Submodel: Expressiveness

The second model's response variable are the counts for all expressive non-standard markers. In the example below, this count would equal 6: only the expressive markers (3 hearts and 3 infatuated faces) are counted, and not the oral *gij*, which is a substandard pronoun, or the non-standard abbreviation *wrs* for *waarschijnlijk* ('probably').

Gij komt wrs met de fiets? 🖤🖤🖤😍😍😍 ('You are probably coming by bike?')

Once again, the best results were obtained with the predictors education and the interaction between age and gender. The model's predictions are visualized in Figure 2. The estimates and standard errors (compared to the reference category: younger girls in General Secondary Education) are presented in Table 5 and the Anova for the overall effects of the factors can be found in Table 6.

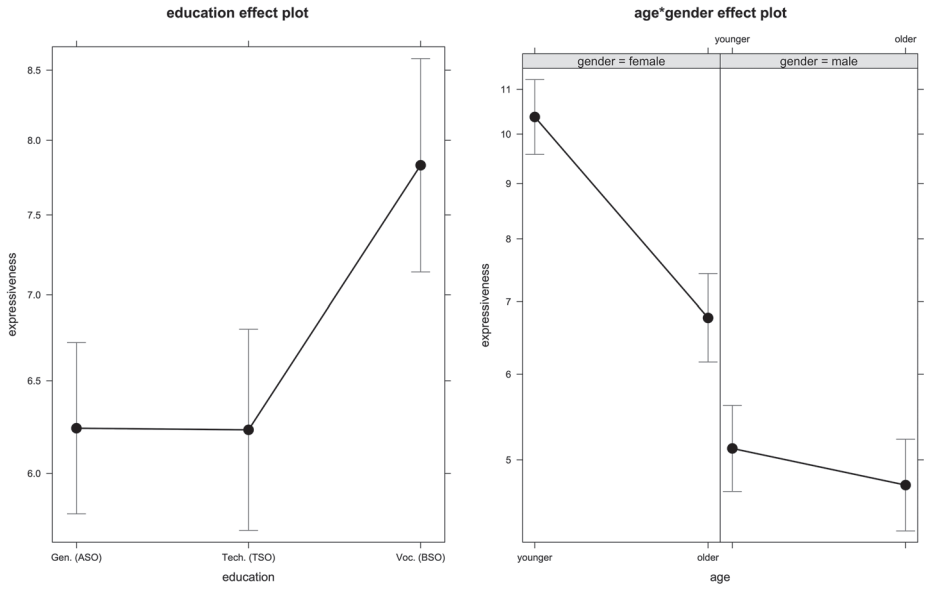

Fig. 2: Expressiveness model: effect plot (predicted counts per 100 tokens)

	Estimate	Std. Error	z value	Pr(>|z|)	Signif.
(Intercept)	-2.325802	0.048525	-47.93	< 2e-16	***
ageOlder	-0.427283	0.058177	-7.34	2.06e-13	***
genderMale	-0.705199	0.061788	-11.41	< 2e-16	***
educationTechnical	-0.001413	0.058264	-0.02	0.980646	
educationVocational	0.227048	0.060059	3.78	0.000157	***
ageOlder:genderMale	0.349235	0.085797	4.07	4.69e-05	***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Table 5: Expressiveness model: fixed effects (reference category: younger girls in General Secondary Education)

	Chisq	Df	Pr(>Chisq)	Signif.
age	38.759	1	4.794e-10	***
gender	126.573	1	< 2.2e-16	***
education	17.143	2	0.0001895	**
age:gender	16.569	1	4.692e-05	***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 6: Expressiveness model: Anova

Table 6 shows that all predictors, including the interaction term, have a significant impact on the adolescents' use of expressive (non-standard) features on social media. As for the effect of educational track, Figure 2 shows that the highest number of expressive markers occurs in the Vocational students' texts (significantly differing from the other educational tracks for every age/gender group), followed by the Technical and General students'. For the latter groups the data render no significant difference (regardless of the youngsters' age and gender).

Again, as the interaction between age and gender is significant, the impact of these factors should be interpreted simultaneously. We can observe the following pattern: in both gender groups, older teenagers use fewer expressive markers, but the decrease is much stronger for the girls. In fact, for the boys, the decrease is marginal and not statistically significant. For the girls, on the other hand, the age difference is significant in all education groups. Furthermore, we see that at whatever age, girls always write in a more expressive way on social media than boys: this pattern holds and is statistically significant in all education groups, at all age points.

4.3 Submodel: Orality

The third model's response variable are the counts for all non-standard features that correspond to the orality maxim. The count for "oral non-standard markers" in the example below would be 1: only the Flemish colloquial pronoun *gij* belongs to the orality category, consequently the expressive markers and the chatspeak abbreviation *wrs* are not included.

Gij komt wrs met de fiets? 🖤🖤🖤😍😍😍 ('You are probably coming by bike?')

The best results were obtained with the following predictors: the interaction between age and gender and the interaction between gender and education. The model's predictions are visualized in Figure 3. The estimates and standard errors (compared to the reference category: younger girls in General Education) are presented in Table 7 and the Anova for the overall effect of the factors can be found in Table 8.

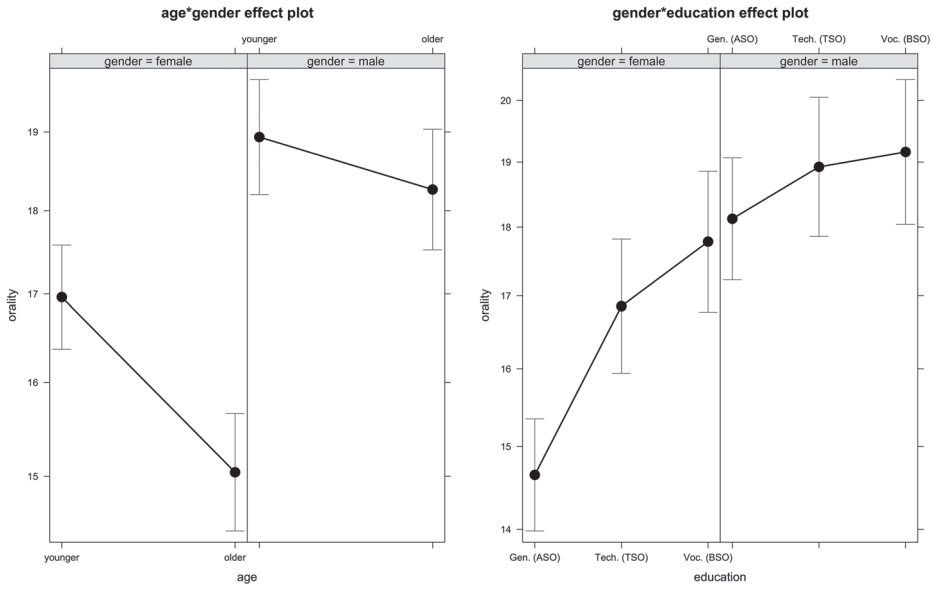

Fig. 3: Orality model: effect plot (predicted counts per 100 tokens)

	Estimate	Std. Error	z value	Pr(>|z|)	Signif.
(Intercept)	-1.86935	0.02521	-74.15	< 2e-16	***
ageOlder	-0.12019	0.02301	-5.22	1.75e-07	***
genderMale	0.17688	0.03776	4.68	2.81e-06	***
educationTechnical	0.14030	0.03719	3.77	0.000161	***
educationVocational	0.19390	0.03813	5.09	3.67e-07	***
ageOlder:genderMale	0.08413	0.03393	2.48	0.013157	*
genderMale:educationTechnical	-0.09709	0.05406	-1.80	0.072532	.
genderMale:educationVocational	-0.13829	0.05540	-2.50	0.012556	*
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Table 7: Orality model: fixed effects (reference category: younger girls in General Secondary Education)

	Chisq	Df	Pr(>Chisq)	Signif.
age	23.2491	1	1.423e-06	***
gender	41.4467	1	1.211e-10	***
education	24.8202	2	4.077e-06	***
age:gender	6.1478	1	0.01316	*
gender:education	6.9440	2	0.03105	*
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 8: Orality model: Anova

Table 8 shows that both higher order terms (i. e. age:gender and gender:education) have a significant impact on the adolescents' use of oral non-standard features on social media.

As for the interaction between age and gender, we can see that in both gender groups, older teenagers use fewer oral features than younger teenagers. For girls in all educational tracks, this decrease is strong and significant, whereas for boys, it is marginal and not statistically significant, in none of the educational tracks.

Regarding the interaction between gender and education, the data reveal a strikingly greater variety among the educational tracks for girls than among the ones for boys. For girls, the General school system is a clear outlier with the lowest scores for orality. The Technical and the Vocational systems overlap slightly. For boys, predictions for all three tracks overlap. Additional significance testing points out that at every age, girls in the General system significantly differ from girls in the other school systems, but that there is never a significant difference between girls in the two most practice-oriented education types. For boys however, at whatever age, no significant education difference can be found.

4.4 Submodel: Brevity

The final model's response variable are the counts for brevity-related non-standard features. The count in the example below would be 1: only the non-standard abbreviation *wrs* (for *waarschijnlijk*, 'probably') is included in the brevity category, and not the expressive hearts and faces or the colloquial pronoun *gij*.

Gij komt wrs met de fiets? 🖤🖤🖤😍😍😍 ('You are probably coming by bike?')

The most complex model that converges and scores best in terms of significance tests, includes age and the interaction between gender and education as predictors. Its predictions are visualized in Figure 4. The estimates and standard errors (compared to the reference category: younger adolescents in General Education) are presented in Table 9 and the Anova for the overall effects of the factors can be found in Table 10.

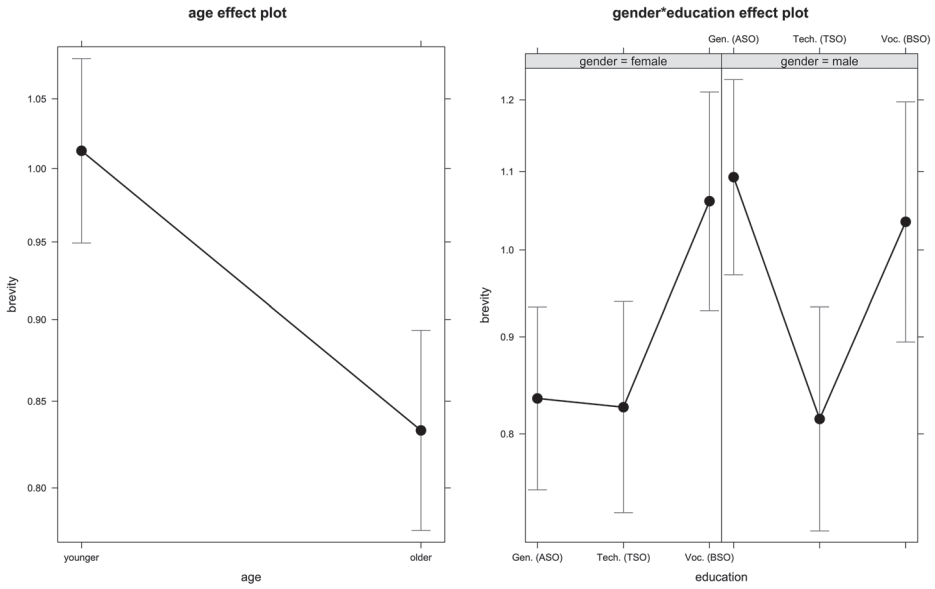

Fig. 4: Brevity model: effect plot (predicted counts per 100 tokens)

	Estimate	Std. Error	z value	Pr(>|z|)	Signif.
(Intercept)	-4.70158	0.05873	-80.05	< 2e-16	***
ageOlder	-0.19539	0.04215	-4.64	3.56e-06	***
genderMale	0.26851	0.08251	3.25	0.00114	**
educationTechnical	-0.01063	0.08623	-0.12	0.90189	
educationVocational	0.23929	0.08778	2.73	0.00641	**
genderMale:educationTechnical	-0.28281	0.12567	-2.25	0.02442	*
genderMale:educationVocational	-0.29353	0.12974	-2.26	0.02367	*
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Table 9: Brevity model: fixed effects (reference category: younger girls in General Secondary Education)

	Chisq	Df	Pr(>Chisq)	Signif.
age	21.4881	1	3.56e-06	***
gender	3.4892	1	0.061769	.
education	13.0270	2	0.001483	**
gender:education	7.1892	2	0.027471	*
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 10: Brevity model: Anova

Table 10 reveals that both age and the interaction between gender and education have a significant impact on adolescents' use of brevity-related features on social media. Young adolescents use more chatspeak abbreviations than older teenagers or young adults. This age difference is significant in all education types and for both girls and boys. The highest frequencies for abbreviations are attested in the data of students in Vocational Education and in those of the boys in General Education. Students in Technical Education score lower. Strikingly, gender differences are only apparent in General Education (in both age groups), with boys using significantly more non-standard abbreviations than girls. In the other educational tracks, no significant gender difference can be found, for none of the age groups.

5. Discussion

While the general model that combines all non-standard features reveals clear large-scale age, gender and education patterns in the data, the more specific models reveal distinct patterns for different kinds of non-standard writing. Below, we will compare and evaluate the results from the four different models.

A very consistent age pattern as well as a consistent interaction between age and gender can be found in the different models. The general model shows that the use of non-standard features in social media messages becomes less popular as teenagers grow older. Moreover, the decrease of non-standard features is much stronger in girls' CMC than in boys'. The submodels confirm this pattern for expressive as well as for oral features. For brevity-related features, however, age and gender do not interact, but the same consistent age pattern can be found, with older adolescents using fewer chatspeak abbreviations and acronyms than younger adolescents. The decreasing preference for non-standard features could be related to changing attitudes towards the linguistic standard or specifically towards standard writing norms. While, on a more subconscious level, these changing attitudes might be related to a decreasing pressure towards nonconformist behavior and an increasing acceptation of adult norms, we hypothesize that the youngsters' main concern is related to self-profiling for the peer group, striving for belonging and demonstrating "cool". As mentioned in Section 1, GRONDELAERS/VAN HOUT/VAN GENT (2016: 130) call the combination of standard language components and "socially meaningful non-standard features" a "linguistic tool for modern self-portrayal". However, the dosage of standard and non-standard features needs to be well-balanced in order for language use (in whatever context) to be perceived as "harmonious" (GRONDELAERS/VAN HOUT 2016: 67). And our results reveal that precisely that balance, and the sense of harmony attributed to it, seems to be different for younger adolescents compared to older ones. While younger adolescents seem to consider the abundant use of a wide range of non-standard features as cool and appear to use them for personal identity construction as well as for inclusion in the peer group (DE DECKER/VANDEKERCKHOVE 2017: 277–278, VERHEIJEN 2015: 129), young adults seem to evaluate this "excessive" use of non-standard markers as childish (VERHEIJEN 2015: 135).

However, while the general model suggests the existence of a significant age difference for all gender-education groups, the submodels for both oral and expressive features nuance this finding, revealing a significant age difference for girls only (in all educational tracks), and not for boys. For the latter only marginal differences can be found, which are insignificant in all education groups. This suggests that girls and boys derive different prestige from standard and non-standard markers in their late teens, and that especially girls turn away from non-standard markers (to some extent). The latter tendency confirms older sociolinguistic findings. TRUDGILL (1983: 182–183) for instance notes that (adult) women's preference for standard linguistic varieties cannot simply be transferred to (teenage) girls, as non-standard speech forms do not only appeal to (adult) men, but to youngsters of b o t h sexes. Since the preference pattern for younger girls and women differs, some sort of linguistic and attitudinal female "shift" must take place when adolescent girls reach adulthood. The strong decrease by age in the girls' non-standard writing attested in our corpus could be interpreted as evidence for such a shift. EISIKOVITS (2006) studies two groups of teenagers that are comparable to our participants in terms of age categories: she analyzes the (either standard or non-standard) realization of grammatical variables by 13-year old versus 16-year old adolescents. She finds largely the same pattern as the one resulting from our analyses, i. e. older girls using the non-standard variants significantly less often than younger girls, and older boys using them just as much or even more frequently than younger boys (EISIKOVITS 2006: 44–47). She ascribes these linguistic differences between adolescent boys and girls to a difference in attitude towards mainstream societal norms by the time the youngsters finish high school: while girls "are increasingly ready to accept external social norms" (EISIKOVITS 2006: 50), boys want to "affirm their own masculinity and toughness and their working class anti-establishment values" (EISIKOVITS 2006: 51). Our findings suggest that these attitudinal differences can be transferred to the online domain of social media: girls appear to aim more for a standard, adult linguistic "appearance" on social media as they grow older, whereas boys barely seem to adapt their online language practices, as far as the use of non-standard markers is concerned.

Interestingly, the submodels reveal strikingly different gender patterns for different types of non-standard writing on social media. While the expressive markers are more popular among girls, the typically oral features score higher among boys, for both genders at any age. For brevity-related features such as chatspeak acronyms and abbreviations, (significant) gender differences can only be attested in the theory-oriented General Education track, with the boys using more abbreviations than the girls. The divergent gender preferences for oral and expressive features might be related to gender-specific preferences for old versus new vernacular (ANDROUTSOPOULOS 2011: 146). Male preference for old vernacular, i. e. traditional, "tough" non-standardness, has been reported in many sociolinguistic studies (see for example EISIKOVITS 2006 quoted above). The current study does not only confirm this classical preference, it also suggests that it transcends genre and medium, and holds on new (digital) media and in new (online) peer networks as well, through the integration of oral features in written

discourse. Furthermore, our findings show a female preference for new vernacular and specifically for expressive chatspeak features, which also corresponds to previous findings: both in older sociolinguistic research and in more recent (CMC) studies, female discourse has been attested to be more expressive and stronger emotionally involved (ARGAMON et al. 2009, BARON 2004: 415, HILTE/VANDEKERCKHOVE/DAELEMANS 2018c, KUCUKYILMAZ et al. 2006: 282, PARKINS 2012: 48, 50–53, SCHWARTZ et al. 2013: 8–9, WOLF 2000: 831, and many more). This well-known gender pattern does not only persist in social media, it actually seems to gain visibility, through the availability of a wide range of relatively "new", explicitly expressive typographic features. Finally, the finding that gender does not impact the use of brevity-related features in Technical and Vocational Education, and that the gender difference in General Education is not very outspoken (odds ratio = 1.33), suggests that these shortening strategies – due to their mainly practical functionality – are indeed "stable markers of the genre" (DE DECKER/VANDEKERCKHOVE 2017: 277–278, see also: HILTE/VANDEKERCKHOVE/DAELEMANS 2018a: 18). In addition, the gender difference among General students indicates that teenage boys do sometimes show a preference for new vernacular features as well, i. e. when these features serve a practical rather than an expressive purpose.

As for the linguistic impact of educational track, a consistent pattern emerges from the different models: all types of non-standard features are more popular among vocational students, i. e. high school students in the most practice-oriented educational track who are trained for a manual (working class) profession. The higher frequency of oral features points towards a stronger adherence to old vernacular, which, once again, is in line with older sociolinguistic findings on social class patterns (LABOV 2001). However, the higher frequency of – mainly typographic – expressive markers and of non-standard abbreviations in the online discourse of these students reveals that these students are also attracted to new vernacular or modern/dynamic manifestations of non-standardness, i. e. non-standard markers that are the product of digital writing culture. Consequently, our findings suggest that teenagers in practice-oriented educational tracks pursue different types of social capital, i. e. both "dynamism" (typically associated with new vernacular) and "localness"/"toughness" (associated with old vernacular). We hypothesize that these correlations with educational track and specifically the relatively high scores for non-standardness in vocational students' CMC are impacted by both attitudinal factors and skills or proficiency. The latter might be explained in terms of the educational priorities in the educational tracks: while correct and formal standard Dutch writing is a major objective in theoretical school systems, it is much less of a priority in the practice-oriented tracks. A weaker familiarity with and possibly also a more limited proficiency in the formal written standard might thus influence these adolescents' writing practices on social media. As for possible attitudinal differences, we note that educational track is not only highly predictive of students' future professional career and social class belonging, on a micro-level, it largely determines their present peer networks and communities of practice. Moreover, offline peer networks (for example class groups) are often reflected in online networks, for example on social network

sites.[11] Since strong networks function as "norm enforcement mechanisms" (COATES 1993: 88) and "support localized linguistic codes" (MILROY/LLAMAS 2013: 409), it need not come as a surprise that different networks display different preferences. However, the patterns we attest here transcend these local networks or local communities of practice, since they seem to apply to entire educational tracks, no matter what class or school pupils come from. In other words, it seems like particular non-standard markers are more attractive, cool or prestigious amongst working class youngsters than amongst their middle-class peers.

In addition to the general education effect found in the different models, a more complicated and nuanced pattern emerges for the oral and brevity-related features. For the non-standard markers related to the principle of expressive compensation, education does not interact with any of the other social variables. For orality- and brevity-based features, however, it significantly interacts with the adolescents' gender. Although for the oral markers the same pattern can be found for girls and boys (i. e. more oral features are used by students in more practice-oriented education types), the tendency is much more outspoken for the girls (see also HILTE/VANDEKERCKHOVE/DAELEMANS 2018b). Among teenage girls, the variation between the three educational tracks is much larger than among boys. Furthermore, the education-related differences for orality markers are only significant for girls' CMC. In other words, girls seem to display a higher sensitivity to status and more status profiling for traditional vernacular features. As for the brevity-features, we note that for girls, Vocational students are outliers with the highest scores, whereas no significant difference can be found among female students in the two more theory-oriented tracks. Interestingly, male students in the most theory- and most practice-oriented tracks use about the same amount of abbreviations and acronyms, whereas boys in Technical Education use them significantly less often. In previous work, we already showed that the Technical students, holding a middle position on the continuum from practice to theory, do not always hold a middle position linguistically, but can also obtain the highest or lowest frequency scores for certain chatspeak features (see HILTE/VANDEKERCKHOVE/DAELEMANS 2018a and HILTE/VANDEKERCKHOVE/DAELEMANS 2018b)

6. Conclusion

The present study aimed at modeling adolescents' online writing practices in a most diverse way so as to lay bare more nuanced patterns of social and linguistic variation (compared to some previous studies with a narrower scope in terms of either the linguistic or social variables). In the end we wanted to find out to what extent different adolescent groups adhere to different social digilects. Therefore, we analyzed correlations between three parameters of the authors' socio-demographic profile (age, gender and education-

11 This phenomenon becomes apparent in our dataset, as many of the donated chat conversations are group conversations among all students of a specific class group.

al track) and their use of a wide variety of non-standard features in a large corpus of instant messages produced by teenagers. The use of generalized linear mixed models enabled the simultaneous inspection of the different predictors' linguistic impact as well as the inclusion of interactions between these predictors. Important contributions of the present study concern its multidimensional conceptualization of the linguistic and social variables, its inclusion of interactions between the social variables, and its systematic operationalization of the distinction between new and old vernacular features, and between expressive, oral and brevity-related non-standard markers.

Four models were fitted: one for all types of non-standardness, and three more specific submodels for features related to the chatspeak principles of expressive compensation, orality and brevity. Each model examined the impact of the adolescents' age, gender and educational track on their online writing practices. We can conclude that the similarities between the three submodels in terms of age, gender and education patterns were captured adequately by the general model. The more subtle but nevertheless important gender differences, however, were obfuscated in this model, and only became apparent when de-clustering the non-standardness category and fitting different models for distinct non-standard writing practices.

The data revealed higher frequencies for non-standard markers in texts written by younger adolescents (compared to older adolescents or young adults – this decrease by age was particularly strong for girls) and in texts written by students in Vocational Education (compared to students in more theory-oriented tracks). In addition, distinct gender preferences were found: while oral features (old vernacular features, such as the use of dialect lexemes) were more popular among teenage boys, expressive markers (new vernacular features, such as the use of emoticons) scored higher among girls. In other words, the toughness of old vernacular features seems to grant boys more "cool" on social media than the expressive markers that are extremely favored by girls, and vice versa. And students in practice-oriented tracks tend to invest stronger in both the toughness or "localness" of traditional vernacular and the dynamism of new digital vernacular than students with other educational backgrounds. So both seem to render them more social capital than their peers in more theory-oriented tracks. However, education appeared to have a stronger impact on girls' than on boys' online writing. Finally, brevity markers to some extent take a separate position, since they yield much less clear social patterns. For example: gender differences are much less outspoken and only reach significance (with low odds ratio) for one educational type. This may be related to the primarily functional rather than expressive nature of these brevity markers. But overall, we can conclude that, although Flemish adolescents may have access to the same pool of non-standard markers, the distinct social patterns for most features reveal that they do not share one and the same social digilect.

This study shows that there is more to the standard or non-standard nature of informal online writing than meets the eye: different social variables are at play and they do not only impact each other but also the selection of distinct strategies of non-standard writing. It may be clear from the above discussion that non-standard online writing cannot be operationalized as one homogeneous cluster of features, but should be consid-

ered in its complexity, as a combination of features representing different writing strategies and serving different purposes. We also argue that social variables cannot (solely) be studied in isolation, but that their combined impact should be examined as well, as potential interactions might emerge, like the ones discussed above between the adolescents' age and gender on the one hand, and between their gender and educational track on the other.

Finally, we note that the different linguistic features included in this study may represent very different kinds of non-standardness. Apart from the distinction between old and new vernacular, one can argue that some features are simply less "non-standard" than others within the genre of informal online writing: for example the insertion of an emoticon can be seen as less non-standard than the use of a dialect word. Some features that are or have become very characteristic of the genre, might even be perceived as the "standard" in informal online messages. One could argue that formal standard Dutch writing – without any typographic or lexical substandard markers – is less "standard" on social media than writing practices that do contain some of these markers of the genre. In this context, GRONDELAERS/VAN HOUT/VAN GENT (2016: 131) note that a conservative standard register does not necessarily sound neutral, but might even be linked to "superiority, and [a] condescending attitude towards chat styles and chat language". Therefore, in future work, we will address the question "what is standard on social media?" through a survey among high school students who match the profiles of the providers of the chat data discussed here. We will verify whether these students can identify and "correct" different non-standard items (including both common spelling mistakes and prototypical chatspeak markers), i. e. whether they can convert utterances that contain any of the linguistic markers discussed above into their formal standard Dutch equivalents. Furthermore, they will be invited to evaluate these markers on several dimensions (ranging from social attractiveness to status factors) and for several contexts (for example school writing versus social media writing). In this way we hope to gain insight in both the language skills of the target population and in their sociolinguistic attitudes. Furthermore, we will be able to examine whether certain prototypical chatspeak markers are still perceived as not belonging to the formal writing standard, or whether they have become the "new standard" in adolescents' eyes. This future study on the p e r c e p t i o n of computer-mediated communication will complement our previous and current work on the p r o d u c t i o n of this varied, fascinating linguistic register, as we will not only try to answer the question of h o w teenagers write on social media, but also w h y they appear to favor certain linguistic markers or styles.

7. Acknowledgements

We are very grateful towards Ella Roelant, Erik Fransen and Giovanni Cassani for their help and advice in the statistical modeling. We also thank Robert Grimm for the German translation of the abstract. Finally, we wish to thank the anonymous reviewers for their pertinent feedback on a previous version of this article.

References

ALLISON, P. D. (1999): Logistic regression using SAS: Theory and application. Cary, NC: SAS Institute Inc.

ANDROUTSOPOULOS, JANNIS (2011): Language change and digital media: A review of conceptions and evidence. In: KRISTIANSEN, TORE / COUPLAND, NIKOLAS (ed.): Standard languages and language standards in a changing Europe. Oslo: Novus, 145–161.

ARGAMON, SHLOMO / MOSHE KOPPEL / JAMES W. PENNEBAKER / JONATHAN SCHLER (2009): Automatically profiling the author of an anonymous text. In: Communications of the ACM. Inspiring women in computing 52 (2), 119–123.

BAMMAN, DAVID / JACOB EISENSTEIN / TYLER SCHNOEBELEN (2014): Gender identity and lexical variation in social media. In: Journal of Sociolinguistics 18 (2), 135–160.

BARON, NAOMI S. (2004): See you online: Gender issues in college student use of instant messaging. In: Journal of Language and Social Psychology 23 (4), 397–423.

BATES, DOUGLAS / MARTIN MAECHLER / BEN BOLKER / STEVEN WALKER (2017): Package 'lme4'. URL: <https://cran.r-project.org/web/packages/lme4/lme4.pdf>, accessed 24 June 2020.

COATES, JENNIFER (1993): Women, men and language. A sociolinguistic account of sex differences in language. London: Longman.

COXE, STEFANY / STEPHEN G. WEST / LEONA S. AIKEN (2009): The Analysis of Count Data: A Gentle Introduction to Poisson Regression and Its Alternatives. In: Journal of Personality Assessment 91 (2), 121–136.

DE DECKER, BENNY (2014): De chattaal van Vlaamse tieners. Een taalgeografische analyse van Vlaamse (sub)standaardiseringsprocessen tegen de achtergrond van de internationale chatcultuur. [Doctoral thesis, University of Antwerp, Antwerp].

DE DECKER, BENNY / REINHILD VANDEKERCKHOVE (2012): English in Flemish adolescents' computer-mediated discourse: A corpus-based study. In: English World-Wide 33 (3), 321–352.

DE DECKER, BENNY / REINHILD VANDEKERCKHOVE (2013): De integratie van Engels in Vlaamse jongerentaal kwantitatief en kwalitatief bekeken: das wel nice! :p. In: Nederlandse Taalkunde 18 (1), 2–34.

DE DECKER, BENNY / REINHILD VANDEKERCKHOVE (2017): Global features of online communication in local Flemish: Social and medium-related determinants. In: Folia Linguistica 51, 253–281.

EISIKOVITS, EDINA (2006): Girl-talk/Boy-talk: Sex Differences in Adolescent Speech. In: COATES, JENNIFER (ed.): Language and Gender. A Reader. Oxford: Blackwell, 42–54.

FMET = Flemish ministry of education and training (2017): Structuur en organisatie van het onderwijssysteem. In: Flemish ministry of education and training: Statistisch jaarboek van het Vlaams onderwijs. Schooljaar 2015–2016, 8–18.

GRONDELAERS, STEFAN / ROELAND VAN HOUT / PAUL VAN GENT (2016): Destandardization is not destandardization. In: Taal en Tongval 68 (2), 119–149.

GRONDELAERS, STEFAN / ROELAND VAN HOUT (2016): How (in)coherent can standard languages be? A perceptual perspective on co-variation. In: Lingua 172–173, 62–71.

GRONDELAERS, STEFAN / DIRK SPEELMAN (2013): Can speaker evaluation return private attitudes towards stigmatised varieties? Evidence from emergent standardisation in Belgian Dutch. In: KRISTIANSEN, TORE / STEFAN GRONDELAERS (ed.): Language (De)standardisation in late Modern Europe: Experimental Studies. Oslo: Novus, 171–192.

HARRISON, XAVIER A. (2014): Using observation-level random effects to model overdispersion in count data in ecology and evolution. In: PeerJ 2, e616.

HILBE, JOSEPH M. (2011): Modeling count data. In: LOVRIC, MIODRAG (ed.): International Encyclopedia of Statistical Science. Berlin: Springer, 836–839.

HILTE, LISA / WALTER DAELEMANS / REINHILD VANDEKERCKHOVE (2018): Predicting Adolescents' Educational Track from Chat Messages on Dutch Social Media. In: Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Association for Computational Linguistics (ACL): Stroudsburg, 328–334.

HILTE, LISA / REINHILD VANDEKERCKHOVE / WALTER DAELEMANS (2018a): Adolescents' social background and non-standard writing in online communication. In: Dutch Journal of Applied Linguistics 7 (1), 2–25.

HILTE, LISA / REINHILD VANDEKERCKHOVE / WALTER DAELEMANS (2018b): Social Media Writing and Social Class: A Correlational Analysis of Adolescent CMC and Social Background. In: International Journal of Society, Culture and Language 6 (2), 73–89.

HILTE, LISA / REINHILD VANDEKERCKHOVE / WALTER DAELEMANS (2018c): Expressive markers in online teenage talk: A correlational analysis. In: Nederlandse Taalkunde 23 (3), 293–323.

HOLMES, JANET (1992): An Introduction to Sociolinguistics. London/New York: Longman.

ISMAIL, NORISZURA / ABDUL AZIZ JEMAIN (2007): Handling Overdispersion with Negative Binomial and Generalized Poisson Regression Models. In: Casualty Actuarial Society Forum, 103–158.

KILLERMANN, SAM (2014): Breaking through the Binary: Gender As a Continuum. In: Issues 107, 9–12.

KRISTIANSEN, TORE (2001): Two Standards: One for the Media and One for the School. In: Language awareness 10 (1), 9–24.

KRISTIANSEN, TORE / PETER GARRETT / NIKOLAS COUPLAND (2005): Introducing subjectivities in language variation and change. In: Acta Linguistica Hafniensia 37 (1), 9–35.

KUCUKYILMAZ, TAYFUN / B. BARLA CAMBAZOGLY / CEVDET AYKANAT / FAZLI CAN (2006): Chat Mining for Gender Prediction. In: International Conference on Advances in Information Systems. Berlin: Springer, 274–283.

LABOV, WILLIAM (2001): Principles of Linguistic Change. Volume 2: Social Factors. Maiden: Wiley-Blackwell.

MILROY, LESLEY / CARMEN LLAMAS (2013): Social Networks. In: CHAMBERS, J. K. / NATALIE SCHILLING (ed.): The Handbook of Language Variation and Change. Second Edition. Oxford: Blackwell, 409–427.

PARKINS, RÓISÍN (2012): Gender and Emotional Expressiveness: An Analysis of Prosodic Features in Emotional Expression. In: Griffith Working Papers in Pragmatics and Intercultural Communication 5 (1), 46–54.

SCHWARTZ, H. ANDREW / JOHANNES C. EICHSTAEDT / MARGARET L. KERN / LUKASZ DZIURZYNSKI / STEPHANIE M. RAMONES / MEGHA AGRAWAL / ACHAL SHAH / MICHAL KOSINSKI / DAVID STILLWELL / MARTIN E. P. SELIGMAN / LYLE H. UNGAR (2013): Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. In: PLoS ONE 8.

TRUDGILL, PETER (1983): Social identity and linguistic sex differentiation. Explanations and pseudo-explanations for differences between women's and men's speech / Sex and covert prestige. Linguistic change in the urban dialect of Norwich. In: TRUDGILL, PETER (ed.): On dialect. Social and Geographical Perspectives. Oxford: Blackwell, 161–185.

VARNHAGEN, CONNIE K. / G. PEGGY MCFALL / NICOLE PUGH / LISA ROUTLEDGE / HEATHER SUMIDA-MACDONALD / TRUDY E. KWONG (2010): Lol: New language and spelling in instant messaging. In: Reading and Writing 23 (6), 719–733.

VERHEIJEN, LIEKE (2015): Out-of-the-ordinary orthography: The use of textisms in Dutch youngsters' written computer-mediated communication. In: GONZÁLEZ TEMER, VERÓNICA / JELENA HORVATIC / DAVID O'REILLY / AIQING WANG (ed.): Issue 2: Proceedings of the second Postgraduate and Academic Researchers in Linguistics at York (PARLAY 2014) conference, 12th September 2014 (York Papers in Linguistics PARLAY Proceedings Series. 1), 127–142.

VERHEIJEN, LIEKE / LAURA DE WEGER / ROELAND VAN HOUT (2018): Code-Mixing with English in Dutch Youths' Online Language: OMG SUPERNICE LOL! In: VANDEKERCKHOVE, REINHILD / DARJA FIŠER / LISA HILTE (ed.): Proceedings of the 6th Conference on Computer-Mediated Communication (CMC) and Social Media Corpora. 17–18 September 2018, University of Antwerp. Antwerp: University of Antwerp, 63–67.

WOLF, ALECIA (2000): Emotional Expression Online: Gender Differences in Emoticon Use. In: CyberPsychology & Behavior 3 (5), 827–833.

ZAPPAVIGNA, MICHELE (2015): Searchable talk: The linguistic functions of hashtags in tweets about Schapelle Corby. In: Global Media Journal: Australian Edition 9 (1).

ZEILEIS, ACHIM / CHRISTIAN KLEIBER / SIMON JACKMAN (2008): Regression models for count data in R. In: Journal of Statistical Software 27 (8), 1–25.

Corpora

A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Brown). (1964, 1971, 1979). Compiled by W. N. FRANCIS and H. KUČERA. Brown University. Providence, Rhode Island, USA.

ANW = Algemeen Nederlands Woordenboek. URL: <http://anw.inl.nl/>

COCA = Corpus of Contemporary American English. URL: <https://corpus.byu.edu/coca/>

DPC = Dutch Parallel Corpus. URL: <https://www.kuleuven-kulak.be/DPC>

Named Entity Recognition (NER) Datasets. CLiPS research center, University of Antwerp. URL: <https://www.clips.uantwerpen.be/conll2002/ner/data/>

Roularta Consortium (2011): Roularta corpus.

SoNaR = Stevin Nederlandstalig Referentiecorpus. URL: <https://ivdnt.org/downloads/tstc-sonar-corpus>

LISA HILTE

Stadscampus, Prinsstraat 13, 2000 Antwerpen

e-mail: <lisa.hilte@uantwerpen.be>

REINHILD VANDEKERCKHOVE

Stadscampus, Prinsstraat 13, 2000 Antwerpen

e-mail: <reinhild.vandekerckhove@uantwerpen.be>

WALTER DAELEMANS

Stadscampus, Prinsstraat 13, 2000 Antwerpen

e-mail: <walter.daelemans@uantwerpen.be>