Faculteit Letteren en Wijsbegeerte
Departement Taalkunde

The social in social media writing:

The impact of age, gender and social class indicators on adolescents' informal online writing practices

Proefschrift voorgelegd tot het behalen van de graad van doctor in de taalkunde aan de Universiteit Antwerpen te verdedigen door

Lisa Hilte

Promotoren prof. dr. Reinhild Vandekerckhove
prof. dr. Walter Daelemans

Antwerpen, 2019

The social in social media writing: The impact of age, gender and social class indicators on adolescents' informal online writing practices
Nederlandse titel: *Sociale patronen in taalgebruik op sociale media: De invloed van leeftijd, gender en sociale-klasse-indicatoren op de informele online schrijfpraktijk van tieners*

The research in this thesis was supported by the FWO (Research Foundation Flanders) under grant G041115N.

Cover art: Lisa Hilte
Cover design: Anita Muys
Printed by: Universitas, Antwerpen

PREFACE

Preface

In many ways, conducting my PhD research has been a team effort. In this preface, I would like to thank the people whose advice, insights, feedback, moral support or company have been of great value to me during these past four years.

First and foremost, I wish to thank my supervisors Reinhild and Walter for guiding me on this journey. It has been a pleasure and an honor to work with them. I consider myself lucky to have been able to learn from the best in both Sociolinguistics and Computational Linguistics, and I believe the interaction we created between these two fields was most fruitful.
Before accepting this job offer, I was told by Reinhild's previous PhD student that if he could go back in time, he would absolutely work on a PhD again, "especially with Reinhild as a supervisor". Now, nearly four years later, I can say that I agree with him completely. I am grateful to Reinhild, with whom I collaborated most closely, for always brainstorming with me, for helping me improve my research skills as well as my academic writing, for always being there when I had any questions or doubts, and for her kindness and understanding at all times.
I want to thank Walter for his valuable help and advice during this project, especially with the more technical and computational aspects of the research, but also for his pertinent remarks and questions with respect to the sociolinguistic components, and for his ever prompt feedback, even at times when I know my timing must have been inconvenient. In addition, I thank Walter for welcoming me into the Computational Linguistics group in the L-building: for a long time, I was the only student of our research group in my building, but I never felt like an outsider at CLiPS.

I also wish to express my gratitude to the members of the jury, Dominiek Sandra, Steven Gillis, Darja Fišer and Roeland van Hout, for investing their time in evaluating my thesis and attending the defense. I appreciate their effort and feedback, and hope that they will enjoy reading this dissertation.

Next, I cannot emphasize enough how crucial the contribution was of each principal, teacher, parent and of course high school student who participated in this research project. Without their enthusiasm, collaboration and trust, this PhD project would have never existed.

An important part of why I enjoyed working on this research project so much was not only the work or the topic, but also the atmosphere on campus.
I am grateful to all my colleagues at CLiPS for creating such a fun and inspiring work environment, as well as for their pertinent feedback on my research. I enjoyed our meetings,

lunch breaks, parties, and the conferences we attended. Thanks go to the labmeeting “crew”, Tim, Pieter, Giovanni, Stéphan, Madhumita, Simon and Nicolae, to the third floor team, Jolien, Hanne, Nathalie, Pietro, Ilke, and Lotte, to my (fairly) new colleagues Hanne, Edwige and Rianne, to Sarah, Mike, Guy, Elyne, Enrique and Ben and to my former colleagues Ihor, Robert and Ben. Special thanks go to Stéphan and Giovanni for their help and advice on more technical aspects of my research on many occasions.
Apart from the colleagues in my own research group, I would like to thank Lieke – who works on a very similar topic in the Netherlands and with whom I attended many conferences – for all the fun we had, in the Low Countries and abroad. I am also grateful to my non-CLiPS colleagues at the university with whom I shared many great moments and who were always there for either serious talks, laughs or shameless celebrity gossip, in particular Elvira, Wouter, Hava and of course Anna, who has been in this PhD journey with me from the very beginning and has been my number one listening ear on campus. Last but not least, I thank my office mate Katrien. I enjoyed all the time we spent discussing books, movies, love, listening to music together and sometimes even being silent and actually getting some work done ☺.

My final thanks go to my parents, for their unconditional support in all my endeavors since 1992, and to my partner Thomas, who has always shown an interest in my research and supported me, whose expertise in computer science was of great help with regard to several technical issues I encountered, and who has been my rock during all these years.

Lisa Hilte
Antwerp, 2019

TABLE OF CONTENTS

Table of contents

CHAPTER 1

Introduction

Over the past decades, processes of digitalization and the creation of new, electronic, media have been of great linguistic importance, as new ways of communicating and new text genres have emerged. The genre of *instant messaging* (i.e. online chat conversations) in particular is highly relevant from a linguistic point of view, as it mirrors (informal) conversational speech more closely than any other written genre and has forever disproved the idea that written language use would always be more formal than spoken language. In addition, it is generally believed that people nowadays produce more written texts than ever before, as interactions on social media and instant messaging platforms such as WhatsApp and Facebook Messenger have become a major aspect of daily communication. Furthermore, some of these platforms or media are extremely popular and easily accessible, and are therefore used by people with diverse socio-demographic profiles (in terms of e.g. age, gender, and social or educational background). Consequently, these chat conversations or instant messages offer highly valuable data sources for variationist sociolinguistic studies.

The present research project specifically focuses on the most ardent users of new media, i.e. the adolescent generation, and aims to lay bare correlations between their online writing practices and various aspects of their socio-demographic profile (including age, gender, and several social class indicators – see below). The above mentioned processes of digitalization and the creation of new media are generally assumed to have led to a "pluralisation of written language norms" (Androutsopoulos 2011, 146; see also Grondelaers et al. 2016, 143), as the new text genres and corresponding linguistic practices that have emerged often deviate from formal writing standards in several respects (e.g. in terms of spelling or typography). Such deviations appear to be especially omnipresent in online texts produced by *youths*. We note that this phenomenon transcends the context of digital media, as non-standard language use – possibly related to a non-conformation to adult standards in general – has repeatedly been reported to peak during adolescence, in both on- and offline contexts (see e.g. Coates 1993; Holmes 1992; Tagliamonte 2016), which is assumed to be due to "group pressure to not conform to established societal conventions" (Nguyen et al. 2016, 17). However, while the wide variety of deviations from standard language norms present in teenagers' online writing might create the impression that 'anything goes' – and informal online writing has e.g. been characterized in terms of "linguistic whateverism" (Baron 2008, 169) – researchers seem convinced that the genre "has its own rules rather than that it follows no rules whatsoever" (Verheijen 2013, 584). Furthermore, previous findings suggest that these 'rules' or conventions are – to a certain degree – socially determined, as distinct social groups appear to favor certain linguistic markers to different extents. However, while some socio-demographic characteristics, such as gender, have often been examined with respect to their

correlation with online writing practices, others remain highly underresearched. The impact of people's – and especially *adolescents'* – social class on their online writing, for instance, has hardly been analyzed at all. Furthermore, certain widely accepted patterns of (e.g. gender-based) sociolinguistic variation have only been examined for middle and upper class youths, while the writing practices of working class youths remain largely unexplored, although divergent tendencies might actually be attested in these adolescents' writing.

Therefore, the main aim of this dissertation is to lay bare correlations between teenagers' online writing practices and their socio-demographic profile through a diversified operationalization of the variables and their interactions, and with a strong focus on underresearched aspects of this subject (such as the linguistic impact of social class factors – see below). In order to obtain a nuanced and complete image of youths' online communication, both the linguistic and social variables included in the research project are operationalized as multifaceted, complex phenomena consisting of multiple parameters.
As for the linguistic variables, a wide variety of deviations from formal standard writing norms is included, as well as some more general (i.e. not specifically bound to new media or adolescent language use) textual properties. The basic premise is that systematic analysis and comparison of adolescents' use of different linguistic markers might reveal different 'rules' of linguistic conduct for different social groups.
As for the social variables, several socio-demographic characteristics are taken into account, i.e. the adolescents' age and gender, but also several parameters of their social class background, such as their educational track, home language and the profession of their parents. While age and – as mentioned above – gender are often the focus of sociolinguistic studies on adolescent speech, the other social variables are less prominent (although there are some well-known exceptions, e.g. Eckert 2000) or nearly completely absent when it comes to online communication. However, these factors are major determinants of adolescents' social profile, and may thus strongly influence this group's communicative and linguistic practices. Furthermore, the participation of a large group of working class teenagers fills a scientific gap with respect to the analysis of online communication.
Another methodological contribution of the dissertation concerns the systematic inclusion of (potential) interactions between the social variables, since the linguistic impact of different social predictors may not be independent (e.g. divergent gender patterns may be observed for teenagers in distinct social classes or in different educational tracks). The inclusion of interactions in the research design may not only complement or nuance the tendencies reported in previous work (in which social determiners are often examined in isolation), but might even challenge common sociolinguistic findings.

The dissertation is a collection of seven research articles (Chapters 2 to 8) each of which focuses on a different aspect of the same main topic – consequently, the dissertation combines multiple perspectives on sociolinguistic variation in youths' online writing. In general, the research project is of an interdisciplinary nature, as it brings together the scientific fields of (variationist) sociolinguistics and (stylometric) computational linguistics

through the quantitative-correlational and computational approach to traditional sociolinguistic research questions. Furthermore, the dissertation includes chapters with a focus on teenagers' *production* as well as a chapter with a focus on their *perception* of the genre. Consequently, we do not only try to answer the question of *how* youths write in an informal online setting, but also *why,* e.g. by verifying to which extent the (in)frequent use of certain linguistic features is actually related to an explicit (non-)appreciation of these features. To a minor extent, the teenagers' linguistic skills (e.g. spelling skills) and their register sensitivity with respect to non-standard markers of chatspeak are examined too. Finally, the dissertation brings together research on teenagers' 'traditional' and 'digital' literacy, by examining the extent to which they use several linguistic repertoires in an informal online setting, i.e. a (traditional) verbal repertoire and a (digital) typographic repertoire. Such a comparison may complement findings on teenagers' adherence or non-adherence to formal writing standards (i.e. the 'school' standard) with findings on their familiarity with and creativity with regards to the new communicative possibilities of digital media.

In the sections below, the main aims and the general design of the research project are presented. In Section 1, the most important research questions of the dissertation are discussed. Section 2 presents the social media corpora that are examined in the different chapters, and explains the data collection and preprocessing procedures. Next, Section 3 gives an overview of the linguistic and social variables included in the research design, and Section 4 briefly summarizes the statistical data processing. Finally, Section 5 presents an overview of the different chapters (i.e. research articles) included in this dissertation.

1. Research questions

This section presents the three main (clusters of) research questions addressed in the present dissertation, and refers to the specific chapters in which they are examined. We also refer to the relevant chapters for discussions of the state of the art concerning the research topics and questions they address. Table 1 at the end of this section summarizes the different research questions and the specific chapters in which they are addressed.

RESEARCH QUESTION 1: Which patterns of sociolinguistic variation can be attested in adolescents' informal online writing with respect to age, gender and social class indicators such as educational track and parental profession? Can significant interactions between these socio-demographic variables be observed?

Linguistic practices tend to be significantly influenced by the socio-demographic characteristics of the speaker or writer. The main aspects of Flemish teenagers' socio-demographic profile that will be examined in this research project are age, gender and social class indicators such as educational track and parental profession (see Section 3.2 for a more detailed discussion of all included social variables). We note that educational track is

operationalized as a separate variable in certain chapters, and as a subfactor of teenagers' social class background in other chapters (see below).

EDUCATIONAL TRACK

Educational track is rarely included as a variable in related (sociolinguistic or computational linguistic) research on (adolescents') online writing practices. However, it is an essential part of youths' social profile: it strongly influences the composition of their (online as well as offline) peer group networks, and it may have a major impact on their future professional career (de Jager, Mok, & Sipkema 2009) – for instance, Glorieux et al. (2014, 77) report that the type of secondary education that youths attend is by far the strongest determinant of their chances of attending higher (tertiary) education. We want to find out to which extent educational track influences teenagers' informal online communication, and whether divergent writing practices can be observed for adolescents with distinct educational backgrounds (for instance w.r.t. how they deal with the communicative possibilities of new media).
The articles presented in Chapters 3 to 5 and Chapter 7 include educational track as an independent variable, whereas in Chapter 6, it is the dependent variable (see research question 2).

SOCIAL CLASS

Educational track is an important (though not the sole) factor of teenagers' social class (see above). While social class is quasi absent as a variable in social media research, it has been included in quite a lot of older sociolinguistic work on spoken language (e.g. Labov 1972; Trudgill 1983), although these studies mostly include *adults'* social class. Only a limited number of analyses can be found in which some notion of *youths'* social class is included: Eckert (2000), for instance, examines the oral language use of two groups of high school students who occupy different social positions in the school system and come from different social backgrounds. Furthermore, previous studies on youths' linguistic practices mainly include participants with middle or upper class backgrounds, though there are some exceptions (e.g. Eisikovits 2006). The participation of a large group of working class youths and the inclusion of adolescents' social class (conceptualized in terms of educational track, home language and parental profession – see Section 3.2) as a variable in the research design enable us to investigate research questions such as whether distinct (online) writing practices can be observed for teenagers with different social class profiles, and whether working class youths connect to the international digital writing culture to the same extent as their peers with a middle or upper class background.
Social class is the main focus of Chapters 3 and 4.

AGE AND GENDER

In addition to educational track and social class, the impact of teenagers' age and gender is examined too. These two variables have been included in previous studies on online

communication, and some tendencies have been attested repeatedly. For instance, a consistent gender difference in informal online communication concerns a more frequent use of emoticons or emoji by women/girls than by men/boys (see e.g. Baron 2004, 415; Herring & Martinson 2004, 436; Kucukyilmaz et al. 2006, 282; Parkins 2012, 52; Schwartz et al. 2013, 8). As for age, it has been attested that younger teenagers use more emoticons and other prototypical (non-standard) chatspeak markers than older teenagers or young adults (see e.g. De Decker 2014, 263-264; Tagliamonte & Denis 2008, 13; Verheijen 2015, 135-136; Verheijen 2016, 283, 285). However, these previous findings may be limited in several respects and deserve further examination (see below).
Age and gender are explicitly included as social variables in Chapters 2, 5 and 7. However, in *all* chapters, we control and correct for their potential linguistic influence.

INTERACTIONS BETWEEN SOCIAL VARIABLES

The previously attested findings on age and gender may be limited in multiple respects. First of all, the social variables are often studied in isolation, whereas we hypothesize that they might rather interact with each other rather than be independent (e.g. gender patterns may be different in distinct age groups). Furthermore, as mentioned above, many studies only include middle and upper class participants, which implies that the reported age and gender tendencies may not be as universal or general as initially thought, since distinct tendencies may actually emerge for youths with different social or educational backgrounds. Consequently, the systematic operationalization and inclusion of interactions between multiple socio-demographic characteristics is a relevant contribution of the dissertation, as it may complement, nuance or even challenge previous (socio)linguistic findings.
Interactions are analyzed in Chapters 5, 7 and 8.

RESEARCH QUESTION 2: Are sociolinguistic variation patterns in youths' informal online writing sufficiently robust to be used in quantitative (descriptive and predictive) modeling?

The second research question concerns the modeling of teenagers' online writing practices. The dissertation addresses the question whether the attested sociolinguistic variation patterns (with respect to age, gender and educational track) in teenagers' online conversations are sufficiently strong and robust to be used in quantitative modeling. Both descriptive and predictive models are explored, as the dissertation includes chapters on models in two directions: i.e. models that use authors' socio-demographic characteristics as predictors to model their linguistic practices, and models that use (linguistic properties of) text samples to predict aspects of authors' socio-demographic profile.

DESCRIPTIVE MODELS: LINGUISTIC PRACTICES AS RESPONSE VARIABLE

First, we examine whether teenagers' online writing practices can be modeled accurately, given relevant parameters of the authors' socio-demographic profile (i.e. age, gender and educational track). Models are built to estimate teenagers' use of different (sets of) linguistic

features based on the authors' social profile. Contributions of the dissertation in this respect are the simultaneous inspection of multiple social variables, the inclusion of potential interactions between these social variables (as mentioned above) and the systematic comparison of different models for different types of linguistic features.
These models are discussed in Chapters 5 and 7.

PREDICTIVE MODELS: EDUCATIONAL TRACK AS RESPONSE VARIABLE

The dissertation also includes a pilot study in which it is verified whether teenagers' educational track can be predicted based on a sample of their online writing. The performance of different types of models, using different feature sets – i.e. linguistic properties derived from the text samples – is compared. The findings with respect to these models and to the ones described above may be complementary, as the fields of variationist sociolinguistics and author profiling (i.e. in which the task is to predict people's profile based on their language use) are inherently related, approaching the same topic or problem from different, opposite perspectives, and using different methods. While social variables such as age and gender are often the focus of studies in author profiling (see e.g. the overview paper by Reddy et al. 2016), education is seldom included, and – to our knowledge – never for youths. In addition, the research focus of (the very limited number of) education profiling studies has up until now never concerned the text genre of social media writing, and Dutch has not been the language of interest. In this respect, the included pilot study fills a gap and opens up paths for further research.
This pilot study is presented in Chapter 6.

RESEARCH QUESTION 3: Do teenagers' attitudes on their peers' online writing practices reflect the attested sociolinguistic patterns? Or do discrepancies emerge between adolescents' production and perception of the linguistic genre?

Finally, after having focused extensively on Flemish adolescents' *production* of informal online writing, we will examine their *perception* of or attitudes on the genre. While various attitudinal studies can be found on people's perception of the potential (negative or positive) impact of informal writing on youths' literacy (see e.g. Verheijen 2018, 40-49, for an extensive overview), research on people's perception of the (characteristics of the) genre itself is mostly absent.

We expand the research question of *how* adolescents write in their informal online communication to *why* they tend to do so. A survey is conducted among high school students, in order to find out more about e.g. adolescents' (non-)appreciation of linguistic markers, or about their adherence to or reluctance towards standard language norms and ideologies. Furthermore, we investigate whether teenagers, when conversing online, tend to follow certain (implicit) rules of linguistic conduct (see above), and, if so, whether these rules differ for distinct groups of youths. In addition, it is verified to what extent teenagers' reported attitudes reflect previously attested sociolinguistic patterns in online writing, since different

degrees of (non-)appreciation of certain features may explain divergent writing practices. So it is investigated whether there are convergent or divergent tendencies for production versus perception.
These questions are addressed in Chapter 8.

Research question	Addressed in
1.1. Linguistic variation w.r.t. age	Chapters 2, 5, 7
1.2. Linguistic variation w.r.t. gender	Chapters 2, 5, 7
1.3. Linguistic variation w.r.t. educational track	Chapters 3-5, 7
1.4. Linguistic variation w.r.t. social class	Chapters 3-4
1.5. Linguistic variation: interactions between social variables	Chapters 5, 7-8
2.1. Descriptive modeling: linguistic practices as response variable	Chapters 5, 7
2.2. Predictive modeling: educational track as response variable	Chapter 6
3. Attitudes on and perceptions w.r.t. informal online writing	Chapter 8

Table 1: Overview of the research questions and the chapters in which they are addressed

2. Corpus

Below, we present the social media data that are analyzed in the dissertation. The data collection (Section 2.1) and preprocessing (Section 2.2) are described, and an overview of the distributions in the final corpus is presented (Section 2.3). We note that since the research papers included in the dissertation are all published, accepted or submitted as separate articles, each of them inevitably briefly discusses the corpus and data collection again.

2.1. Data collection

The collection of a large and representative corpus of teenagers' private conversations on social media platforms proved to be one of the biggest challenges of the research project, and took one year and a half to be completed. Below, we describe how the social media texts and the metadata were obtained (Section 2.1.2), but first, we briefly present an older corpus which just like the new corpus was compiled within the research group CLiPS (University of Antwerp) (Section 2.1.1). This older corpus is examined in the first research paper of the dissertation.

2.1.1. Corpus 2007-2013

PROPERTIES

The first research article that is included in this dissertation (Chapter 2) presents linguistic analyses conducted on a corpus that was collected prior to this doctoral research project, by

students and researchers[1] of the University of Antwerp. This corpus contains more than 400,000 informal online posts (over 2 million tokens), produced between 2007 and 2013 by Flemish adolescents aged 13 to 20. The corpus consists of both instant chat messages (produced on MSN / Windows Live Messenger and Facebook Messenger) and social media posts (produced on the social network site Netlog). While the first group of texts are private, synchronous (i.e. real-time) messages, the second group mostly consists of public, asynchronous (i.e. not real-time) posts. Finally, the corpus contains metadata on the informants' profile, more specifically on age, gender and regional background. For a detailed overview of this dataset, we refer to Chapter 2 and to De Decker (2014, 23-28).

LIMITATIONS

While this corpus contains a large, diverse and highly relevant sample of youths' informal online writing, the collection of a more recent dataset was necessary, since the genre of (especially teenagers') informal online communication is constantly evolving and changing. Furthermore, the research questions of the present dissertation (see Section 1) required additional metadata to be included in the new corpus, such as information on the teenagers' educational and social background. However, the availability of the 2007-2013 corpus enabled us to start working on the code for the automated feature extraction and test it out on actual social media texts, and to learn and apply new statistical techniques while collecting the new dataset.

2.1.2. Corpus 2015-2016

TARGET GROUP

In order to collect a large sample of high school students' spontaneous chat conversations, we collaborated with twelve secondary schools, all situated in the province of Antwerp. Consequently, region (operationalized as the province in which the participants live) is a (quasi-)constant. The students of eleven out of these twelve schools almost exclusively lived in the province of Antwerp. One secondary school with an excellent reputation for culinary training had a more interregional profile and attracted some students from other provinces too. While keeping region a constant, we tried to reach a varied group of youths in terms of age, gender and educational track. The participation of a large group of Flemish adolescents with varied socio-demographic profiles does not only render a more representative sample, it was also required in order to address the research questions (see Section 1). We visited class groups in the three main tracks of Belgian secondary education, i.e. General, Technical and Vocational Secondary Education, which range from a very strong theoretical to a very strong practical orientation (see Section 3.2 for more information on these tracks). With respect to the teenagers' age, a varied audience was targeted too, as we visited class groups

[1] We are grateful to Benny De Decker and Guy De Pauw for allowing us to analyze the corpus they collected, and for their help and technical support whenever we had questions about the data.

in all six (or seven[2]) years of secondary education. The participants in the corpus are between thirteen and twenty years old. We note that while most secondary school students are under the age of nineteen, we included slightly older teenagers too, as long as they were still in secondary school. Finally, with respect to gender, we visited class groups that were fairly balanced (containing more or less the same number of female and male students) as well as class groups with a strong imbalance. These 'imbalanced' class groups mostly concerned practice-oriented educational tracks related to more prototypically "gendered" professions, such as hair dressing or car mechanics.

INITIAL PHASE: SCHOOL VISITS

In the initial phase of the data collection, we sent out letters to principals of secondary schools in the province of Antwerp to introduce ourselves and the research project and ask permission for school visits. The principals who agreed sent us the contact information of teachers that were willing to participate. Most of these teachers taught Dutch (in General and Technical Secondary Education), professional communication (in certain vocational tracks) or *PAV*[3] (in Vocational Secondary Education). They generally offered us a timeslot of one hour, during which we met the students and gave a presentation about our project and related research. We introduced the students to the topics of sociolinguistics and computational linguistics, and to research on online communication in particular. Next, we informed the students about the setup and goals of our own study and about our data collection, and demonstrated how donated utterances were processed and anonymized (see Section 2.2). Finally, we showed the students how they could donate their own chat conversations. We note that participation was entirely voluntary: the students were free but not obliged to submit material. We asked the students who donated conversations (and if they were minors, their parents or guardians too) for consent to store and analyze their anonymized utterances. The students who participated filled in a form in which they were asked to provide the relevant metadata (e.g. their year of birth, their gender, etc.). These metadata were required to analyze but also to anonymize the texts, as the students' names were deleted from the database, but the relevant profile information was kept (see Section 2.2). All parents received a letter containing our contact information, information about the research project, and a consent form. Some parents actually contacted us and asked to be kept informed about our findings. Finally, we note that the participating students, at all time, kept the right to withdraw their submission – which implied that their text samples would be deleted from the dataset. However, this situation occurred only once.

[2] Some practice-oriented tracks offer an additional seventh (specialization) year.

[3] *PAV* stands for *Project Algemene Vakken* or 'Project general courses', and integrates a variety of general courses (e.g. Dutch, mathematics, history, etc.). The course has a practical focus, as skills and knowledge related to the different general courses are to be used and applied in student projects.

SUBMISSIONS

We welcomed all chat conversations or instant messages that were produced on the social media platforms (and/or smartphone apps) of Facebook Messenger and WhatsApp. These private (i.e. non-public) conversations could be dyadic (i.e. one-on-one) chats or group chats (i.e. including more than two interlocutors). The students were free to submit their own selection of either entire conversations or parts of conversations, as long as these were produced before the time of our school visits, and as long as the main language of the conversation was Dutch. The former condition was meant to exclude the observer's paradox. The latter condition implied that, while codeswitching could still occur in the messages (e.g. the insertion of English words or phrases in Dutch utterances), entire conversations in a language other than Dutch were excluded, as these fall outside the scope of the present research. All data were collected between 2015 and 2016. Consequently, most of the chat conversations were also produced during this time span; more recent messages naturally do not occur in the dataset, but a small proportion of older messages does (i.e. 12% of all messages), as the participants were free to search their chat history and donate older conversations.

CITIZEN SCIENCE

This procedure of intense collaboration with schools and personal contact with students and teachers resulted in the collection of a corpus of more than 400,000 messages (over 2.5 million tokens), produced by more than 1000 adolescents in the context of private and spontaneous online chat conversations. Furthermore, the corpus contains detailed metadata: each participant's year of birth, gender and educational track is known. For a subset of participants, additional information on their social background is known, such as the language(s) they speak at home and their parents' profession. The final corpus and its distributions are presented in Section 2.3.

We conclude that during the entire process of data collection, it was indispensable to create goodwill and trust – this relates to the principals and teachers as well as to the students and their parents. We did not only try to do so before and during our school visits, but also afterwards. For instance, after having conducted some initial analyses, we reached out to the teachers and principals again (and also to some parents who had asked to be kept informed – see above) and communicated our findings to them. This resulted in a fruitful interaction with the secondary schools. In addition, through this intense bidirectional collaboration with the school communities, we were able, while still collecting the data, to fill 'gaps' when we noticed that particular subgroups were underrepresented, as the teachers allowed us to visit some additional class groups. Furthermore, we recruited the adolescents who participated in the anonymous survey (see Chapter 8) in the same secondary schools: 168 high school students with various socio-demographic profiles filled in an online survey that examined their linguistic attitudes and awareness of sociolinguistic patterns in informal online communication, as well as – to a minor extent – relevant language skills. No sensitive or

personal data were collected for this study, and participation was voluntary and completely anonymous, since the teenagers were never asked to fill in their name or class group. We refer to Chapter 8 for more information on the survey and its analysis.

We can conclude that citizen science, i.e. the active engagement of citizens in a scientific project, has been an indispensable part of the present research design and is a promising path for future sociolinguistic projects, especially with respect to the collection of reliable, representative and informative datasets.

2.2. Data preprocessing

SUBMISSION FORMAT

In the initial stages of the data collection, the secondary school students submitted their online conversations in various formats, ranging from screenshot images to text pasted in Word documents. Soon, we decided to optimize this procedure. From then on, we allowed two possible formats, depending on the social media platform on which the conversations were produced. For WhatsApp, the students could easily forward entire conversations via the app's 'export chat' setting. The conversations were automatically converted to plain text files and attached to an e-mail. While the plain text format kept all text, including special characters such as emoji, the students could opt to automatically delete all media files (e.g. pictures, video or audio files inserted in the chat conversations) from the email attachment. We note that all remaining media files (only applicable if the students did not select the delete-option) were automatically removed by us, since their analysis falls outside the scope of the project. For Facebook Messenger, the students were instructed to copy their conversations from the Facebook website and paste them to our submission website[4]. These pasted texts were immediately and automatically converted to a plain text format too, from which all media files were removed, but in which all text and special characters such as emoji were kept.

ANONYMIZATION

All relevant metadata were provided by the participants, who were asked to fill in their gender, year of birth, educational track, the province they lived in, their home language(s) and the profession of their parents. This meta-information was used to anonymize the corpus. In the original submissions, each message was linked to the author's name. These author names were deleted and replaced by unique and anonymous identifiers (e.g. author '502') that were still linked to the authors' socio-demographic profile. The link between the identifiers and the original names, however, was deleted from the corpus, and stored in a secure separate file (so we could still delete participants if they requested this, as they held a

[4] We thank Guy De Pauw and Ben Verhoeven for creating this website.

'right to be forgotten'[5] – see above). Consequently, each message in the anonymized corpus is associated with an author identifier, with a set of socio-demographic characteristics (e.g. the author's gender, educational track, age at the time of production of the message, etc.) and a set of additional properties (e.g. the platform on which and the year in which the message was written). We note that two properties were annotated and added to the corpus, but ultimately fell outside the scope of the research project: i.e. the specific conversational context in terms of the number of interlocutors and the interlocutors' gender. When annotating these variables, we made a distinction between dyadic (i.e. one-on-one) conversations and group chats (i.e. with more than two interlocutors), and between same-gender conversations (i.e. girls or boys only) and mixed-gender conversations (i.e. including at least one participant of both sexes). While the sociolinguistic analysis of these metadata falls outside the scope of the present research project, it is highly relevant, and is an interesting path for further research (see the concluding section, Chapter 9).

The anonymization of the data consisted of two steps. The deletion of the authors' names (described above) was the first step. The second step concerned the contents of the utterances: a script was written to automatically detect and replace personal (contact) information in the social media messages. (We recall that all media files, such as photos, videos or audio files, were already removed at this point.) Occurrences of first and last names, of towns or cities, street names, phone numbers, email addresses and specific urls were replaced by a 'placeholder' (see the examples (1) to (4) below). This procedure allowed us to anonymize the utterances without losing their general message or content. Below, we illustrate the results of the anonymization procedure with some authentic examples from the anonymized corpus.

(1) *Ge moogt woensdag om 7 uur naar de XSTRAATNAAMX komen, we gaan bij mij thuis hamburgers eten.*
'You can come to the XSTREETNAMEX on Wednesday at 7 o'clock, we are going to eat hamburgers at my place.'

(2) *ok merci maat hier is et e-mailadres : XEMAILADRESX*
'ok thanks buddy here is the email address: XEMAILADDRESSX'

(3) *Gaat gy graag vrijdag mee bbq'en in XGEMEENTEX?*
'Would you like to join us on Friday to barbecue in XTOWNX?'

(4) *XNAAMX komt denk ik wel*
'XNAMEX is coming, I think'

We detected persons' names and names of towns with predefined lists, such as name lists published by the Belgian government and lists of location names, to which all Flemish towns and cities were added manually. Phone numbers, street names, email addresses and urls were detected automatically with regular expressions, as they have a fixed and recognizable format.

[5] See General Data Protection Regulation (GDPR) article 17, 'Right to be Forgotten': https://eugdpr.org/the-regulation/

DELETION OF MULTIPLE OCCURRENCES OF IDENTICAL MESSAGES

Finally, we note that the corpus that is analyzed in Chapters 3 and 4 is slightly larger than the one examined in Chapters 5 to 8. The latter is a subset of the former, from which some multiple occurrences of identical messages or conversations (produced by the same participants) were deleted. These multiple occurrences are a consequence and complication of the data collection procedure. First of all, (parts of) conversations could accidentally be donated more than once by the same participants (e.g. we visited some class groups and students twice over the course of eighteen months). In addition, because the teenagers were free to make their own selection of (parts of) conversations, they sometimes selected certain parts more than once when scrolling through their chat history and copy pasting. As soon as we discovered these multiple occurrences, we wrote a script to clean up the corpus in an automated way. We note that a large amount of utterances *genuinely* occur many times – i.e. they are actually produced in an identical way by the same authors, but at different times or in different conversational settings. This is often the case for short, pragmatic utterances, such as *OK, ja* 'yes', *no* 'no', *haha*, or messages consisting of a single emoticon. Consequently, it was important to find the right balance between deleting too many (genuine) identical messages and too few (unintended) identical messages. Therefore, we added three conditions: first of all, we only treated messages as truly identical when they were produced in an identical way by the same participant and in the same conversational setting (condition 1). Furthermore, multiple occurrences were only deleted when the utterance contained more than 5 tokens and was part of a recurring 'block' of utterances, that was at least 3 posts long (conditions 2 and 3). The first condition was included to avoid the deletion of identical utterances that were produced by different authors, or by the same author but in a different conversational context. The second condition was meant to avoid the removal of genuine repetitions of short posts (see above), and the third one aimed at keeping the (genuine) repetition of the instant messaging version of 'chain' letters or emails (i.e. typically long chat utterances with an 'important' message, that are deliberately copied from one conversation and pasted to many others, in order to 'spread the word').

For the analyses that are conducted on the larger, unfiltered corpus, we recalculated all tendencies on the final corpus, and observed no changes in the detected patterns, nor in the effect sizes or levels of significance. We also note that, naturally, the number of participants was not reduced (as messages were only considered to be truly 'identical' when produced by the same participant), and that the multiple occurrences appeared to be equally frequent for all groups of youths (so the proportions and balances in the corpus remained the same before and after filtering).

Table 2 presents an overview of the different (versions of the) corpora and indicates in which chapters of the dissertation they are analyzed.

Corpus	Nr. of posts	Relevant available metadata	Analyzed in
Corpus 2007-2013	400 808	Age, gender, medium	Chapter 2
Corpus 2015-2016: unfiltered	488 014	Age, gender, educational track (For a subset of participants: Home language, parental profession)	Chapters 3-4
Corpus 2015-2016: filtered	434 537	Age, gender, educational track (For a subset of participants: Home language, parental profession)	Chapters 5-8

Table 2: Overview of the corpora and the chapters in which they are analyzed

2.3. Overview of the corpus

Table 3 presents an overview of the distributions in the final corpus with regards to the relevant socio-demographic variables, in terms of tokens[6], posts (i.e. instant messages), and participants or authors. We note that in order to protect the participants' privacy, and following the guidelines of our university's ethical committee, the collected dataset cannot be made publicly available.

[6] Tokens are visual units in a text, separated by whitespaces. In our corpus of social media posts, a token can be a word, but also e.g. an emoticon or an isolated punctuation mark.

Variable	Variable levels	Tokens	Posts	Participants
Educational track	General Secondary Education	739 831 (29%)	120 839 (28%)	596 (43%)
	Technical Secondary Education	1 151 684 (46%)	197 534 (45%)	393 (28%)
	Vocational Secondary Education	639 839 (25%)	116 164 (27%)	395 (29%)
	Unknown	0	0	0
Gender	Girls	1 696 517 (67%)	282 940 (65%)	717 (52%)
	Boys	834 837 (33%)	151 597 (35%)	667 (48%)
	Unknown	0	0	0
Age	Younger teenagers (13-16)	1 360 898 (54%)	244 807 (56%)	1 234[7]
	Older teenagers / young adults (17-20)	1 170 456 (46%)	189 730 (44%)	897
	Unknown	0	0	0
Home language	Dutch only	2 242 653 (89%)	380 064 (87%)	1154 (83%)
	Dutch + other language(s)	155 259 (6%)	24 782 (6%)	87 (6%)
	Other language(s) only	124 704 (5%)	27 720 (6%)	105 (8%)
	Unknown	8 738 (0.35%)	1 971 (0.45%)	38 (3%)
Parental Profession	'upper class'	358 698 (14%)	57565 (13%)	99 (7%)
	'middle class'	655 838 (26%)	114964 (26%)	214 (15%)
	'working class'	361 135 (14%)	59598 (14%)	87 (6%)
	Unknown or unclear	1 155 683 (46%)	202410 (47%)	984 (71%)
Total		2 531 354	434 537	1 384

Table 3: Distributions in the corpus

3. Linguistic and social variables

This section presents a concise overview of the linguistic and social variables included in the dissertation (Sections 3.1 and 3.2, respectively). We note that for all variables, in-depth discussions can be found in the chapters in which they are included.

3.1. Linguistic variables

The linguistic variables included in the dissertation concern different aspects of Flemish teenagers' informal online writing practices. Five main sets of features are introduced below. We briefly discuss how the relevant occurrences were automatically extracted from the

[7] The number of younger and older participants does not add up to the total number of participants, but to a higher number (which is why we did not add percentages for age). We recall that the same participants can occur in the corpus at different age points if they submitted recent chat conversations as well as older ones.

corpus, but for a more detailed explanation (as well as error analyses and evaluations of the software's performance), we refer to the chapters in which the features are addressed. At the end of this section, Table 4 summarizes all sets of linguistic features included in the dissertation as well as the chapters in which they are examined.

DEVIATIONS FROM THE FORMAL WRITING STANDARD

Most research papers included in the dissertation focus on a variety of deviations from the formal written standard present in teenagers' instant messages. Most of these deviations belong to one of three groups, corresponding to different 'maxims' or 'principles' of chatspeak, i.e. implicit rules of linguistic conduct for (informal) online interaction, that are distinguished by e.g. Androutsopoulos (2011, 149) and Thurlow and Poff (2013, 176): *brevity*, *orality* and *expressive compensation*.

DEVIATIONS FROM THE FORMAL WRITING STANDARD: EXPRESSIVE COMPENSATION

The principle of *expressive compensation* concerns the use of – predominantly typographic – markers and strategies in chatspeak that compensate for the lack of certain non-verbal expressive cues that are present in face-to-face communication (e.g. facial expressions, voice volume, intonation). A well-known example is the use of emoticons, which can, to a certain extent, represent facial expressions. All typographic expressive markers were automatically extracted from the corpus using regular expressions, as they are variations on fixed, typographic patterns.
These expressive features are the sole linguistic focus of Chapter 2, and are included among other types of features in all other chapters.

DEVIATIONS FROM THE FORMAL WRITING STANDARD: ORALITY

The *orality* principle concerns the inclusion of typical spoken language features in written online communication, in order to make the genre more speech-like. An example concerns the inclusion of non-standard (e.g. dialect or colloquial) renderings of Dutch lexemes in chat conversations, or the insertion of English words in Dutch conversations. These language- or register-specific oral features were detected in the data using predefined word lists or dictionaries. A pipeline approach was applied: first, each word's presence was checked in a standard Dutch word list (including named entities), and then in a standard English word list. If the word did not occur in any of these two dictionaries, it was classified as 'non-standard Dutch'[8]. A more detailed description and an evaluation of this dictionary-based approach can be found in the chapters in which oral features are examined, i.e. Chapters 3 to 6.

[8] We note that the vast majority of words in the corpus are Dutch, since entire conversations in other languages were excluded (see Section 2.1.2, 'submissions'). While codeswitching to English frequently occurs, words or phrases in other languages (e.g. Arabic) are much more rare. Consequently, the 'non-standard Dutch' category contains mostly words in non-standard (colloquial or regional) Dutch, as well as e.g. some misspellings.

DEVIATIONS FROM THE FORMAL WRITING STANDARD: BREVITY

The third and final maxim presented by e.g. Androutsopoulos (2011, 149) and Thurlow and Poff (2013, 176) is the *brevity* principle, which consists in maximizing typing speed and minimizing typing effort, so as to mimic the 'flow' and pace of an actual face-to-face conversation. An example is the use of (non-standard) abbreviations and acronyms. Just like the oral features, prototypical chatspeak abbreviations and acronyms were detected using handcrafted lists of common instances too.
These features are included in the research design in Chapters 3 to 6.

DEVIATIONS FROM THE FORMAL WRITING STANDARD: DISCOURSE MARKERS

A final set of features that diverge from the formal writing standard does not truly belong to any of the above mentioned categories but is nevertheless typical of informal online communication: i.e. the discourse markers hashtags ('#') and mentions ('@'). These can be used to indicate a topic or express a feeling about it (hashtags) or to address a specific person in a group chat (mentions). While these features emerged (and are especially relevant) on the microblogging platform Twitter, they have become popular on other social media platforms too (Zappavigna 2015, n.p.).
Just like the typographic expressive markers, these features were automatically extracted using regular expressions, as they are variations on fixed, typographic patterns.
These discourse markers are included in the research design in Chapters 4 to 6.

GENERAL TEXTUAL FEATURES: LEXICAL PATTERNS AND RELATED PARAMETERS

A final set of linguistic variables concerns more general (i.e. not specifically bound to digital media) textual features, with a focus on lexical patterns. An example of such features is lexical richness (i.e., a measure of the ratio of *different* words used in a text). For the analysis of some of these more general text features, the noisy social media texts needed to be normalized or standardized first (e.g. in order not to mistake orthographic variation for lexical variation, with respect to the variable of lexical richness): non-standard elements, i.e. deviations from formal writing norms, needed to be converted to their standard Dutch equivalent (e.g. non-standard spellings of standard Dutch words) or simply deleted (e.g. emoji). This normalization procedure and its accuracy are discussed elaborately in Chapter 7. Finally, the analysis of teenagers' top favorite – i.e. most frequently used – words (which were automatically counted and ranked) and their associated topics was done manually: we manually inspected the top-500 lexemes used by different subgroups of adolescents (e.g. boys compared to girls), to identify potentially relevant differences and similarities.
These more general textual features are the focus of Chapter 7 of the dissertation.

Linguistic variable	Examined in
Non-standard features: Expressive compensation	Chapters 2-7
Non-standard features: Orality	Chapters 3-6
Non-standard features: Brevity	Chapters 3-6
Non-standard features: Discourse markers	Chapters 4-6
General textual features	Chapter 7

Table 4: Overview of the linguistic variables and the chapters in which they are examined

3.2. Social variables

The social variables included in this dissertation all represent different aspects of teenagers' socio-demographic profile. As Table 3 shows (see Section 2.3), three variables are known for *all* participants: age, gender and educational track. For the vast majority of the participants, we also have information on their linguistic home context, as we know which language(s) they speak at home, and for a limited subset of participants, we have information on their parents' profession. The teenagers' social class, finally, is conceptualized as a combination of multiple socio-demographic characteristics (see below), and is available for a subset of teenagers. Below, we briefly describe each of these social variables. At the end of this section, Table 5 summarizes the social variables and the chapters in which they are included.

AGE

All participants in the corpus are secondary school students between thirteen and twenty years old. In the analyses, we systematically make a distinction between two groups of adolescents, i.e. younger teenagers, aged 13 to 16, and older teenagers or young adults, aged 17 to 20. The decision to treat age as a categorical (binary) variable rather than as a continuous one is based on theoretical grounds. In multiple sociolinguistic studies, it has been suggested that non-conformist behavior and, in (socio)linguistic terms, the use of 'non-standard' language by teenagers 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). Furthermore, a distinction between two similar age groups of younger and older youths is often made in related research (see e.g. De Decker 2014; Verheijen 2018).
Age is explicitly included as a social variable in Chapters 2, 5, 7, and 8. In all other chapters, additional tests are conducted to correct for its potential influence.

GENDER

Gender is treated as a binary variable too (i.e. as sex), since a non-binary approach, in which gender is operationalized as a continuum, was not feasible given the profile information we had access to. For alternative operationalizations, we refer to e.g. Bamman, Eisenstein and Schnoebelen (2014), who linguistically approach gender as consisting of multiple gender-

oriented (language) clusters, and Killermann (2014) who conceptualizes gender identity as a combination of values on four continuums, relating to identity, attraction, expression and sex. Just like age, gender is explicitly included as a social variable in Chapters 2, 5, 7, and 8, whereas in all other chapters, additional tests are conducted to correct for its potential influence.

EDUCATIONAL TRACK

All participants are students in one of the three main secondary educational tracks in Belgium: General, Technical and Vocational Secondary Education. These three tracks can be situated on a continuum, ranging from a very strong theoretical to a very strong practical orientation. General Secondary Education (in Dutch *Algemeen Secundair Onderwijs* or *ASO*) is the most theory-oriented track, in which students are prepared for higher (tertiary) education – which most students indeed attend after graduating from high school (i.e. approximately 96%, see Glorieux et al. 2014, 79). Vocational Secondary Education (in Dutch *Beroepssecundair Onderwijs* or *BSO*) is the most practice-oriented track, in which students are taught a specific (often manual) profession. Most of these students start their professional career after graduation. We note that the completion of this educational track does not offer (direct) access to higher education. Technical Secondary Education (in Dutch *Technisch Secundair Onderwijs* or *TSO*), finally, holds a middle position on the continuum from theory to practice, as it has a practical as well as a theoretical orientation. The main focus is on technical courses. After graduating, these students can either start their professional life or proceed to higher education. For more information on these three educational tracks and on additional tracks that fall outside the scope of the present research, we refer to the Flemish Ministry of Education and Training (2017). We note that the inclusion of this social variable in the research design is highly relevant, as adolescents' educational track strongly impacts their current and future (adult) social networks as well as their future professional career (de Jager, Mok, & Sipkema 2009; Glorieux et al. 2014) (see above).
Education is included in the research design of the studies presented in Chapters 3 to 8.

HOME LANGUAGE

For the vast majority of the participants, we have information on their linguistic background, i.e. we know which language(s) they speak at home. We make a distinction between three groups: teenagers who only speak Dutch at home (i.e. the official language in Flanders and the language of education), teenagers who speak both Dutch and one or multiple other languages at home, and finally teenagers who do not speak Dutch at home, but one or multiple other languages. Home language – which often, but not necessarily, coincides with native language – is an important aspect of youths' social background. It is a socio-cultural factor that may indicate a migration background and that can be highly relevant in educational settings, since, apart from the linguistic skills of the students themselves, it may for instance indicate the presence or absence of a parent who can easily connect with the (Dutch) school context and support children with school-related communication or tasks.
This variable is examined in Chapters 3 and 4.

PARENTAL PROFESSION

For a subset of 400 participants, we have detailed information on their parents' profession. As Table 3 shows, this information is 'unknown' or 'unclear' for the majority of the participants, for several reasons. First of all, many teenagers simply left this field blank when donating material, either because they did not know the answer or because they were reluctant to provide it. Furthermore, many teenagers filled in answers that were too vague to be used in the analyses (e.g. some students only provided the name of a company, without a job description and quite a lot of them used very vague labels, such as 'harbor'). Finally, a limited number of social positions (e.g. housewives/-men, retired or unemployed people) falls outside the scope of the sociological scheme that we applied in our classifications (see below).

We make a distinction between three groups: typical 'working class', 'middle class', and 'upper class' professions. This distinction and the classification of the specific professions was based on a widely applied sociological classification scheme in which professions are ranked in terms of autonomy, supervision, required level of education or skills, etc. (Erikson, Goldthorpe & Portocarero 1979, 420; see also Vranken, Van Hootegem, Henderickx & Vanmarcke 2017, 318).
Just like home language, parental profession is included as a social variable and discussed more elaborately in Chapters 3 and 4.

SOCIAL CLASS

For a subset of participants, i.e. all teenagers whose home language and parental profession is known (and clear), we operationalized social class. This complex phenomenon is conceptualized as a combination of home language, parental profession and educational track in Chapter 3, and as a combination of parental profession and educational track only in Chapter 4 (i.e. the improved operationalization). While we make a major distinction between prototypical 'upper class', 'middle class' and 'working class' teenagers, we also take into account adolescents with a more hybrid social profile (see Chapter 4).

Social variable	Examined in
Age	Chapters 2, 5, 7-8
Gender	Chapters 2, 5, 7-8
Educational track	Chapters 3-8
Home language	Chapters 3-4
Parental profession	Chapters 3-4
Social class	Chapters 3-4

Table 5: Overview of the social variables and the chapters in which they are examined

4. Statistical data processing

In the seven research papers included in this dissertation, a variety of statistical methods is used. In each chapter, the applied tests and techniques are discussed. Below, we give a brief overview. Table 6 at the end of this section summarizes the methods and the chapters in which they are applied.

CHI-SQUARE TESTS

In the first research paper of the dissertation (Chapter 2), chi-square tests are used in order to test the significance of attested sociolinguistic differences. As these tests are generally less robust in the sense that they are not normalized for differences in sample size (and the large size of the corpus might, to a certain extent, artificially boost statistical significance in this respect), we improve them by adding a bootstrap procedure[9]. This procedure is explained in Chapter 2. In addition, we correct for imbalances in the dataset with respect to the social variables by conducting subtests for distinct groups of participants, always keeping all social variables constant except for one (i.e. the variable at interest). In later chapters, we correct for potential gender and age imbalances by conducting *weighted* chi-square tests (Chapters 3 and 4) or by inspecting the impact of the different social variables simultaneously through (generalized) linear mixed model analyses (Chapters 5, 7 and 8).

(GENERALIZED) LINEAR MIXED MODELS

The analysis of adolescents' online writing practices through the use of (generalized) linear mixed models (GLMMs and LMMs, for categorical and continuous response variables, respectively) enabled the simultaneous inspection of the effect of multiple social variables (e.g. the teenagers' age, gender and educational track) on a linguistic response variable (e.g. the use of expressive chatspeak features), as well as the inclusion of potential interactions between these social predictors[10]. As mentioned above, the linguistic impact of different predictors need not necessarily be independent, but might be correlated, e.g. the effect of age might be different for girls and boys. The inclusion of a random effect for authors or participants takes into account the individual impact of the teenagers, which makes the analyses more robust and accurate. In addition, differences in sample size between the participants are accounted for too.

For more detailed information on these models, we refer to Chapters 5 (for generalized linear mixed models, with a categorical response variable) and 7 (for linear mixed models, with a continuous response variable). We note that these models are also used to a minor extent in Chapter 8, in order to reveal potential influences of teenagers' social profile on their survey responses.

[9] We thank Giovanni Cassani and Dominiek Sandra for their help and advice.
[10] We thank Ella Roelant and Erik Fransen from Statua, and Giovanni Cassani, Dominiek Sandra and Koen Plevoets for their help and advice.

MACHINE LEARNING ALGORITHMS

Chapter 6 has a stronger computational orientation, as it concerns the prediction of teenagers' educational track based on a sample of their online writing. This is a traditional profiling or classification problem, for which regression models did not yield good results. Therefore, well-known classification algorithms (i.e. machine learning algorithms) are used instead.

Statistical method/technique	Applied in
Chi-square tests	Chapters 2-4
Generalized linear mixed models	Chapters 5, (8)
Linear mixed models	Chapter 7
Machine learning algorithms	Chapter 6

Table 6: Overview of the statistical techniques and the chapters in which they are applied

5. Outline of the dissertation

This dissertation contains seven research papers each of which addresses different aspects of the main topic, i.e. the sociolinguistic variation in Flemish adolescents' informal online writing. This section presents an overview and summary of these seven papers or chapters.

Chapters of the dissertation:

CHAPTER 2: Expressive markers in online teenage talk: A correlational analysis
(Published in *Nederlandse Taalkunde*)

The research paper presented in Chapter 2 concerns the sociolinguistic variation in Flemish adolescents' production of a wide range of so-called *expressive* markers in informal online writing. These markers are predominantly (but not exclusively) typographic features that enhance or add the expression of emotional or social involvement in social media texts, and can – to a certain extent – compensate for the lack of non-verbal emotional cues present in face-to-face communication (see above). Correlations are examined between the use of these expressive markers and the teenagers' age and gender. In addition, the impact of the specific social media platform on which the texts were produced is analyzed too.

CHAPTER 3: Adolescents' social background and non-standard writing in online communication
(Published in *Dutch Journal of Applied Linguistics*)

Chapter 3 presents a pilot study on the linguistic impact of adolescents' social background on their informal online writing. Three aspects of teenagers' social background are included: their educational track, the profession of their parents and their home language(s). The

impact of the social variables is examined for a selection of three linguistic markers that each represent a different category of prototypical chatspeak markers.

CHAPTER 4: Social media writing and social class: A correlational analysis of adolescent CMC and social background
(Published in *International Journal of Society, Culture & Language*)

Chapter 4 is a follow-up analysis on the pilot study presented in Chapter 3. Again, correlations are examined between adolescents' social media writing and their social class background, but the operationalization of both the independent and dependent variables is adapted and improved, based on the findings presented in the pilot study. A much wider range of linguistic features is included, and the adolescents' social class is operationalized as a combination of educational track and parental profession only. In addition to 'prototypical' social groups (i.e. prototypical upper, middle, and working class youths), teenagers with a hybrid social class profile are now included in the linguistic analyses too. Furthermore, potential interactions between the teenagers' age, gender and social class are examined.

CHAPTER 5: Modeling adolescents' online writing practices: The sociolectometry of non-standard writing on social media
(Accepted with minor revisions in *Zeitschrift für Dialektologie und Linguistik* – revised version included in the dissertation)

Chapter 5 is a research paper on the statistical (descriptive) modeling of adolescents' online writing practices using generalized linear mixed models. Four models are presented that were trained to model or estimate the teenagers' use of four different sets or types of chatspeak features. We examine the impact of the adolescents' age, gender and educational track (incl. potential interactions) on their use of expressive chatspeak features (e.g. emoticons), 'oral' markers (e.g. the use of regional features) and brevity-related features (e.g. chatspeak acronyms). In addition, we make a systematic distinction between teenagers' use of new and old vernacular, i.e. 'digital' versus 'traditional' types of non-standardness. It is examined whether distinct sociolinguistic variation patterns emerge for the different feature sets and/or for the different vernaculars.

CHAPTER 6: Predicting adolescents' educational track from chat messages on Dutch social media
(Published in *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*)

Chapter 6 presents a pilot study on education profiling, i.e. the prediction of teenagers' educational track based on a sample of their social media texts and its linguistic properties. The models discussed in this paper can thus be seen as the 'reverse' of the models presented in Chapter 5. Rather than (generalized) linear regression models, typical (machine learning) classification algorithms are used (e.g. Naive Bayes).

CHAPTER 7: Lexical patterns in adolescents' online writing: The impact of age, gender and education
(Manuscript submitted)

Chapter 7 complements the findings of the previous chapters on teenagers' use of prototypical chatspeak features and deviations from the formal writing standard with an analysis of more general (i.e. not specifically bound to new media) linguistic properties of their online writing. It mainly focuses on lexical patterns (e.g. lexical richness). Consequently, aspects of the adolescents' 'traditional literacy' are analyzed in the informal setting of social media and compared to their exploitation of 'digital literacy'.

CHAPTER 8: Adolescents' perceptions of social media writing: Has non-standard become the new standard?
(Manuscript submitted)

Chapter 8 complements the findings of previous chapters on teenagers' *production* of informal online communication by examining their *perception* of or attitudes on the genre. We report the findings from a survey conducted among Flemish secondary school students. The survey was designed in order to investigate the participants' awareness of previously attested sociolinguistic patterns in online writing, as well as their attitudes with respect to standard language use in different settings and their appreciation of specific chat utterances or features. To a minor extent, it focused on youths' formal spelling skills and examined their register sensitivity. The teenagers' replies to the survey are systematically compared to the findings on youths' actual online writing practices, in order to lay bare similarities or discrepancies between the perception and production of the genre.

CHAPTER 9: Conclusion

Finally, in Chapter 9, the findings presented in the different chapters are summarized and evaluated. We also discuss the broader relevance of the dissertation and present some paths for further research.

References

Androutsopoulos, Jannis. (2011). Language change and digital media: A review of conceptions and evidence. In Tore Kristiansen, & Nikolas Coupland (Eds), *Standard languages and language standards in a changing Europe* (pp. 145-161), Oslo: Novus.

Bamman, David, Jacob Eisenstein, & Tyler Schnoebelen. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18(2), 135-160.

Baron, Naomi S. (2004). See you online: Gender issues in college student use of instant messaging. *Journal of Language and Social Psychology* 23(4), 397-423.

Baron, Naomi S. (2008). *Always on: Language in an online and mobile world.* Oxford: Oxford University Press.

Coates, Jennifer. (1993). *Women, men and language. A sociolinguistic account of sex differences in language*. London: Longman.

De Decker, Benny. (2014). *De chattaal van Vlaamse tieners. Een taalgeografische analyse van Vlaamse (sub)standaardiseringsprocessen tegen de achtergrond van de internationale chatcultuur.* Antwerp: University of Antwerp (doctoral thesis).

De Decker, Benny, & Reinhild Vandekerckhove. (2017). Global features of online communication in local Flemish: Social and medium-related determinants. *Folia Linguistica* 51, 253-281.

de Jager, Hugo, Albert Louis Mok, & G. Sipkema. (2009). *Grondbeginselen der sociologie*. Groningen / Houten: Noordhoff.

Eckert, Penelope. (2000). *Linguistic variation as social practice.* Malden / Oxford: Blackwell.

Eisikovits, Edina. (2006). Girl-talk/boy-talk: Sex differences in adolescent speech. In Jennifer Coates (Ed.), *Language and gender: A reader* (pp. 42-54), Oxford: Blackwell.

Erikson, Robert, John H. Goldthorpe, & Lucienne Portocarero. (1979). Intergenerational class mobility in three Western European societies: England, France and Sweden. *The British Journal of Sociology* 30(4), 415-441.

Estival, Dominique, Tanja Gaustad, Son Bao Pham, Will Radford, & Ben Hutchinson. (2007). Author profiling for English emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics* (pp. 263-272).

Flemish Ministry of Education and Training. (2017). *Statistisch jaarboek van het Vlaams onderwijs. Schooljaar 2015-2016.* Brussels: Department of Education and Training.

Glorieux, Ignace, Ilse Laurijssen, & Olaf Sobczyk. (2014). De instroom in het hoger onderwijs van Vlaanderen: Een beschrijving van de huidige instroompopulatie en een analyse van de overgang van secundair onderwijs naar hoger onderwijs. Research paper SSL/2013.16/4.1.2, Leuven: Steunpunt SSL.

Grondelaers, Stefan, Paul van Gent, & Roeland van Hout. (2016). Destandardization is not destandardization. Revising standardness criteria in order to revisit standard language typologies in the low countries. *Taal en Tongval* 68(2), 119-149.

Herring, Susan C., & Anna Martinson (2004). Assessing gender authenticity in computer-mediated language use: Evidence from an identity game. *Journal of Language and Social Psychology* 23, 424-446.

Holmes, Janet. (1992). *An introduction to sociolinguistics*. London / New York: Longman.

Killermann, Sam. 2014. Breaking through the binary: Gender as a continuum. *Issues* 107, 9-12.

Kucukyilmaz, Tayfun, B. Barla Cambazogly, Cevdet Aykanat & Fazli Can. (2006). Chat mining for gender prediction. In: *International Conference on Advances in Information Systems* (pp. 274-283), Berlin: Springer.

Labov, William. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Marsh, Ian (Ed.). (2000). *Sociology. Making sense of society.* Harlow: Prentice Hall.

Nguyen, Dong, A. Seza Doğruöz, Carolyn P. Rosé, & Franciska de Jong. (2016). Computational sociolinguistics: A survey. *Computational Linguistics* 42(3), 537-593.

Parkins, Róisín (2012). Gender and emotional expressiveness: An analysis of prosodic features in emotional expression. *Griffith Working Papers in Pragmatics and Intercultural Communication* 5(1), 46-54.

Reddy, T. Taghunadha, B. Vishnu Vardhan, & P. Vijayapal Reddy. (2016). A survey on authorship profiling techniques. *International Journal of Applied Engineering Research* 11(5), 3092-3102.

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. *PLoS ONE* 8(9).

Tagliamonte, Sali. (2016). *Teen talk: The language of adolescents.* Cambridge: Cambridge University Press.

Tagliamonte, Sali A., & Derek Denis. (2008). Linguistic ruin? Lol! Instant messaging and teen language. *American Speech* 83(1), 3-34.

Thurlow, Crispin, & Michele Poff. (2013). Text messaging. In Susan Herring, Dieter Stein, & Tuija Virtanen (Eds), *Pragmatics of computer-mediated communication* (pp. 163-190), Berlin / New York: Mouton de Gruyter.

Trudgill, Peter. (1983). *On dialect. Social and geographical perspectives.* Oxford: Blackwell.

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

Verheijen, Lieke. (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).

Verheijen, Lieke. (2016). De macht van nieuwe media: Hoe Nederlandse jongeren communiceren in sms'jes, chats en tweets. In Dorien Van De Mieroop, Lieven Buysse, Roel Coesemans, & Paul Gillaerts (Eds), *De macht van de taal: Taalbeheersingsonderzoek in Nederland en Vlaanderen* (pp. 275-293), Leuven / The Hague: Acco.

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

Vranken, Jan, Geert Van Hootegem, Erik Henderickx, & Luc Vanmarcke. (2017). *Het speelveld, de spelregels en de spelers? Handboek sociologie.* Leuven / The Hague: Acco.

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

CHAPTER 2

This chapter was published as a journal article. Reference:

Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans. (2018). Expressive markers in online teenage talk: A correlational analysis. *Nederlandse Taalkunde* 23(3), 293-323.

Expressive markers in online teenage talk:

A correlational analysis

Abstract
This paper discusses the expression of emotional involvement in informal computer-mediated communication (CMC). While related research is quite fragmentary through its exclusive focus on a limited number of expressive markers or the inclusion of just one independent variable, the present study includes a wide range of expressive markers and three independent variables. The data reveal strikingly consistent age and gender correlates across all expressive markers and a strong correlation between the preferences of younger adolescents and girls. Furthermore, the study highlights a major impact of medium type. It calls for a refinement of the operationalization of the variable medium, as apart from its inherent characteristics (private/public, synchronous/asynchronous), the nature and goal of the interaction (which is also partly related to the type of social media that people use) trigger specific linguistic practices.

Keywords: CMC, youth language, expressive markers, sociolinguistics, gender, age, medium

1. Introduction

Since the rise of informal computer-mediated communication (CMC), both laymen and linguists have been fascinated by the prototypical features of several forms of digital writing (see Crystal 2001). Androutsopoulos (2011, 149) relates these features to three dimensions (also called maxims or principles): orality, compensation, and economy. While orality refers to the use of spoken language features in written discourse[1] and economy covers all strategies to shorten messages, the "semiotics of compensation" "includes any attempt to compensate for the absence of facial expressions or intonation patterns" (Baron 1984, 125 as cited in Androutsopoulos 2011, 149). De Decker and Vandekerckhove (2017, 278) stress the importance of making a distinction between economical and expressive chatspeak features in CMC research, as both groups of features appear to correlate differently with the variables of age, gender and medium. While they found age and medium correlates for several chatspeak features, they did not identify significant gender patterns, except for the only expressive variable that was part of their analyses. Consequently, the authors concluded that "[their] findings call for further refinement of the operationalization of emotional

[1] For a more elaborate and nuanced view on the dichotomy between written and spoken language, we refer to Koch & Oesterreicher (2001, 584-585; 2011, 3-4), who take both *Medium* ('realization': either phonic or graphic) and *Konzeption* ('register': spoken/informal register or written/formal language) into account to create four combinations on what they call the *continuum* between spoken and written language. In the case of informal CMC, the medium may be a written medium, but the discourse is often to a large extent conceptually oral (see also Schlobinski 2005).

expressiveness in CMC and a broader selection of expressive markers". The present paper, which focuses exclusively on that type of markers in Flemish online teenage talk, meets these requirements. It does not only include typographic features that are prototypically associated with the maxim of compensation, such as emoticons or the capitalization of words and utterances, but also a lexical and an onomatopoeic variable – namely the use of intensifiers and the onomatopoeic rendering of laughter (e.g. *haha*). The notion of expressiveness is thus used as a cover term for the expression of (strong) involvement, in most cases emotional involvement. The following example contains four of the eight features that function as the dependent variables in the present study (i.e. the onomatopoeic *hahaha*, the capitalization of *super*, repetition of the exclamation mark and the emoticon *:D*):

(1) *Hahaha SUPER!!! :D*

Our main research question relates to the potential correlation between the use of the selected expressive markers and the sociolinguistic profile of the chatters. All informants are adolescents from Dutch-speaking northern Belgium, i.e. Flanders. The social variables operationalized in the present study are their age and gender. The main goal is to identify the most expressive subgroup: do women and younger adolescents outperform men and older adolescents respectively in the use of expressive markers, or do these groups show distinct preferences for specific expressive markers? These research questions are inspired by the related research that will be discussed in Section 3. Apart from age and gender, Section 3 also discusses the potential impact of different digital media. Since our data contain both largely public asynchronous and private synchronous online messages, this variable had to be included in the research design. Moreover, the combination of these three variables distinguishes the present study from much of the related research. Before discussing potential determining factors, we will present the expressive markers themselves and previous literature on each of them (Section 2). Section 4 is devoted to the experimental setup: it describes the corpus, the participants and the methodology of the data extraction and processing. The following section (5) contains the results of the analyses and the final one (6) presents the conclusion.

2. The expressive markers

There are many ways of expressing emotional involvement, both in speech and in written language. The most obvious way of doing so is by literally articulating emotions, e.g.: 'I feel sad'. In many cases, however, feelings and emotions are expressed in a more indirect way, for example through particular facial expressions. The absence of such facial expressions, but also of other forms of body language (e.g. hand gestures), of voice volume and pitch in textual computer-mediated communication leads to the compensatory strategies which we referred to above (see also Thurlow & Poff 2013, 176, who use the term *paralinguistic restitution*, and Kucukyilmaz, Cambazogly, Aykanat & Can 2006, 276). These compensatory typographic

features represent the majority of the expressive markers that we selected for the present study. We refer to them as the *expressive chatspeak features* and discuss them in Section 2.1. Section 2.2 concerns the onomatopoeic rendering of laughter, which is not a typographic feature but can be considered typical of chatspeak too. Section 2.3, finally, presents a lexical feature which is not typical of CMC, but which certainly can be considered a marker of expressiveness that functions in much the same way as some of the typographic markers, i.e. the use of intensifiers.

2.1. Typographic expressive markers: Expressive chatspeak features

Androutsopoulos (2011, 149) distinguishes several compensational features: "emoticons, abbreviations that signify various types of laughter, simulation of expressive prosody by iteration of letters and punctuation". All of these typographic markers are included in the present study, but we added two more: capitalization of entire words or utterances and the use of the letter(s) *x* or *xo*[2] (or several instances of both) to symbolize kisses versus hugs and kisses respectively. In the next paragraphs, we will briefly discuss each of these features.

The first marker in the present research design is so-called *flooding*:[3] the deliberate repetition of letters or punctuation marks (both are present in example (2)).

(2) *ik ben s**uuuuu**per hyper**!!!!***
'I am s**uuuuu**per hyper[active]**!!!!**'

Flooding can be interpreted as a way of symbolically emphasizing a word (letter flooding) (De Decker & Vandekerckhove 2017, 265) or an entire utterance (punctuation flooding). Parkins (2012, 52) states that letter flooding serves both expressiveness and creativity: "The manipulation of letters, such as the repetition of a certain vowel or consonant, can be used creatively in many situations to represent emotional stances such as pondering, disappointment, doubt, frustration, sarcasm, and happiness". As for punctuation flooding in particular, she adds that it is used "to indicate a degree of intensity in what the author had to say" (2012, 50), rather than for grammatical purposes, as is the case for standard punctuation. Unlike De Decker and Vandekerckhove (2017, 265), we make a distinction between the repetition of letters and the repetition of punctuation marks. For letter flooding, we worked with a threshold of three or more[4] identical graphemes. Repetitions of the letter *x* were

[2] We note that single occurrences of *xo* are rather ambiguous, as they could be used (and perceived) as both a kiss and a hug, or as an emoticon representing a facial expression with an open mouth. We opted for the first interpretation, but that might be the wrong choice in some cases. However, as these occurrences are extremely rare in the corpus (only 0.4% of all kisses, and less than 0.01% of all tokens) and their impact consequently is negligible, we did not exclude them from the analyses.
[3] Different terms are used to indicate the phenomenon of flooding, like *reduplication* (Verheijen 2015, 132), *additional letters* (Parkins 2012, 52) or *letter repetition* (Darics 2013).
[4] We note that there is no 'rule' that decides which number of repetitions is needed for character repetition to be counted as flooding, nor is there the certainty that some occurrences of flooding were not just typed by mistake.

excluded, as they are generally used to render 'kisses' and thus serve a different function (see below). For punctuation flooding, we used a threshold of two or more repetitions and restricted the selection to question and exclamation marks.

Apart from punctuation flooding, combinations of question and exclamation marks (example (3)) are also included as a distinct variable:

(3) *wat**?!?***
'what**?!?**'

Another way to express emotion or involvement in written CMC is the use of unconventional capitalization. The most common and probably most expressive application consists in writing entire words or utterances in capital letters (also called *allcaps*), which seems to be a visual, typographic representation of shouting. The following extract from a conversation between two Flemish chatters corroborates this interpretation:

(4) chatter A: *NIE ZO RAP KAN NI VOLGEN*
'not so fast, I can't keep up'
chatter B: *nie schreeuwe*
'don't shout'

Just like shouting in a face-to-face conversation, capitalizing entire words in an online conversation often is intuitively perceived as an expression of anger. However, it can just as well express other emotions, such as excitement and happiness (Parkins 2012, 51):

(5) *ik zal morgen **ALLES** vertellen*
'tomorrow, I will tell **EVERYTHING**'
(6) ***IT WAS SO GOOD THOUGH**! I'll have to show you so you can buy it :P* (Parkins 2012, 51, emphasis added)

Finally, it can be used as a more neutral emphasizer, which draws attention to parts of the utterance:

(7) *wie gaat er nu **ZEKER** mee?*
'who is coming along **FOR SURE**?'

These cases of allcaps were included in the present study, but other unconventional ways of capitalization, like alternating upper and lower case letters (e.g. *hElLo* instead of *hello*, see Herring 2012, 2) were not, because they seem to have a primarily fun-oriented and creative function, rather than a strictly (emotionally) expressive one. For the detection of allcaps, we only selected words that contain more than one letter, in order to reduce noise.

Furthermore, emoticons (short for "emotional icons", Wolf 2000, 828) or smileys are quite explicit expressive markers, as many of them literally are (typo-)graphic representations of facial expressions. Emoticons are very popular in CMC (Wolf 2000, 828). Parkins (2012, 52) even states that "[they] are the most frequently used prosodic features to express emotion online". Originally, typographic characters (mainly punctuation marks) were combined to

create a stylized image of a human face. Among these original smileys, both Western and Asian (also called Japanese) variants can be distinguished. The main difference is that the Western ones (examples (8) and (9)) are rotated – one must tilt one's head to the left or sometimes to the right to read them (Wolf 2000, 828) –, whereas the Asian ones, called 'kaomoji', are not (examples (10) and (11)).

(8) :) (smiling face)
:-O (surprised face)
;-) (winking face)
XD (face laughing, eyes closed)
(9) *Sgoe* ***:)*** 'alright **:)**'
(10) ^^ (closed, smiling eyes)
T__T (crying face, tears streaming from eyes)
o_O (surprised face, confused)
-__- (unamused face, frustrated)
(11) *kvin een papier nimeer* **-_-**
'I can't find a sheet of paper **-_-**'

Punctuation marks, letters, numbers and other symbols can also be combined to create images other than human faces, like (rotated) hearts:

(12) *ik mis em ook* ***<3***
'I miss him too **<3**'

These manually composed smileys are the oldest, i.e. first-generation emoticons. More recent than these traditional smileys are all kinds of Unicode or ASCII encodings which today are called emoji. Instead of actually composing the desired emoticon, the chatter simply selects emoji (as proper images) from a list. In the present paper, we use the term emoticon as a cover term for both classic emoticons and expressive emoji. Some examples of the latter can be found below: (13) contains one that was bound to the data produced on the synchronous chat platform, whereas (14) contains one typical of the asynchronous data produced on a social media site (see Section 3.3).

(13) *kheb toch gratis smse* 
'I can send free text messages anyway' 
(14) *Mrciii* 
'Thank you  '

Emoticons can express a whole range of feelings. Wolf (2000, 830) distinguishes the following categories: "teasing/sarcasm,[5] humor, sadness, despair, confusion, to offer an apology, a

[5] Wolf (2000, 832) points out that "whether [sarcasm and teasing] constitute an emotion is debatable". We will not focus on that debate here, as the expression of sarcasm or teasing increases the overall expressiveness in CMC just as well as the expression of 'unambiguous' emotions does.

positive feeling or thanks, or to express solidarity/support". She adds a separate category for emoticons with an unclear or no apparent purpose.

Finally, we added a typographic feature which dates back to pre-digital times, but which, judging from the Flemish chat conversations, seems to enjoy a renewed and intense popularity nowadays: the use of one (or more) instances of the letter *x* (sometimes capitalized) to symbolize a kiss (or several kisses). Many adolescents do not only use this symbol at the end of their conversations, by way of greeting, but insert the *x*'s in their discourse continuously or quite frequently. For the sake of completeness, we also included the sequence *xoxo* (and variants: *xoxoxo,* …), which stands for 'hugs and kisses'. Examples are shown below.

(15) *hey snelle cv metj* ***xx***
'hey handsome, everything alright **xx**'
(16) *hey !!!* ***xoxo***

Summing up, these are the six typographic expressive markers that function as variables in the present study: (1) flooding of letters, (2) flooding of punctuation, (3) combinations of exclamation mark and question mark, (4) capitalization of words or entire utterances, (5) emoticons, (6) rendering of kisses or hugs and kisses.

2.2. Onomatopoeic expressive marker: Onomatopoeic rendering of laughter

An alternative for one of the most common emoticons, i.e. the smiling face, are the onomatopoeic utterances *haha* and *hihi* (and variants: *hahaha*, *whaha, hihihihi,* …). These utterances may not be prototypical chatspeak features, but for two reasons we decided to include them: first of all, they seem to be the equivalent of smileys that express laughter (see example (17)). Secondly, they are fairly frequent in the Flemish corpus. Therefore, it seemed somewhat incongruent to include laughing smileys but exclude their onomatopoeic equivalents, so we chose to include both.

(17) ***Haha*** *Grappig profiel*
'**Haha** funny profile'

2.3. Lexical expressive marker: Intensifiers

The concept of intensifiers is quite ambiguous. Symptomatic in this respect is the fact that there is no real consensus among linguists concerning the appropriate terminology. Some of the names and terms used in previous research are *intensives* (Stoffel 1901), *amplifiers* (Quirk, Greenbaum, Leech, Svartvik & Crystal 1985, 590), *maximizers* and *boosters* (Quirk et al. 1985, 591). We adopt both the terminology and the definition used by Stenström, Andersen & Hasund (2002, 139), and see intensifiers as "items that amplify and emphasize the meaning of an adjective or adverb". This definition captures both their function and their grammatical

'compatibility'. In Dutch, intensifiers can either be adverbs, as illustrated in example (18), or intensifying prefixes, as shown in example (19).

(18) *Auwtch daswel **heel** vroeg*
'Ouch that is **very** early'
(19) ***kei**schattig!!*
'**very** cute!!'

Intensifiers are not typical of computer-mediated communication. However, they can be considered markers of expressiveness and they often function in much the same way as the other expressive features. According to Peters (1994, 271), people mainly use intensifiers to captivate the interlocutor or reader by displaying linguistic creativity, and to express emotional involvement. Both functions apply to most of the other expressive features as well. Compare, for instance, an utterance like *you are BEAUTIFUL* with *you are so beautiful*. In the former utterance, the speaker stresses his involvement through the capitalization of the adjective, in the latter through the insertion of the intensifier *so*. By using an intensifier, the speaker shows that his enthusiasm, disappointment, happiness, appreciation, etc. is not just moderate or mediocre, but intense. Typographic features like flooding and capitalization generally have the same effect.

Since we are focusing on the correlation between the frequency of intensifier use and authors' age, gender and medium, we will not be dealing with the actual appearance of the many variants, but we note that they are fascinating objects of linguistic study for several reasons, one of them being that they are very dynamic and marked by constant renewal and change (Quirk et al. 1985, 590; Pyles & Algeo 1993, 250; Peters 1994, 271; Méndez-Naya 2003, 372; Tagliamonte 2008, 391 and references therein). Moreover, they are often subject to delexicalization or grammaticalization, i.e. the process in which a word gradually loses lexical content but gains grammatical functionality (Partington 1993, 183; Lorenz 2002, 144).

3. The independent variables: Gender, age and medium

In this section, we discuss the results of previous research on the linguistic impact of gender (3.1), age (3.2) and medium (3.3). We will focus on expressiveness and include both sociolinguistic and stylometric[6] research. Following the discussion of the related research, we will present our hypotheses (3.4).

[6] Stylometry is a subdiscipline of computational linguistics: "The basic research question for computational stylometry seems then to describe and *explain* the causal relations between psychological and sociological properties of authors on the one hand, and their writing style on the other" (Daelemans 2013, 1, emphasis in original).

3.1. Gender

Sociolinguistic and stylometric research reveal parallel tendencies with respect to patterns in male and female language[7] related to expressiveness. Female discourse is said to be more expressive and emotional, in offline (i.e. face-to-face) as well as in online communication (Jespersen 1922, 251; Wolf 2000, 831; Kucukyilmaz et al. 2006, 282; Parkins 2012, 48, 50, 53). These findings contradict hypotheses about "online gender swapping", i.e. women and men adopting different roles in online communication than in face-to-face interaction and thus possibly communicating in new, non-stereotypical ways (Wolf 2000, 827). While women are found to use more emotional language or language expressing social involvement – talking and writing more about personal, social and emotional processes like feelings and thoughts – , men appear to use more informative language – focusing more on specific facts, objects and events (Jespersen 1922, 251; Argamon, Koppel, Fine & Shimoni 2003, 323, 334; Baron 2008, 51; Newman, Groom, Handelman & Pennebaker 2008, 223, 229, 232-233; Argamon, Koppel, Pennebaker & Schler 2009; Schwartz et al. 2013, 9).

With respect to the expressive markers that are subject of the present study, women (or girls) have been found to use significantly more intensifiers than men (or boys) (Stenström et al. 2002, 142 and references therein). Apart from this quantitative discrepancy, a qualitative difference has been found as well, with men and women preferring different intensifiers (Tagliamonte & Roberts 2005, 289; Xiao & Tao 2007, 251; Tagliamonte 2008, 388). While teenage girls may be more expressive quantitatively, from a qualitative perspective, the teenage boys seem to outperform the girls as they opt more often for strong intensifiers (e.g. *extremely*) and taboo words (e.g. *fucking*) (Stenström et al. 2002, 139, 143).

Furthermore, CMC research generally reveals a higher frequency of emoticons in female utterances (Baron 2004, 415; Herring & Martinson 2004, 436; Kucukyilmaz et al. 2006, 282; Parkins 2012, 52; Schwartz et al. 2013, 8). Moreover, Wolf (2000, 833) points to a functional expansion of smileys in female discourse: "Females have expanded on the male definition of emoticons and their use, adding other dimensions including solidarity, support, assertion of positive feelings, and thanks". Huffaker and Calvert (2005), on the one hand, and Wolf (2000), on the other, however, challenge and nuance the findings concerning the gender-dependent rate of emoticon use. Huffaker and Calvert (2005, n.p.) report the opposite effect among adolescent chatters, i.e. boys using more emoticons than girls. Wolf's nuance concerns the interlocutors: she found that in mixed-gender conversations, "both males and females display an increase in emoticon use", resulting in an insignificant gender difference (2000, 831-832). Moreover, her findings also reveal convergence with respect to the communicative function of the emoticons. According to Wolf, women mostly use smileys for humorous purposes, while men deploy them more for teasing or expressing sarcasm. In mixed-sex conversations, this difference is levelled out to some extent (Wolf 2000, 832). However, while the corpus for

[7] Gender is generally reduced to a binary variable (male vs. female). For criticism of this approach and for alternative views, see Bing and Bergvall (1996) and Coates (1993).

the present case study contains both mixed-sex and single-sex conversations, this variable was not included in the research design.

Finally, Parkins (2012, 48, 50-53) reports a higher frequency in online female communication for several of the expressive markers that are subject of the present study: letter and punctuation flooding, capitalized text, emoticons and expressions of laughter. Varnhagen et al. (2010, 729) and Baron (2004) also report a higher frequency of typical chatspeak features and markers of emotional involvement in girls' CMC.

3.2. Age

As for the linguistic impact of age and adolescence, it is widely accepted that creativity, language innovation and non-standard language use peak during puberty (Eckert 1997, 163; Androutsopoulos 2005, 1499; De Decker 2014, 44; Peersman, Daelemans, R. Vandekerckhove, B. Vandekerckhove & Van Vaerenbergh 2016, 16-17). However, adolescence is no homogeneous linguistic period, since the so-called 'adolescence peak' tends to be situated at the ages of 15 and 16. The use of non-standard language is supposed to culminate at that age and to decrease as youngsters age (Wolfram & Fasold 1974 as mentioned in Eisikovits 2006, 42; Holmes 1992, 184; Coates 1993, 94; De Decker & Vandekerckhove 2017, 277). As for CMC specifically, younger teenagers are said to use more typical chatspeak features in their online messages than older adolescents (Tagliamonte & Denis 2008, 13). A possible explanation could lie in changing attitudes concerning deviations from the linguistic standard: whereas adolescents seem to consider them as cool and use them for 'belonging' as well as for identity construction (Verheijen 2015, 129; De Decker & Vandekerckhove 2017, 278), young adults might see these deviations as "somewhat childish" (Verheijen 2015, 135).

In general, younger people's and particularly teenagers' language use is considered to be more expressive and emotionally loaded than that of the older generations: many of the (stylistic) innovations typical of adolescent talk are hypothesized to "primarily serve expressive and interactive purposes" (Androutsopoulos 2005, 1499). Pennebaker (2011, 61-63) adds that younger people use more negative and fewer positive emotion words than older people. On a content-based level, teenagers often talk and write about how they feel (Argamon et al. 2009, n.p.). Quite surprisingly, however, adolescent speech is generally found to contain fewer intensifiers than adult language (Paradis 2000, 154; Stenström et al. 2002, 141; Pertejo & Palacios Martínez 2014, 218). Stenström et al. even report that in their corpus, "the adults use intensifiers almost twice as frequently as the teenagers" (2002, 141). Paradis (2000, 154) ascribes this quantitative difference to a different choice of intensifying strategies. Yet intensifiers often function as a *groups binder* in adolescent peer groups: the use of a specific (set of) variant(s) can serve not only speaker but also *group* identification and signal in-group membership, at least until the variant becomes more widely popular and gets picked up by other groups (Peters 1994, 271; Lorenz 1999, 24-25). Furthermore, research

indicates qualitative differences between adolescents' and adults' use of intensifiers (Tagliamonte 2008, 388), with the former showing a greater preference for new, informal, regional and non-standard variants (Eckert 2003, 116; Androutsopoulos 2005, 1497).

CMC research suggests that teenagers generally use more stylistic (chatspeak) features than older chatters (Argamon et al. 2009, n.p.; Goswami, Sarkar & Rustagi 2009, 215; Schwartz et al. 2013, 9). This also holds for some of the expressive markers included in the present study: they appear to be more frequent in teenagers' CMC than in older people's chat messages. Youngsters have been found to use more emoticons than adults (Argamon et al. 2009, n.p.; Schwartz et al. 2013, 9), while young adolescents apply more flooding than adolescents at the end of their teens (De Decker & Vandekerckhove 2017, 265).

Verheijen (2015; 2016) distinguishes two age groups: younger adolescents versus older adolescents or young adults. She reports that in instant messages, emoticons and unconventional spelling forms were used much more often by teenagers than by young adults (Verheijen 2015, 135-136; Verheijen 2016, 283, 285). Strikingly, the opposite effect was noted for emoticons in (telephone) text messages: young adults used more emoticons than adolescents (Verheijen 2016, 285).

3.3. Medium

The final independent variable relates to the medium on which the online communication took place. We distinguish four main types of CMC based on (the possible combinations of) two parameters: synchronicity of the medium and number of interlocutors[8] (see Table 1 for an overview). *Synchronous* CMC (instant messaging) consists of real-time chat sessions in which all interlocutors are online at the same time (Baron 2004, 298). In *asynchronous* CMC (or non-instant messaging), only the emitter is online and not the receiver, or at least not necessarily so (Herring 2001). Both types can contain one-to-one just as well as one-to-many messages.

	One-to-one	**One-to-many**
Synchronous	Instant messaging with two interlocutors	Instant messaging with multiple interlocutors: group chats
Asynchronous	Email, private messages, texts, …	Public posts or reactions on social media or online fora

Table 1: Different types of CMC (De Decker 2014, 3)

[8] We note that other typologies are possible too, as the two selected parameters are not the only ones, nor are they necessarily the most influential ones for all phenomena or markers: e.g. the type of keyboard or electronic device – computer or mobile device such as smartphone or tablet – can have a large influence as well. However, our choice is determined by practical constraints, as these are the only parameters we have information on.

Both the synchronicity of the medium and the public versus private nature of the communication can impact on language use. Different hypotheses can be found in related research, relying on different views on digital media platforms and different theories about the ease or automaticity with which people use standard language.

As for the impact of the synchronicity of the electronic medium, some linguists argue that people write in a more standard-oriented way on asynchronous platforms, as they experience less time pressure[9] than in synchronous communication and therefore have more time to check and edit their posts (Herring 2001, 617; Gheuens 2010, 17-18; Verheijen 2015, 134). Others, however, hypothesize that chatters might use the extra time in asynchronous posts for experimenting and linguistic innovation (De Decker 2014, 64; De Decker & Vandekerckhove 2017, 256).

As for the public versus private dimension, Verheijen (2015, 134) notes that the public (one-to-many) character of some asynchronous channels could encourage people to turn to more standard orthography, to avoid "being chided for their spelling". But De Decker and Vandekerckhove (2017) add that even though private conversations with close peers can be more comfortable, "this need not imply that private interaction favors experimenting more than public interaction, since self-presentation on public networking sites might also be a trigger for creative language use" (256) and the use of chatspeak features "might raise [youngsters'] personal attractiveness to outsiders" (277). The 'showing-off' function is also identified by Verheijen (2016, 289) who observes abundant use of English in public tweets of Dutch youngsters who enjoy demonstrating to a large audience how cool they are.

Verheijen (2015, 133-134) generally observed a strong impact of medium in Dutch online communication. Instant messages appeared to contain much more non-standard writing than text messages and tweets (microblogging). The latter had the lowest score for non-standard forms. De Decker and Vandekerckhove (2017, 277-278) call for a distinction between expressive or playful CMC features and highly functional economical spelling choices: abbreviations appeared to be more frequent in synchronous data, whereas the expressive marker of flooding scored higher in asynchronous interaction. With respect to intensifier use, Herring (2001, 617) observes that synchronous media trigger a higher frequency of intensifiers because communication there is less formal than on asynchronous media.

As may be deduced from the above, impact of the medium is hard to predict. Moreover, apart from the public versus private character of the medium and the degree of synchronicity, there are other determining factors, such as the above-mentioned formality of the interaction and the contents of the messages. There may be huge differences between several asynchronous media with respect to these parameters. The tweets analyzed by Verheijen (2015; 2016) are often quite neutral in terms of formality, but the asynchronous messages examined in the

[9] Verheijen (2015, 129) notes that although the speed principle may not hold for asynchronous media, brevity can still be important, as some asynchronous genres have limited message size (e.g. tweets). In this paper, however, message size is no (sub)variable, as none of the medium variants represented in our corpus have limitations with respect to the length of the messages.

present study (and in De Decker 2014 and De Decker & Vandekerckhove 2017) have been extracted from a social media site which triggered quite personal and in most cases highly informal communication between youngsters (see Section 4.1).

3.4. Hypotheses and research questions

The discussion of the related research in the previous sections leads to the following hypotheses: since girls are generally supposed to have a stronger focus on establishing social and emotional connections, we assume they will produce more expressive markers than boys. In view of the fact that the older adolescents (see 4.1) are beyond the adolescent peak period, whereas the younger ones are in the midst of it or heading towards it, we hypothesize that the younger adolescents will outperform the older group in the frequency of use of the expressive markers. Finally, we assume that both due to the importance of linguistic self-presentation in the selected public asynchronous media and due to the greater time pressure in the synchronous media, fewer expressive markers will be used in the synchronous chat conversations.

The strength of the present study lies in the fact that it combines several independent variables and includes a wide range of expressive markers. The former enables us to discover the relative strength of several factors: what variable displays the strongest correlation with the use of expressive markers? What about the relative impact of the others? What are the implications for future CMC research? Furthermore, the inclusion of several types of expressive markers allows for a more detailed analysis of the preferences for specific markers by particular groups or in particular media. For instance, irrespective of the potential gender differences with respect to the overall frequency of the expressive markers, boys and girls might display distinct preferences with respect to choice or even realization of particular markers. Do these findings corroborate or nuance the overall age, gender or medium preferences?

4. Methodology

In this section, we will describe our corpus (Section 4.1) and the data extraction and processing (Section 4.2).

4.1. Corpus and participants

The corpus contains CMC data produced between 2007 and 2013 by Flemish adolescents aged 13 to 20.[10] So some adolescents are in their late teens or even rather young adults. We

[10] Apart from some additions, the corpus largely corresponds to the one used in De Decker & Vandekerckhove (2017). It was composed by the research group CLiPS of the University of Antwerp. Numerous students of the University of Antwerp contributed to the data collection (we note that these contributions were filtered, so that

take into account this discrepancy between young teenagers and adolescents nearing adulthood when dealing with the variable age. Furthermore, all of them are Dutch-speaking teenagers living in the north of Belgium. The entire corpus consists of 400 808 posts (i.e. utterances, delimited by carriage returns) or 2 066 521 tokens.[11] The utterances were produced on both synchronous and asynchronous electronic media. The synchronous or Instant Messaging (IM) media were MSN (i.e. Windows Live Messenger), which does not exist anymore, and Facebook Chat (Messenger). The rest of the corpus consists of posts produced on the – at that time – very popular Belgian social networking site Netlog. For some time, Netlog was considered the European equivalent of Facebook, but in recent years it could no longer compete with Facebook and the site closed in December 2014. Unlike the Facebook data in the present corpus, the Netlog data in our corpus do not only contain chat conversations, but also and predominantly data from asynchronous communication, such as blog posts, profile texts and comments on pictures. In other words, whereas the IM-corpora only cover data from synchronous conversations in real-time, data from mainly asynchronous and to a minor extent synchronous communication are mixed within the Netlog-corpus. Moreover, the Netlog-posts generally have a more public character: the posts and the reactions on the posts reach a wider audience (of peers) than the private IM-conversations. Therefore, we distinguish the private synchronous instant messages from the (largely) asynchronous public messages on the social media site Netlog (see Table 2).

For the age variable, we distinguish a younger group (aged 13-16) and an older group (aged 17-20) of adolescents. Table 2 shows the distribution of the tokens over the age and gender groups and the two media. Although there is an imbalance in the amount of data available for all three social variables (e.g. more male than female material), the smaller subcorpora are always sufficiently large and thus do not exclude valid testing for the three variables.

only utterances from teenagers aged 13-20 remained): they collected data in their own networks and donated these data, together with the information on the demographic profile of the chatters (age, gender, region, and in some cases also educational track). The data that were collected in this way were mainly produced on private synchronous media. The Netlog data (see below) were originally collected for the CLiPS project 'A safer internet: (Semi)automatically recognizing internet paedophilia in multilingual online social networks'. For more information on the project and the data collection, see Peersman, Daelemans & Van Vaerenbergh (2011). All of the data were anonymized: the information on the social profile of the chatter is no longer linked to the name of the chatter, nor can the names be traced back. More information on the entire corpus can be found in De Decker (2014, 23-28).

[11] These tokens are the result of splitting the text on whitespace. They were counted automatically. A token can be a word, an emoticon or isolated punctuation marks.

	Girls		Boys		
	Younger	**Older**	**Younger**	**Older**	**Total**
Private SYNC.	118 694	176 233	29 146	973 061	1 297 134
Public ASYNC.	463 277	67 257	162 077	76 776	769 387
Total	581 971	243 490	191 223	1 049 837	2 066 521

Table 2: Distribution of variables in the corpus

4.2. Data extraction and processing

The present section provides some explanation on the automatic extraction and quantitative processing of the tokens for the expressive markers.

4.2.1. Typographic and onomatopoeic expressive markers

All occurrences of the typographic and onomatopoeic expressive markers were detected and counted automatically by using Python scripts. The software's performance was evaluated and judged accurate on a test set of 1000 randomly chosen posts (5595 tokens) from the corpus by comparing a human annotator's decisions to the software's output. For the seven automatically detected expressive variables, the average precision – i.e. the (relative) number of detected occurrences of a marker that actually are legal occurrences of that marker – is very high: 98%. The average recall – i.e. the (relative) number of occurrences of a marker that were actually detected as occurrences of that marker – is high as well: 95%.

4.2.2. Intensifiers

The intensifiers were automatically extracted using a predefined list covering most of the lemmas (and their variants) present in our corpus. Yet, this method is not exhaustive, as less popular or less obvious intensifying modifiers are not retrieved, nor are intensifiers containing unexpected spelling mistakes or typographical errors. Because of the large size of the corpus, however, the impact of such errors can be assumed to be minimal. With respect to the final selection, we added a frequency cutoff: only lemmas (types) that occurred at least fifteen times in the entire corpus, of which at least five times as an intensifier, were preserved. This cutoff resulted in a list of 23 intensifiers.[12]

[12] In alphabetical order: (1) *bere*, (2) *echt*, (3) *echt wel*, (4) *erg*, (5) *fucking*, (6) *gans*, (7) *heel*, (8) *kei*, (9) *kweetniehoe*, (10) *loei*, (11) *mass(as)*, (12) *massiv*, (13) *mega*, (14) *muug*, (15) *over*, (16) *overdreven*, (17) *so*, (18) *super*, (19) *vies*, (20) *vree*, (21) *zeer*, (22) *zo*, (23) *zot.* Adding the frequency cutoff was needed in the original study (Vercammen 2014-2015), where the use of intensifiers (on its own) was correlated with several variables: age, gender, region. Therefore, we needed enough tokens in each of the cells. Moreover, in view of the size of the corpus, it seems sensible not to include intensifiers with an extremely low occurrence.

We did not select intensifiers that appeared in a negative or interrogative context (cf. Ito & Tagliamonte 2003, 264 and Palacios Martínez & Pertejo 2012, 779). In these contexts, the adjectives or adverbs that follow the intensifiers are often mitigated rather than intensified or emphasized, as illustrated by example (20).

(20) *ma je moet **ni super**veel prentjes ebbn*
'but you **don't** need **that** many images'

After automatic extraction, we manually screened and filtered the software's output, i.e. for each utterance, we checked if the intensifying words were truly used as an intensifier. This finally rendered 14 269 tokens for the selected set of intensifiers. A test set of 700 intensifiers in context was screened by two annotators, who obtained a disagreement of only 1.57% (i.e. the percentage of truly ambiguous utterances containing an intensifier).

5. Results and discussion

This section presents the results of the analyses. It starts with the general findings (Section 5.1) and is followed by a more detailed discussion of some of the patterns on the level of the individual markers (Section 5.2). To verify the statistical significance of our quantitative findings, we combined chi-square tests with a bootstrapping approach (Monte Carlo resampling).[13] With this approach, we can obtain more solid results than when performing one single chi-square test on the entire data set, because we can estimate the (sampling distribution of the) statistics: we first calculated the statistics of interest (chi-square value, p-value, etc.) for each sample and stored them, and finally, we computed the average values (as well as the corresponding standard deviations and confidence intervals). The statistical values reported in the next paragraphs are the mean values for all bootstrap samples.

5.1. General findings

An overview of all expressive markers in the corpus in terms of relative and absolute frequency is shown in Table 3. We note that the use of some markers is more limited than others, depending on their function (e.g. an emoticon can be inserted almost anywhere in an utterance, whereas several grammatical constraints limit intensifier use).

[13] Bootstrapping is a statistical technique in which the original data set is resampled by picking *n* bootstrap samples randomly and with replacement, in order to estimate (the sampling distribution of) a statistic (Efron & Tibshirani 1998, 12; Field 2009, 782). By doing so, one is "treating the data as a population from which smaller samples are taken" (Field 2009, 782). We resampled our corpus by picking 10 000 random samples, each containing 100 000 tokens, chosen with replacement (a same token could thus occur more than once in one sample).

	Absolute number	Percentage of all markers	Percentage of total number of tokens (*) or question and exclamation marks (**)[14]
Laughter (*)	11 412	3.87 %	0.55 %
Emoticons (*)	150 895	51.13 %	7.30 %
Allcaps (*)	15 029	5.09 %	0.73 %
Kisses (*)	45 129	15.29 %	2.18 %
Flooding letters (*)	40 479	13.72 %	1.96 %
Flooding punctuation marks (**)	17 213	5.83 %	12.18 %
Combinations of question and exclamation marks (**)	701	0.24 %	0.50 %
Intensifiers (*)	14 269	4.83 %	0.69 %
Total	**295 127**	**100 %**	

Table 3: Absolute and relative frequencies for each expressive marker in the entire corpus

For the analyses, we quantified the degree of expressiveness by dividing the number of expressive markers in the (sub)corpus by the total number of tokens in the (sub)corpus. This approach led to relative expressiveness scores or ratios. The entire data set contained 295 127 expressive markers, which is a ratio of 14.28%. An overview of the ratios per independent variable is shown in Table 4. These percentages should be interpreted with caution. A score of 9.30% does not imply that 9.30% of all tokens in the relevant subcorpus contains an expressive marker. In fact, a smaller percentage of all tokens actually contains an expressive marker, since some tokens contain more than one expressive feature (e.g. combinations of letter flooding and allcaps in one word: *SUUUUUPER*). Yet these scores present a reliable indication of the relative representation of expressive markers in the adolescent groups and media. The asynchronous posts contain the highest ratio of expressive markers (28.35%), followed by the younger participants' texts (25.23%) and the girls' texts (21.77%).

Female	**Male**
21.77%	9.30%
Younger (13-16)	**Older (17-20)**
25.23%	7.74%
Public/Asynchronous posts	**Private/Synchronous posts**
28.35%	5.94%

Table 4: Overview of expressiveness ratios per subcorpus

[14] As the use of some markers is more limited than others (e.g. because of grammatical constraints), they will naturally occur less frequently. We partially normalized these quantitative differences by counting features related to punctuation in a different way than the other markers. The relative frequency of punctuation flooding and of combinations of question and exclamation marks was obtained by dividing the absolute counts not by the number of tokens in the (sub)corpus (which was done for all other markers), but by the number of occurrences of question and exclamation marks in the (sub)corpus. This increased the otherwise very low relative frequency of these expressive markers

General tendencies for the social variables are that girls use significantly more expressive markers than boys ($p < .001$, chisq. = 3044.57, df = 1) and that younger teenagers integrate significantly more of them than older ones ($p < .001$, chisq. = 5850.01, df = 1). Furthermore, expressive markers score much higher on the public/asynchronous medium than on the private/synchronous media ($p < .001$, chisq. = 9274.18, df = 1). In view of the imbalance of several subgroups in relation to particular variables (e.g. older boys are dominant in the synchronous data), we also tested the impact of each independent variable while keeping the other variables constant (for every possible combination of subgroups). Apart from one exception, the observed tendencies were confirmed and turned out to be significant.[15] Moreover, these general tendencies also hold for *each* of the expressive markers: every single expressive marker occurs more frequently in female, younger and public / asynchronous texts than in male, older and private / synchronous texts respectively.

In order to assess the strength of the association between the linguistic and independent variables, we looked at the Cramer's V scores (here identical to Phi scores),[16] which rank from 0 to 1 (Field 2009: 699). The strongest association is found for medium (Cramer's V = 0.31), followed by age (Cramer's V = 0.24) and gender (Cramer's V = 0.17). Apart from that, we took into account the effect size – i.e. a 'measure of the magnitude of observed effect' (Field 2009: 56) – by calculating the odds ratio scores[17] per experiment. These ratios rank from 1 to infinite (or, in the inversed notation, from 0 to 1): 'an odds ratio of 1 would indicate that the *odds* of a particular outcome are equal in both groups' (Field 2009: 790). The odds ratios appear to display the same order as the Cramer's V or Phi scores: medium has the largest effect size (odds ratio = 6.27), followed by age (odds ratio = 4.02) and gender (odds ratio = 2.71). In other words, the odds that a token contains an expressive marker are 6.27 times higher if the token is produced within the asynchronous medium than when produced within the synchronous media in our corpus.[18] Medium definitely appears to be the strongest determinant of expressiveness. The correlation with the linguistic variables appears to be very strong and the effect size is much larger than for the other variables.

[15] We ran 12 subtests: 4 per social variable. We will illustrate our approach for gender. We compared the younger **girls'** synchronous data to the younger **boys'** synchronous data (test 1), the younger **girls'** asynchronous data to the younger **boys'** asynchronous data (test 2), the older **girls'** synchronous data to the older **boys'** synchronous data (test 3) and finally the older **girls'** asynchronous data to the older **boys'** asynchronous data (test 4). In these subtests, gender is always the only variable that changes; medium and age remain constant. The only subtest in which the observed tendency was not significant, was the final gender test: older girls used more expressive markers in asynchronous posts than their male peers, but not significantly so.

[16] Cramer's V and Phi are "measures of the strength of association between two categorical variables" (Field 2009, 695). In our experimental setup (with two categorical variables per experiment, each containing two subcategories), the two measures are identical (Field 2009, 698), and are "calculated by taking the chi-square value and dividing it by the sample size and then taking the square root of this value" (Field 2009, 695).

[17] Field (2009, 790) defines odds ratio as "the ratio of the *odds* of an event occurring in one group compared to another".

[18] Note that these numbers differ from the ratios reported in Table 3. Although both numbers express a similar concept, the calculation behind them is different, as sample sizes of both subcorpora are taken into account to calculate odds ratios and not to calculate the straightforward percentages.

Some markers produce remarkably high odds ratios. This is the case for letter flooding (deliberate, expressive letter repetition) and the rendition of kisses (e.g. *xxx*), especially with regards to medium. The odds ratios are 51.85 (kisses – medium) and 16.33 (letter flooding – medium): the odds of a token containing a rendition of kisses (letter flooding, resp.) are 51.85 times higher (16.33, resp.) when that token is produced in a public/asynchronous utterance instead of in a private/synchronous post. Markers that were strongly associated with the two other independent variables were letter flooding (CV 0.11, OR 5.53) for gender, and letter flooding (CV 0.14, OR 8.99) and kisses (CV 0.13, OR 6.20) for age. In other words, girls and young adolescents show a strikingly stronger preference for letter flooding than boys and older adolescents, and *x*'s representing kisses are much more frequent in younger adolescents' CMC than in that of the older ones.

5.2. Patterns on the level of the individual markers

5.2.1. General tendencies

The data display some striking constants across all different subgroups with respect to certain patterns or preferences on the level of the individual markers. The present section presents a selection of the dominant tendencies. While the percentages reported in the next paragraphs are the relative counts for the *entire* corpus, the same tendencies were actually found in all six[19] subcorpora.

The most popular expressive markers in all groups are punctuation flooding and emoticons (with relative frequencies of 12.18% and 7.30% resp.). For punctuation flooding, the difference may be (partly) ascribed to the fact that the ratio was not calculated in the same way as for the other markers (see footnote 14). Since we relate the tokens of flooding of exclamation and question marks to all occurrences of these punctuation marks instead of to all tokens in the corpus, the ratio inevitably is higher than for the other markers. However, this does not apply to the emoticons. A possible explanation for their popularity is that these features are very explicit expressive markers: emoticons often literally represent a facial expression. They are very obvious and consequently favored expressive markers.

Another tendency concerns letter flooding: in all subgroups, mainly vowels are repeated (91% of all occurrences of this expressive marker) and hardly ever plosives (2%). Liquids, fricatives and nasals occupy an intermediate position in this respect. This supports the hypothesis that flooding is the (CMC-specific) orthographic representation of an oral phenomenon (Darics 2013, 144), i.e. the lengthening of sounds, which is most natural for vowels and impossible for plosives. Concerning the nature of the words that were emphasized through letter flooding, we found that many of the top lexemes are positively qualifying adjectives (30% of the top 100 types containing letter repetition), mainly variants of the Dutch adjective *mooi* ('beautiful') (22% of the top 100 types). While adolescent language generally has a strong

[19] The six subcorpora are: female texts, male texts, younger texts, older texts, and synchronous and asynchronous posts.

focus on how people and things are valued and experienced (Taylor 2001, 299) with an abundance of evaluative vocabulary (Androutsopoulos 2005, 1497), the nature of the asynchronous data certainly contributes to the top position of the positive qualifiers in the flooding data: a large part of these social media posts are positive reactions to other users' profile pictures, which often involve some degree of pleasing or even flirting (see also Section 5.2.2). A similar tendency could be found for the use of intensifiers: the adjective *mooi* represents 18% of the intensified adjectives and adverbs.

With respect to the use of allcaps as an expressive marker, we note the top position of the Dutch first person singular pronoun *ik* ('I') (1.44% of all capitalized lexemes in the entire corpus, and the type that was most often written in capitals by all subgroups). Function words are generally used more frequently than content words (Newman et al. 2008, 216; Pennebaker 2011, 27), but the top position of *ik* might be symptomatic of the intense personal self-expression of the teenagers. Quite often, the pronoun is integrated in an utterance that is consistently written in allcaps.[20]

Furthermore, we note a preference for 'simple' variants for the rendition of laughter and kisses and the combination of question and exclamation marks. The three most popular ways of expressing laughter were *haha*, *hahaha* and *hihi*, which are the shortest variants (55% of all onomatopoeic renditions of laughter). The three most popular ways of expressing kisses were *x*, *xx*, *xxx*, also the shortest variants (95% of all renditions of kisses), and finally, the most popular combinations of question and exclamation marks were simply *!?* and *?!* (58% of all occurrences of this feature). These preferences could be interpreted in terms of the speed principle: typing the compact variants is more economical. Apart from that, the less elaborate variants simply seem to be highly conventional, even beyond CMC-contexts: *haha*, for instance, is a very international and common way to express laughter. An interesting (though not academic) tool to estimate the degree of conventionalization and 'internationalization' in informal 'speech' worldwide is the representation and interpretation of these features on Urban Dictionary.[21] The lemma *haha* for instance is identified as a 'short quick way of letting somebody know you are laughing, most likely at them' while its longer variant *hahaha* gets a deviant and more specific interpretation.[22] The same accounts for *x*, *xx* and *xxx*: they are all being identified as kisses on Urban Dictionary, but longer variants as *xxxx* or *xxxxx* are not defined as such. While this type of source has to be handled extremely carefully, it gives a

[20] Manual screening of the output revealed that the impact of typographic errors for this phenomenon is negligible: *IK* was almost always capitalized deliberately (i.e. either integrated in an entirely capitalized sentence (58 out of the 61 cases), or emphasized in a lowercased sentence, in contrast with another pronoun, i.e. *JIJ* ('you') (1 occurrence)). Only in two cases it could not be excluded that the chatter capitalized the entire pronoun unintentionally. So the potential 'mechanical' influence of capitalizing digitally (i.e. accidentally capitalizing not only the first letter at the beginning of a sentence, but the next one as well) for this particular token appeared to be very small.

[21] urbandictionary.com

[22] "To express on aim when something was funny, because just 'haha' isn't that dramatic and can be used as just aknowledging [sic] when someone has said something."

clue with respect to the extent to which particular features are universal and mainstream in informal (online) communication.

Concerning emoticons, finally, we found that the Western variants are the most popular ones among all groups of participants (68% of all emoji in the corpus). They are among the oldest ones (together with the manually composed Asian variants) and are used worldwide, contrary to some of the emoticons that were typical of the Dutch-Flemish social medium of Netlog and of MSN.[23] Western emoticons are (at least for our Western participants) also quite easy to interpret and to create: they are simple visual representations of facial expressions. The more recent and highly popular Unicode emoji are not yet present in our corpus, which dates back to 2007-2013. The most popular Western emoticons in the corpus are:

:P or :p	(sticking out tongue)
:D	(laughing)
:)	(smiling)

These variants figure in the top five emoticons for the entire corpus as well as in the top ten for each of the subcorpora, and thus appear to be very popular among all gender and age groups and on all platforms.

5.2.2. Correlations between gender, age and medium

Finally, the in-depth analyses for each of the expressive markers also lay bare correlations between the independent variables. Strikingly, parallel tendencies could be noted for texts written by female participants, by younger teenagers, and on the public/asynchronous medium. What these have in common is, for instance, that they contain many more expressive markers related to love and friendship than those of their male, older adolescent and private/synchronous counterparts. The most popular emoticons (top 3 or 2 for each of these three groups) were all related to love (e.g. the heart-emoticon *<3* or love-related emoticons bound to the specific chatroom and social media site used in this study). Heart-variants specifically figured quite frequently in these posts (9 to 10% of the emoticons used by each of the three groups). Furthermore, many of the top lexemes that were written in allcaps concerned love or friendship (at least 10% of the top 100 lexemes written in allcaps for each group) (e.g. *LOVEYOU*, *BFF*: 'best friend forever'). The same holds for the lexemes that contained letter flooding: 8 to 11% of the top 100 lexemes containing letter flooding for each group were love- or friendship-related (e.g. *iloveyouuu*).

These results manifest a strong discrepancy with boys' and older adolescents' CMC and with practices on the private/synchronous media. In these subcorpora, the top emoticons were not related to love or friendship, nor were heart-variants popular emoticons. On the contrary,

[23] We note that on these two platforms, the traditional facial expressions can be typed manually (e.g. smiling face as :)) and are then converted to a pictogram. For subtler or more elaborate expressions, the platform-specific images need to be selected from the interface. This explains why, for instance, the most popular Netlog variants do not contain smiling faces (as these are not Netlog-specific), but include hugging figures and a blushing face.

they were even the least favored variants (0.40% to 2% of the emoticons used by each of these groups). Only few of the top lexemes containing letter flooding concerned love or friendship (0 to 5% of the top 100 flooded lexemes for each group), and even fewer of the lexemes written in capitals (1 to 3% of the top 100 allcaps lexemes for each group). These three groups' top emoticons contained more representations of negative emotions (e.g. :(, -_- and :/ , respectively a sad, frustrated and confused face). Many of the lexemes written in allcaps were exclamations (*YEAH*, *WOW*, *BAM*) and 'tougher' words, such as curse words, insults and taboo words (*FUCK*, *ASS*, *GAY*, *GVD* – short for *godverdomme* 'goddamn it'). Finally, most of the boys' and older teenagers' flooded words were positively evaluating adjectives concerning appearance (e.g. *mooooi* for *mooi* 'beautiful'). These lexemes' relatively low frequency in the synchronic posts suggests that the positive evaluations primarily concern (profile) pictures, typical of the asynchronous medium. For boys and older teenagers, and on the synchronous medium specifically, interjections are often flooded (*pfffff*, *ooooh*, *aaaah*) as well as exclamations and greetings (*heeey*).

However, some caution might be needed when interpreting these correlations, as there is an imbalance in our dataset which could (partially) influence our results: young female participants in public asynchronous CMC are overrepresented in our corpus, and so are older male participants in private synchronous CMC (see Table 2). Still, similar correlations between gender and age were reported on before (Argamon, Koppel, Pennebaker, & Schler 2007; Pennebaker 2011; Schwartz et al. 2013). Stylistic correlations concern the use of function words: men and older people use them in similar ways (using more articles and prepositions), as do younger people and women (using more pronouns, conjunctions and auxiliary verbs) – a tendency which seems to hold across cultures, languages and time (Argamon et al. 2007, n.p.; Pennebaker 2011, 66; Schwartz et al. 2013, 8-9). On a content-related note, correlations between the same two age and gender groups can be distinguished. Argamon et al. report that men and older people prefer topics like politics, religion and business, whereas women and younger people prefer discussing home, romance and fun (2007, n.p.). These findings correspond to the younger and female teenagers' preferences for expressive markers related to love and friendship.

As for medium, however, no correlations have been reported between the way people write on certain platforms and their gender or age. This could thus be an artefact of the imbalance in our dataset. Another possible explanation lies in the nature of our asynchronous texts. Although many posts on the asynchronous medium are public, the interaction often has a largely personal character. Many comments on this social medium involve pleasing and even flirting (e.g. in positive reactions to other users' pictures). In this respect, our asynchronous medium differs from other public social media, like Twitter, where the communication is less personal and more targeted at informing a wider audience, rather than at bonding or pleasing. The latter focus prevails in our public-asynchronous data: the medium is not only used for intensifying existing bonds, but also for establishing new network connections, friendship ties and even for dating. By using the love and friendship-related expressive

markers (and the other ones), young adolescents acquire social capital. This might explain the higher rate of these markers in the public medium than in the private media, in which people interact with friends from their existing peer group network.

6. Conclusion

This paper discussed linguistic expressiveness in a corpus of Flemish adolescents' computer-mediated messages. We included typographic CMC features (e.g. emoticons), an onomatopoeic variable (the rendition of laughter) and a lexical feature (the use of intensifiers) and looked for possible correlations between these linguistic variables and the author's profile (gender, age) versus the synchronicity and the public versus private character of the CMC medium.

Girls used more expressive markers than boys, and so did the younger adolescents compared to the older ones. The results were extremely consistent in this respect: the same tendencies could be observed for each of the expressive markers. Quite strikingly, however, medium appeared to have the largest impact (more expressive writing in asynchronous and largely public social media posts than in synchronous and mainly private instant messages). Furthermore, the qualitative analyses show that girls and younger teenagers produce more love-related expressive markers than boys and older adolescents. And again, remarkably, these types of correlations were found for medium too (with more love-related markers used in the public/asynchronous than in the private/synchronous posts).

The present research differs from previous research into expressive markers in CMC in that it includes a wider range of expressive markers (both lexical and typographic) combined with three independent variables (age, gender and medium). While gender and, to a minor extent, age have received ample attention in related research, the present findings highlight the importance of the variable medium. They call for refinement of this variable, since apart from (a)synchronicity and the public versus private character of the medium, the character and goal of the interaction seem to be determinant factors too and consequently need to be operationalized in future research. The behavior of the expressive markers is quite revealing in that respect. De Decker and Vandekerckhove (2017) included only one expressive marker (i.e. flooding)[24] and were struck by the consistent age, gender and medium correlations for this variable. They suggested follow-up research with a wider inclusion of expressive markers in order to enhance insights in the operationalization of (emotional and social) expressiveness in CMC and the way it functions as an identity marker or as an identifying factor for specific subgroups. The present research does not only reveal that expressive markers are particularly age and gender-sensitive but suggests that they serve specific goals: bonding, pleasing, building up social capital, etc. Consequently, their use culminates on media in which these are the main driving forces of the interaction, as was the case with the Netlog medium in the

[24] Without making a distinction between letter flooding and punctuation flooding.

present study, and in social groups that tend to invest heavily in these activities or goals. Young adolescents are intensively engaged in identity construction and extremely sensitive to peer group evaluation. Much of their interaction is driven by a 'need of acceptance and fear of rejection' (Taylor 2001, 298). Expressive markers (that often accompany positively qualifying adjectives – see above) seem to be favored tools in that process of both identity and social network construction, in which angling for approval may be a major determining factor. With respect to gender, female discourse is supposed to have a stronger focus on the establishment of social and emotional connections. The consistent gender findings suggest that girls stress their involvement through the use of expressive markers and especially through the use of features that express friendship and love. In view of young adolescents' dependence on peer group approval, it is hardly surprising that they share the latter preference with women/girls.

Our final conclusion, which concerns suggestions for future research, is that the impact of CMC media deserves more attention. We pointed to the importance of the goals and nature of the interaction, or the general communicative function of the medium. Apart from these aspects, other medium-related properties might be incorporated in the research design as well, like the technology or device that is used (e.g. smartphone or pc), the potential impact of spelling checkers or autocorrection, and limitations in message size. Furthermore, other features or devices for emotional expression could be included in future research, like lexical and syntactical expressions. Finally, it might be challenging to disentangle explicit expressions of positive or negative emoticons from subtler implicit ironic or sarcastic connotations.

Acknowledgments

We thank Giovanni Cassani and Dominiek Sandra for their help and advice in the statistical aspect of the paper. We are also grateful towards Jens Vercammen for the manual data processing of the lexical variable. Finally, we thank the editorial board of *Nederlandse Taalkunde*/*Dutch Linguistics* and the two anonymous reviewers for their helpful and constructive feedback on an earlier draft of this paper. This work was supported by the FWO (Research Foundation Flanders).

References

Androutsopoulos, Jannis. (2005). Research on youth language. In Ulrich Ammon, Norbert Dittmar, Klaus J. Mattheier, & Peter Trudgill (Eds), *Sociolinguistics: An international handbook of the science of language and society (vol. 2)* (pp. 1496-1505), Berlin: Mouton de Gruyter.

Androutsopoulos, Jannis. (2011). Language change and digital media: A review of conceptions and evidence. In Tore Kristiansen, & Nikolas Coupland (Eds), *Standard languages and language standards in a changing Europe* (pp. 145-161), Oslo: Novus.

Argamon, Shlomo, Moshe Koppel, Jonathan Fine, & Anat Rachel Shimoni. (2003). Gender, genre, and writing style in formal written texts. *Text* 23, 321-346.

Argamon, Shlomo, Moshe Koppel, James W. Pennebaker, & Jonathan Schler. (2007). Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday* 12, n.p.

Argamon, Shlomo, Moshe Koppel, James W. Pennebaker, & Jonathan Schler. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM. Inspiring Women in Computing* 52, 119-123.

Baron, Naomi S. (1984). Computer mediated communication as a force in language change. *Visible Language* 18, 118-141.

Baron, Naomi S. (2004). See you online: Gender issues in college student use of instant messaging. *Journal of Language and Social Psychology* 23, 397-423.

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

Bing, Janet M., & Victoria L. Bergvall. (1996). The question of questions: Beyond binary thinking. In Jennifer Coates (Ed.), *Language and gender: A reader* (pp. 495-510), Oxford: Blackwell.

Coates, Jennifer. (1993). *Women, men and language: A sociolinguistic account of gender differences in language.* London / New York: Longman.

Crystal, David. (2001). *Language and the internet*. Cambridge: Cambridge University Press.

Daelemans, Walter. (2013). Explanation in computational stylometry. In *International conference on intelligent text processing and computational linguistics* (pp. 451-462), Berlin: Springer.

De Decker, Benny, & Reinhild Vandekerckhove. (2017). Global features of online communication in local Flemish: Social and medium-related determinants. *Folia Linguistica* 51, 253-281.

De Decker, Benny. (2014). *De chattaal van Vlaamse tieners: Een taalgeografische analyse van Vlaamse (sub)standaardiseringsprocessen tegen de achtergrond van de internationale chatcultuur*. Antwerp: University of Antwerp (doctoral thesis).

Eckert, Penelope. (1997). Age as a sociolinguistic variable. In Florian Coulmas (Ed.), *The handbook of sociolinguistics* (pp. 151-167), Oxford: Blackwell.

Eckert, Penelope. (1998). Gender and sociolinguistic variation. In Jennifer Coates (Ed.), *Language and gender: A reader* (pp. 64-75), Oxford: Blackwell.

Eckert, Penelope. (2003). Language and adolescent peer groups. *Journal of Language and Social Psychology* 22, 112-118.

Efron, Bradley, & Robert J. Tibshirani. (1998). *An introduction to the bootstrap.* Boca Raton / London / New York / Washington D.C.: Chapman & Hall / CRC.

Eisikovits, Edina. (2006). Girl-talk/boy-talk: Sex differences in adolescent speech. In Jennifer Coates (Ed.), *Language and gender: A reader* (pp. 42-54), Oxford: Blackwell.

Field, Andy. (2009). *Discovering statistics using SPSS.* Los Angeles / London / New Delhi / Singapore / Washington DC: SAGE.

Herring, Susan C. (1996). Two variants of an electronic message system. In Susan C. Herring (Ed.), *Computer-mediated communication: Linguistic, social and cross-cultural perspectives* (pp. 81-106), Amsterdam: John Benjamins.

Herring, Susan C. (2001). Computer-mediated discourse. In Deborah Schiffrin, Deborah Tannen, & Heidi E. Hamilton (Eds), *The handbook of discourse analysis* (pp. 612-634), Malden / Oxford: Blackwell.

Herring, Susan C., & Anna Martinson. (2004). Assessing gender authenticity in computer-mediated language use: Evidence from an identity game. *Journal of Language and Social Psychology* 23, 424-446.

Herring, Susan C. (2012). Grammar and electronic communication. In Carol A. Chapelle (Ed.), *Encyclopedia of applied linguistics*, *S.l.*: Wiley.

Holmes, J. (1992). *An introduction to sociolinguistics*. London / New York: Longman.

Huffaker, David A., & Sandra L. Calvert. (2005). Gender, identity, and language use in teenage blogs. *Journal of Computer-Mediated Communication* 10, n.p.

Ito, Rika, & Sali A. Tagliamonte. (2003). Well weird, right dodgy, very strange, really cool: Layering and recycling

in English intensifiers. *Language in Society* 32, 257-279.

Jespersen, Otto. (1922). *Language: Its nature, development and origin.* London: George Allen & Unwin.

Koch, Peter, & Wulf Oesterreicher. (2001). Gesprochene Sprache und geschriebene Sprache / Langage parlé et langage écrit. In Günter Holtus, Michael Metzeltin, & Christian Schmitt (Eds), *Lexikon der Romanistischen Linguistik (vol. 1:2)* (pp. 584-627), Tübingen: Max Niemeyer Verlag.

Koch, Peter, & Wulf Oesterreicher. (2011). *Gesprochene Sprache in der Romania: Französisch, Italienisch, Spanisch.* Berlin / New York: De Gruyter.

Kucukyilmaz, Tayfun, B. Barla Cambazogly, Cevdet Aykanat, & Fazli Can. (2006). Chat mining for gender prediction. In *International conference on advances in information systems* (pp. 274-283), Berlin: Springer.

Lorenz, Gunter R. (1999). *Adjective intensification. Learners versus native speakers: A corpus study of argumentative writing*. Amsterdam / Atlanta: Rodopi.

Lorenz, Gunter R. (2002). Really worthwhile or not really significant? A corpus-based approach to the delexicalisation and grammaticalisation of intensifiers in Modern English. In Diewald Wischer (Ed.), *Typological studies in language (vol. 49): New reflections on grammaticalization* (pp. 143-161), Amsterdam / Philadelphia: John Benjamins.

Méndez-Naya, Belén. (2003). On intensifiers and grammaticalization: The case of swithe. *English Studies* 84, 372-391.

Newman, Matthew L., Carla J. Groom, Lori D. Handelman, & James W. Pennebaker. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes* 45, 211-236.

Paradis, Carita. (2000). It's well weird. Degree modifiers of adjectives revisited: The nineties. In John M. Kirk (Ed.), *Corpora galore. Analyses and techniques in describing English* (pp. 147-160), Amsterdam: Rodopi.

Parkins, Róisín. (2012). Gender and emotional expressiveness: An analysis of prosodic features in emotional expression. *Griffith Working Papers in Pragmatics and Intercultural Communication* 5, 46-54.

Partington, Alan. (1993). Corpus evidence of language change: The case of intensifiers. In Mona Baker, Gill Francis, & Elena Tognini-Bonelli (Eds), *Text and technology: In honour of John Sinclair* (pp. 177-192), Amsterdam / Philadelphia: John Benjamins.

Peersman, Claudia, Walter Daelemans, & Leona Van Vaerenbergh. (2011). Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on search and mining user-generated contents* (pp. 37-44).

Peersman, Claudia, Walter Daelemans, Reinhild Vandekerckhove, Bram Vandekerckhove, & Leona Van Vaerenbergh. (2016). The effects of age, gender and region on non-standard linguistic variation in online social networks. <http://arxiv.org/abs/1601.02431>

Pennebaker, James W. (2011). *The secret life of pronouns: What our words say about us.* New York: Bloomsbury press.

Peters, Hans. (1994). Degree adverbs in early modern English. In Dieter Kastovsky (Ed.), *Studies in early modern English* (pp. 269-288), Berlin / New York: De Gruyter.

Pyles, Thomas, & John Algeo. (1993). *The origins and development of the English language*. Boston: Heinle.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, Jan Svartvik, & David Crystal. (1985). *A comprehensive grammar of the English language*. New York: Longman.

Schlobinski, Peter. (2005). Mündlichkeit/Schriftlichkeit in den Neuen Medien. In Ludwig Eichninger, & Werner Kallmeyer (Eds), *Standardvariation: Wieviel Variation verträgt die Deutsche Sprache?* (pp. 126-142), Berlin: De Gruyter.

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. *PLoS ONE* 8, n.p.

Stenström, Anna-Brita, Gisle Andersen, & Ingrid Kristine Hasund. (2002). Non-standard grammar and the trendy use of intensifiers. In Anna-Brita Stenström, Gisle Andersen, & Ingrid Kristine Hasund (Eds), *Trends in teenage talk: Corpus compilation, analysis and findings* (pp. 131-163), Amsterdam: John Benjamins.

Stoffel, Cornelis. (1901). *Intensives and down-toners*. Heidelberg: Carl Winter.

Tagliamonte, Sali A., & Chris Roberts. (2005). So weird; so cool; so innovative: The use of intensifiers in the television series Friends. *American Speech* 80, 280-300.

Tagliamonte, Sali A. (2008). So different and pretty cool! Recycling intensifiers in Toronto, Canada. *English Language and Linguistics* 12, 361- 394.

Tagliamonte, Sali A., & Derek Denis (2008). Linguistic ruin? Lol! Instant messaging and teen language. *American Speech* 83, 3-34.

Thurlow, Crispin, & Michele Poff. (2013). Text messaging. In Susan Herring, Dieter Stein, & Tuija Virtanen (Eds), *Pragmatics of computer-mediated communication* (pp. 163-190), Berlin / New York: Mouton de Gruyter.

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. *Reading and Writing* 23, 719-733.

Vercammen, Jens. (2014-2015). *"Echt massiv sketchy". Versterkers in de chattaal van Vlaamse jongeren: Een corpusonderzoek naar gender-, leeftijds- en regionale verschillen.* Antwerp: University of Antwerp (unpublished master thesis).

Verheijen, Lieke. (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).

Verheijen, Lieke. (2016). De macht van nieuwe media: Hoe Nederlandse jongeren communiceren in sms'jes, chats en tweets. In Dorien Van De Mieroop, Lieven Buysse, Roel Coesemans, & Paul Gillaerts (Eds), *De macht van de taal: Taalbeheersingsonderzoek in Nederland en Vlaanderen* (pp. 275-293), Leuven / The Hague: Acco.

Wolf, Alecia. (2000). Emotional expression online: Gender differences in emoticon use. *Cyberpsychology & Behavior* 3, 827-833.

CHAPTER 3

This chapter was published as a journal article. Reference:

Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans. (2018). Adolescents' social background and non-standard writing in online communication. *Dutch Journal of Applied Linguistics* 7(1), 2-25.

Adolescents' social background and non-standard writing in online communication

Abstract
In a large corpus (2.9 million tokens) of chat conversations, we studied the impact of Flemish adolescents' social background on non-standard writing. We found significant correlations between different aspects of social class (level of education, home language and profession of the parents) and all examined deviations from formal written standard Dutch. Clustering several social variables might not only lead to a better operationalization of the complex phenomenon of social class, it certainly allows for discriminating social groups with distinct linguistic practices: lower class teenagers used each of the non-standard features much more often and in some cases in a different way than their upper class peers. Possible explanations concern discrepancies in terms of both linguistic proficiency and linguistic attitudes. Our findings emphasize the importance of including social background as an independent variable in variationist studies on youngsters' computer-mediated communication.

Keywords: chatspeak, computer-mediated communication, social class, computational sociolinguistics, adolescents

1. Introduction

Many sociolinguistic studies have reported on the impact of age and gender. As for age, non-standard language use is said to be highest during adolescence, peaking around the age of fifteen – i.e. the 'adolescent peak' (Holmes 1992, 184; Peersman, Daelemans, R. Vandekerckhove, B. Vandekerckhove & Van Vaerenbergh 2016, 16-17) – due to peer "group pressure to not conform to established societal conventions" (Nguyen, Doğruöz, Rosé & de Jong 2016, 17). As teenagers age, their language use converges towards the adult standard, since for adults "social advancement matters and they use standard language [or linguistic varieties which more closely approach the standard language] to be taken seriously" (Nguyen et al. 2016, 17 – our insertion). As for gender, female language use has often been found to be more standardized: women are said to prefer the 'overt prestige' of standard language (associated with status, ambition, social mobility, etc.), and men the 'covert prestige' of the vernacular[1] (associated with values such as solidarity, toughness, kindness, humor, etc.) (Coates 1993, 80-83). However, when it comes to deviations from formal writing norms in computer-mediated communication (CMC), women appear to make greater use of particular features (e.g. expressive markers) than men (Hilte, Vandekerckhove & Daelemans 2016, 31-

[1] For possible explanations (more particularly differences in social positions and in attitudes towards language and its goal), we refer to Coates (1993, 82-85) and Trudgill (1983a, 162-168).

32; Kucukyilmaz, Cambazogly, Aykanat & Can 2006, 282; Parkins 2012, 48, 50, 53; Wolf 2000, 831).

Studies on the linguistic impact of social class are more scarce, at least when it comes to adolescent speech and CMC. However, the following findings might be relevant, although they do not relate to CMC practices: both Coates (1993, 80-82) and Trudgill (1983b, 172, 177-178) report a shared preference by middle class and female participants for standard language forms and by working class and male participants for the vernacular. Eckert also points out an interaction between gender and social class for two groups of Detroit high school students (both occupying different social positions in the school system and coming from different social backgrounds) (2000, 48, 55): while the lower class (oriented) adolescents, and especially the girls, led a general vowel shift, the language behavior of the upper class (oriented) youngsters and the boys appeared to be more conservative (2000, 219). Furthermore, Trudgill reports an association between social class and age in youngsters' language practices, noting that "very high covert prestige is associated with WC [working class] speech forms by the young *of both sexes"* (1983b, 182). Finally, sociological research calls for the inclusion of not only social class, but also "other major dimensions of occupational inequality such as sex, age and race" (Crompton 2010, 159).

In a large corpus (2.9 million tokens) of informal online chat conversations, we study the impact of several aspects of Flemish[2] adolescents' social background on different kinds of deviations from standard formal writing norms. The paper is structured as follows: in Section 2, we describe the operationalization of the independent (sub)variable(s) and discuss sociological and governmental studies on the matter. In Section 3, we discuss the different linguistic variables. Next, in Section 4, we describe the corpus and methodology, and in Section 5, we discuss and evaluate the results.

2. Operationalization of social background

In spite of large-scale social changes in the past decades[3], social class is still an issue today:

> "although it is pointless to attempt to deny, or ignore, this individualistic societal shift […] '[c]lass' still persists as systematically structured social and economic disadvantage, which is reproduced over the generations" (Crompton 2010, 157)

As there is not *one* correct way to define class (Braham 2013, 30; Crompton 2010, 155), we treat it as a multidimensional variable with several subfactors, in order to represent its

[2] I.e. living in Flanders, the Dutch speaking part of Belgium.

[3] In comparison to when literature on social class first emerged, several large-scale social changes have taken place, such as (but not exclusively) a shift towards more individualistic societies (Crompton 2010, 155-157; Goldthorpe & Breen 2007, 25), globalization, an increase in female employment (Crompton 2010, 159), as well as a growing importance of education and knowledge (Goldthorpe & Breen 2007, 45; de Jager, Mok & Sipkema 2009, 243), which will be discussed more elaborately in the next paragraph.

complexity and capture its different potential determinants. In the next paragraphs, we discuss these subfactors: the adolescents' level of education, their home language and the profession of their parents, each representing one or more aspects of social background (cultural, economical, etc.). Our operationalization is based on sociological research and on governmental documents by the Flemish Ministry of Education and Training (from now on *FMET*).

A first important aspect of teenagers' social background is their level of education: it affects their current and future (adult) social networks, and is indicative of their future professional career (de Jager, Mok & Sipkema 2009, 253). As today's society has evolved towards a knowledge-based *meritocracy* – i.e. "social stratification based on personal merit" (Macionis 2011, 206) – education and obtained degrees have become an increasingly important aspect of social status and position (de Jager et al. 2009, 243, 247). In the present case study, we include the three main levels of Belgian secondary education[4] (FMET 2017, 10):

- *General Secondary Education* (in Dutch 'ASO' or 'Algemeen Secundair Onderwijs') is the most theory-oriented type. Students are being prepared for higher education, which most of them indeed pursue after graduating.
- *Technical Secondary Education* (in Dutch 'TSO' or 'Technisch Secundair Onderwijs') is quite practice-oriented but still has a large theoretical side to it. After graduating, students can go to higher education or start working.
- *Vocational Secondary Education* (in Dutch 'BSO' or 'Beroepssecundair Onderwijs') is the most practice-oriented type, preparing students for a specific manual profession. In order to obtain the required degree to get access to higher education institutions, an additional (specialization) year must be taken.

Youngsters tend to spend more years in school than they did a few decades ago (i.e. fewer youngsters drop out of high school before graduating) and go to higher education. This *educational expansion* influences social class patterns but surely does not erase them, as the association between class origin and family background on the one hand and youngsters' chances and levels of attainment in (higher) education on the other continues to exist (Goldthorpe & Breen 2007, 45-46). These social differences do not only affect performance at school, but also decisions within the educational track (e.g. type of education, quitting school before graduating), as these are influenced by limitations and opportunities typically faced in different social classes (Goldthorpe & Breen 2007, 45-47).

The second subfactor we include is the adolescents' home language(s). This is both a cultural and educational factor, as it indicates a potential migration background and the presence or absence of a parent who can easily connect with the (Dutch) school context and support children with school-related communication or tasks. We distinguish the following three categories:

[4] For the types that fall outside the scope of this study (Secondary Education in the Arts and Special Secondary Education), see FMET (2017, 10).

- Dutch only: the teenager only speaks Dutch at home (i.e. the official (education) language in Flanders)
- Dutch and a foreign language: communication at home proceeds both in Dutch and in a foreign language
- Foreign language only: the teenager does not speak Dutch at home

We note that the label 'Dutch' as a home language in reality covers a range of varieties: many adolescents grow up with a regiolectic variant of Dutch rather than with the standard register. However, we did not operationalize 'vernacular' registers as separate home language varieties (although they might, of course, influence adolescents' vernacular writing on social media), since previous research has shown that by far most autochthonous Antwerp adolescents speak more or less the same variety at home, i.e. so-called 'tussentaal', which is a variety in between dialect and standard language: Only 8% of the Antwerp high school students in De Decker and Vandekerckhove (2012) reported to use dialect at home, 9% indicated standard language was their home language and 83% opted for 'tussentaal'. Therefore, we can assume that the school population is quite homogeneous in that respect.

We also note that we categorize every language which is not Dutch indiscriminately as a 'foreign' language. However, several languages may be indicative of different ethnic backgrounds and social class belonging. For instance, while Arabic as a home language is often indicative of quite recent migration, speaking French at home can be (at least in Flanders) indicative of traditional autochthonous upper class belonging. Though we are well aware of the social significance of these differences, they fall outside the scope of this paper. In most cases the foreign language listed by the teenagers actually is not French, but a language which points to a migration background.

The third and final subfactor of adolescents' social background is the profession of the parents. For the classification, we use a threefold division (which is a regrouping of the original seven categories) of the EGP-scheme[5] (Table 1), in which professions are ranked in terms of autonomy, supervision, required level of education or skills, etc. (Erikson, Goldthorpe & Portocarero 1979, 420; Vranken, Van Hootegem, Henderickx & Vanmarcke 2017, 318). We note that we cannot classify certain social positions which fall outside the scope of this scheme, such as unemployed or retired people or house wives/men (Marsh 2000, 291). Finally, whenever the profession of both parents is given, we select the one that ranks highest, since we assume that the highest ranked profession may have a major impact on the general family situation, e.g. in terms of financial resources and consequent spending patterns, leisure activities, consumption of cultural goods etc.

[5] We slightly adapted the scheme by dividing the second class into two subclasses: 2a for professions which require a university degree (e.g. teachers in the highest grade of General Secondary Education), and 2b for professions which require a higher education but not university degree (e.g. nurses).

Final cats.	Class	Label	Description
I	1	Upper service class	Higher-grade professionals, administrators, and officials; managers in large industrial establishments; large proprietors
	2	Lower service class	Lower-grade professionals, administrators, and officials; higher-grade technicians; managers in small industrial establishments; supervisors of non-manual employees
II	3	Routine non-manual workers	Routine non-manual employees in administration and commerce; sales personnel; other rank-and-file service workers
	4	Petty bourgeoisie and farmers	Petty bourgeoisie: small proprietors and artisans, etc., with and without employees Farmers: farmers and smallholders and other self-employed workers in primary production
	5	Supervisors etc.	Lower-grade technicians; supervisors of manual workers
III	6	Skilled manual workers	Skilled manual workers
	7	Semi- and unskilled manual workers	Non-skilled workers: semi- and unskilled manual workers (not in agriculture, etc.) Agricultural laborers: agricultural and other workers in primary production

Table 1: EGP class scheme (Erikson & Goldthorpe 1992, 38-39), with our final categorization added in the leftmost column

The profession of the parents does not only impact on the family's financial situation and social status, but is also a determinant factor in youngsters' choice for particular educational tracks. Vranken et al. discuss several studies showing this correlation (2017, 319-325), which is confirmed by our dataset (see below)[6]. Although in theory, Flemish children can choose any educational track regardless of their social background, in practice, social 'stagnation' or 'immobility' – people staying in the same social class as their parents – is still frequent. Ironically, it is education that holds the power to break this pattern as "people who gain schooling and skills may experience social mobility" (Macionis 2011, 206). Social 'mobility' consists in people ending up in a different social layer than their parents, either lower or higher (resp. 'downward' or 'upward' mobility) (de Jager et al. 2009, 254; Vranken et al. 2017, 314-315, 319). In general, parents want to avoid downward mobility for their children (Goldthorpe & Breen 2007, 53).

[6] Or to be more specific, which is confirmed by *a subset of our data*, as we only have information on the parents' profession for 29% of the participants (400 out of 1384 – cf. Table 3) (while we have information about the educational track of 100% of our informants). This is due to three reasons. First of all, many participants left this field blank when donating chat conversations, either because they did not want to give this information or because they did not know. Second, as mentioned above, some positions fall outside the scope of the EGP-scheme (e.g. 'retired', 'housewife', 'unemployed'). Third, some participants' responses were too vague to classify (e.g. they would fill in 'restaurant', or 'harbor', or the name of a company, without specifying their parent's job).

In our dataset, we find a significant and strong correlation between the participants' level of education and their parents' profession category (chisq. = 99.638, $p < 0.0001$, Cramer's V[7] = 35%). In general, 'upper class' professions (cat. I) correlate with General Education, 'middle class' professions (cat. II) with Technical Education and 'working class' professions (cat. III) with Vocational Education. Table 2 shows the number of participants per combination of the different profession and education categories. Half of the participants 'stagnate' (51.25%): their level of education corresponds to their parents' profession type. A quarter of the participants move down (23.50%) and a quarter move up (25.25%) the social ladder, their level of education likely leading to a 'lower' resp. 'higher' type of profession than their parents'.

		Education child			
		General (ASO)	Technical (TSO)	Vocational (BSO)	
Profession parents	Cat. I	**17.50% (70)**	**4.75% (19)**	**2.50% (10)**	**99**
	Cat. II	**17.50% (70)**	**19.75% (79)**	**16.25% (65)**	**214**
	Cat. III	**2.00% (8)**	**5.75% (23)**	**14.00% (56)**	**87**
		148	**121**	**131**	400

Legend:	Social stagnation	Upward social mobility	Downward social mobility

Table 2: Overview of participants per combination of profession and education category

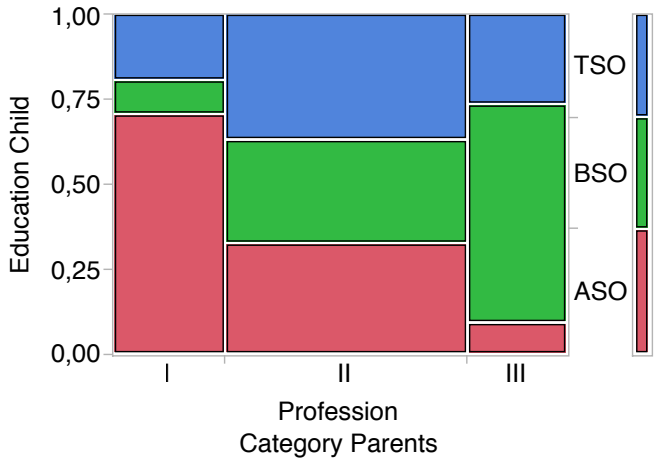


Figure 1: Mosaic plot of the correlation between the adolescents' level of education and their parents' profession

Figure 1 visualizes the distribution for the different education and profession types. The shares of 'extreme' social mobility (cat. I and Vocational Education/BSO, cat. III and General Education/ASO) are smallest. For all education types, stagnation is very frequent, with the

[7] Normalization of the chi-square value for sample size and dimension: square root of the chi-square value divided by the sample size and by the minimum dimension minus one.

profession of most students' parents corresponding to precisely the education type that most probably might lead to the same type of profession. As for the profession categories, stagnation is clearly most frequent for the upper class (cat. I) and working class (cat. III), followed by slight downward or upward mobility respectively. However, for the middle class professions (cat. II), the three possibilities (stagnation, slight upward and slight downward mobility) are more balanced. In other words, children whose parents have a typical middle class profession appear to be the least affected by their social background when it comes to their level of education.

Finally, our data reveal correlations between the participants' home language and their education type on the one hand and the profession of their parents on the other. The correlation between home language and education is significant (though less strong than the one between education and profession of the parents) (chisq. = 23.249, $p < 0.0001$, Cramer's V = 9%; performed for 1346 out of 1384 participants, i.e. the ones for whom we have information on home language) and suggests that it is harder for children from non-Dutch speaking families to get access to more theoretical education systems. The following patterns can be found in our dataset: Adolescents for whom Dutch is the only home language are more likely to attend the theoretical General Education track (ASO) than adolescents with Dutch in combination with a so-called 'foreign' home language or only a foreign home language: 45% of the former category attend ASO compared to 32% (Dutch + foreign) versus 34% (foreign) of the latter type of adolescents. The data for the Vocational education track (BSO) are even more striking: only 26% of the students with Dutch as their only home language attend BSO compared to 46% of the students with a combined Dutch-foreign language profile and 39% of the students with an exclusive foreign language profile. The orientation towards Technical Education (TSO) is comparable for all of the groups: 29% for the 'only-Dutch' students, 22% for the 'Dutch+foreign' students and 27% for the 'foreign language' students.

The home language of the participants does not only correlate with their educational track, it also correlates significantly with their parents' profession (chisq. = 16.138, p = 0.0028, Cramer's V = 14%; performed for 398 out of 1384 participants, i.e. the ones for whom both profession of the parents and home language are known). The following pattern emerges: working class professions seem more common and upper class professions less common amongst the parents who speak a foreign language. Most parents in a Dutch-only home context have a middle class profession (55%), followed by upper class (27%) and working class (18%) professions. In the families where both Dutch and a foreign language are spoken, middle class professions are still the most common category (52%), but working class professions are far more prominent than in the families where Dutch is the only home language (31%) and upper class professions are less well represented (17%). Finally, in the families where only a foreign language is spoken, most parents have a working class profession (50%), followed by middle class (36%) and upper class professions (14%).

3. Operationalization of non-standardness

We examine three deviations from standard (formal) writing norms, each corresponding to a different chatspeak 'maxim'. These maxims (i.e. the underlying principles which explain the particular properties of informal CMC) are "orality, compensation, and economy" (Androutsopoulos 2011, 149). Orality ('write as you speak') refers to the use of colloquial speech, vernacular or other types of non-standard speech (e.g. dialect, regiolect) in written communication. We operationalized this maxim by selecting non-standard Dutch lexemes and words which render non-standard Dutch pronunciation or morphology. E.g.:

(1) original post: *Ja* ***das goe*** *voor* ***effe ma nie*** *constant he*
('Yes that is okay for a while but not all the time')
(1') standard Dutch: *Ja* ***dat is*** *goe****d*** *voor* ***even*** *ma****ar*** *ni****et*** *constant he*

However, since we automatically selected all non-standard lexemes, this category also includes non-standard forms triggered by other factors, e.g. spelling mistakes and words in a language other than Dutch or English. We will address this heterogeneity when discussing the results of the analyses (Section 5.4.2). We note that we treat the integration of English words or phrases in a Dutch chat conversation as a separate phenomenon (falling outside the scope of this paper), although some might argue that it essentially belongs to the orality maxim. However, while quite a lot of English words are very common in Flemish oral adolescent talk (e.g. popular lexemes and phrases as *cool*, *nice, say what?*), the use of many other English terms is still more of an 'online' phenomenon, typical of international chat culture. The occurrence of lexemes in languages other than Dutch or English is generally limited to conversations between participants with a non-Dutch speaking background (e.g. a chat conversation between two teenagers who were (partly) raised in Arabic contains lexemes and phrases in this shared home language). The presence of other languages besides English and Dutch appears to be fairly limited in the present corpus, but this certainly deserves further investigation. At present, all non-Dutch and non-English words have been included in the 'orality' category, since they generally seem to reflect participants' oral communication patterns.

The economy principle ('type as fast as you can'), also called the speed or brevity principle, consists in maximizing typing speed, in order to approach the speed of a face-to-face conversation. We analyze the use of typical chatspeak (i.e. non-standard) abbreviations, which can either be acronyms (in which several words are reduced to a single word, consisting of the first letter of each of the words in the original phrase) or other shortened word forms, as shown in examples (2) and (3).

(2) original post: ***Omg*** *yes*
(2') full version: ***Oh my god*** *yes*
(3) original post: *Ja* ***idd*** ('Yes indeed')
(3') full version: *Ja* ***inderdaad***

The expressive compensation principle ('compensate for the absence of facial expressions, intonation, etc.') leads to a wide variety of expressive strategies. We selected the most prototypical one: the use of emoticons. All possible variations were taken into account (illustrated in examples (4) to (6)): either facial expressions (*smileys*) or other symbols, such as hearts, whether composed by the user with punctuation marks and letters/numbers, or selected as small pictograms from the platform's keyboard interface (*emoji*).

(4) *dammn we look so hot* 😍🔥💋❤
(5) *dat is veel.* ***O_O*** ('that is a lot. O_O')
(6) *haha cutie* ***XD***

4. Experimental setup

In this section, we first discuss the corpus and participants (4.1) and subsequently our methodology (4.2).

4.1. Corpus and participants

The corpus consists of 2 885 084 tokens[8] (488 014 posts) of Flemish teenagers' informal online written language. The number of participants is 1384. Table 3 shows the distribution of the social variables in terms of tokens. Profession of the parents was the hardest information to get: many participants left this field blank or filled in unclear answers which we could not classify (e.g. only a company name, without a job description).

[8] The tokens can be words, but also emoticons or isolated punctuation marks, as they were obtained by splitting the utterances in the corpus on whitespaces.

Variable	Subgroups	Tokens
Education	General Secondary Education (ASO)	920 114 (34%)
	Technical Secondary Education (TSO)	1 213 483 (42%)
	Vocational Secondary Education (BSO)	751 487 (26%)
Home language	Dutch only	2 563 096 (89%)
	Dutch + foreign language	216 558 (8%)
	Foreign language only	93 978 (3%)
	Unknown	11 452 (0%)
Profession of parents	Category I ('upper class')	415 965 (14%)
	Category II ('middle class')	743 952 (26%)
	Category III ('working class')	392 215 (14%)
	Unknown	1 332 952 (46%)
Total		2 885 084

Table 3: Distribution (in terms of tokens) in the corpus for the participants' level of education, home language and profession of their parents

All participants' level of education is known (see Table 3) as well as their gender (66% girls and 34% boys) and their age (55% 'younger' teenagers, aged 13-16, and 45% 'older' teenagers, aged 17-20). In the analyses, we will control for gender and age influences. Almost all tokens (over 96%) are collected from participants living in the same dialect region in the center of Flanders, Antwerp-Brabant, which makes region a (quasi-)constant. The same holds for medium and year: almost all tokens (over 99%) are extracted from instant (i.e. synchronous) messages on Facebook/Messenger, WhatsApp or iMessage, and the vast majority of the tokens (87%) were produced in 2015-2016 (compared to 10% in 2013-2014 and 2% in 2011-2012).

All data were collected in a school context but the conversations delivered by the students were produced outside of school. Students were free to participate and could voluntarily donate chat conversations. We asked for permission of the students and (for minors) their parents to store and analyze their anonymized utterances.

4.2. Methodology

Occurrences of the selected features were automatically extracted from the corpus. We detected emoticons through pattern recognition and abbreviations with predefined lists. For non-standard Dutch, we first checked whether a token was a valid word (and not, for instance, an isolated punctuation mark). For the valid words, a dictionary-based approach was used to check whether they occurred in standard Dutch or English corpora or in a list of named entities. If not, they were classified as non-standard Dutch. We note that word choice was, to some extent, treated independently from other linguistic phenomena. For instance, if a

chatter deliberately repeated a letter within a word for expressive purposes (i.e. *letter flooding*), this did not affect the word choice function. The token *mooooi* (standard Dutch: *mooi*, 'beautiful'), for instance, was classified as *lexical* standard language use, combined with *typographic* non-standardness.

The software's performance was evaluated by comparing the automatically generated output to manual annotations for a test set of 200 randomly selected posts (1257 tokens). Table 4 lists the precision and recall scores per feature. Precision expresses the percentage of detected occurrences of a feature that are indeed valid occurrences of that feature. Recall expresses the percentage of all occurrences of a feature present in the corpus that were detected as such. Here, both measures are (equally) important, as we want our software to be precise in its detections without missing relevant occurrences.

Feature	Precision	Recall
Chatspeak abbreviations	100% = $\frac{19\ detected\ correctly}{19\ detected}$	90% = $\frac{19\ detected\ correctly}{21\ in\ corpus}$
Emoticons	100% = $\frac{51\ detected\ correctly}{51\ detected}$	100% = $\frac{51\ detected\ correctly}{51\ in\ corpus}$
Non-standard Dutch words	95% = $\frac{199\ detected\ correctly}{210\ detected}$	70% = $\frac{199\ detected\ correctly}{285\ in\ corpus}$

Table 4: Evaluation of the software's output per feature in terms of precision and recall

We performed an error analysis on this test set to examine the lower recall score for non-standard Dutch words. Most of the software's mistakes (88.66% or 86 out of 97 errors) 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. For example, the token *me* can either be the standard Dutch pronoun *me* ('me', example (7)) or the written representation of the Flemish non-standard pronunciation of the preposition *met* ('with', example (8)).

(7) *vind je **me** leuk?* ('do you like **me**?')
(8) *ik rij **me** hem mee* ('I'm catching a ride **with** him')

The same mistake can occur for certain typos or spelling errors, if the incorrect form happens to be an existing standard Dutch lexeme. Less frequently, the software incorrectly labeled a token as a non-standard lexeme, i.e. *false positives* (11.34% or 11 out of 97 errors). Many of these misclassified lexemes were very specific named entities, such as the name of a local dance school.

5. Results and discussion

We briefly present the results for the three social variables separately (Sections 5.1 to 5.3). Next, we describe the results for the combined variables in a more detailed way, focusing not only on quantitative tendencies but also on qualitative patterns and possible explanatory factors (Section 5.4).

In general, we report the 'raw' analyses. However, we performed additional analyses to control for age and gender influences by assigning weights to the different subgroups in the data, thus adjusting for possible imbalances in the dataset. The results of these additional analyses are reported where relevant.

5.1. Level of education

Table 5 shows the results per educational track. They reveal a clear distinction between the most theoretical and most practical school system: students in Vocational Education (BSO) use each of the non-standard features much more often than their peers in General Education (ASO). Interestingly, the Technical Education, which occupies an 'intermediate' position on the continuum from theory to practice, does not occupy an intermediate linguistic position but has its own distinct properties. Partial chi-square tests also show the relevance and distinctiveness of all three levels, and the impossibility of further clustering, as the differences between the groups are too salient. When gender and age imbalances are corrected for, the observed patterns are slightly strengthened: the difference between the General and the Vocational System becomes more outspoken and the Technical System stands out even more clearly as a separate group, with the lowest frequencies for all features.

For all three linguistic features, the impact of education is statistically significant, but the correlation strength (calculated as Cramer's V, normalized chi-square value) is very small for chatspeak abbreviations. Stronger correlations can be found for non-standard Dutch words and especially emoticons. When controlling for age and gender influences, the impact of education on the three features remains equally strong or becomes stronger, both in terms of statistical significance and strength of correlation.

The difference in non-standard word choice could be related to the different level of linguistic proficiency that is aimed for in the education types: the more theoretical, the larger the focus on correct standard Dutch writing. However, different attitudes towards vernacular versus standard language might offer an alternative explanation. The difference in emoticon use may tell us something about the (socially determined) expression of emotional involvement in the teenagers' writing. Furthermore, for the chatspeak features (abbreviations and emoticons), there is also the factor of (contemporary) 'prestige': which youngsters perceive which features as 'cool' resp. ridiculous? We will come back to these hypotheses in Section 5.4.

	Tokens	Abbreviations	Emoticons	Non-standard Dutch words
General Secondary Education (ASO)	920 114	1.00%	6.14%	14.04%
Technical Secondary Education (TSO)	1 213 483	1.01%	3.50%	17.75%
Vocational Secondary Education (BSO)	751 487	1.26%	9.05%	17.53%
Significance of correlation (p)		**p < 0.0001**	**p < 0.0001**	**p < 0.0001**
Chisq.		338.353	26 518.16	5 993.251
Strength of correlation (Cramer's V)		1.08%	**9.59%**	**4.56%**

Table 5: Relative counts for all features per level of education, and results chi-square analyses

5.2. Home language

Table 6 shows the results per language category. The use of all three non-standard features gradually increases from the 'Dutch only' to the 'Dutch and foreign language' and finally to the 'foreign language only' category. Even though these gradual differences may suggest that the middle group truly holds an 'intermediate' position and could be clustered with one of the other levels, partial chi-square tests show that all three categories are relevant and that clustering is not possible, as significant differences within the clusters remain.

For all features, the differences in relative frequency between the groups are much smaller and the correlations much weaker than they were between the education levels. Consequently, the linguistic impact of home language appears smaller, though still highly significant. Interestingly, emoticon use is once again affected the most. When controlling for gender and age interference, the same tendencies can be found with the same levels of significance.

As for interpretation, the more frequent use of non-standard Dutch words could indicate a lower proficiency of the standard language. This could be related to the absence of a Dutch speaking parent, as was suggested by the FMET (see Section 2). However, other possible explanations will be discussed in Section 5.4.

	Tokens	Abbreviations	Emoticons	Non-standard Dutch words
Dutch only	2 563 096	1.02%	5.38%	16.30%
Dutch + foreign language	170 689	1.41%	8.91%	17.92%
Foreign language only	139 847	1.61%	9.78%	18.75%
Significance of correlation (p)		**p < 0.0001**	**p < 0.0001**	**p < 0.0001**
Chisq.		633.358	7914.388	816.783
Strength of correlation (Cramer's V)		1.48%	**5.25%**	1.69%

Table 6: Relative counts for all features per language category, and results chi-square analyses

5.3. Profession of the parents

	Tokens	Abbreviations	Emoticons	Non-standard Dutch words
Category I ('upper class' professions)	415 965	0.83%	4.98%	14.89%
Category II ('middle class' professions)	743 952	1.10%	6.36%	16.13%
Category III ('working class' professions)	392 215	1.15%	6.73%	18.34%
Significance of correlation (p)		**p < 0.0001**	**p < 0.0001**	**p < 0.0001**
Chisq.		249.098	1282.129	1817.297
Strength of correlation (Cramer's V)		1.27%	2.87%	3.42%

Table 7: Relative counts for all features per profession category, and results chi-square analyses

Table 7 shows the results per profession category. For all three non-standard features, relative frequencies increase gradually from category I to III. We note that further clustering of this variable (merging the two highest or two lowest levels) is not desirable from a sociological point of view as too much information would be lost, and a threefold class division is generally accepted. Partial chi-square tests also indicate that clustering is not possible as differences within the clusters are just as significant as differences between the clusters and the third remaining group. When controlling for age and gender interference, the same tendencies were observed, with the same levels of significance.

Although profession of the parents has a significant impact on the use of all three features, the correlation strengths are very small. One could conclude that this variable has the smallest linguistic impact (compared to education and language). However, we argue that its *direct* linguistic impact may be small, but that its *indirect* impact is not: in Section 2, we showed that profession of the parents is strongly correlated with the child's educational track.

5.4. Social background (clustered)

Finally, we combine the three subfactors of adolescents' social background (level of education, home language and profession of the parents) and compare two groups with opposite positions on the social spectrum. The first one consists of youngsters with a 'higher' social background: they study General Secondary Education (ASO), they only speak Dutch at home (i.e. the official education language), and their parents have an upper class profession (category I). The second group consists of youngsters with a 'lower' social background who study Vocational Secondary Education (BSO), only speak a foreign language at home, and have parents with a working class profession (category III). In the next two sections, we present the results of the quantitative (5.4.1) and more qualitatively oriented in-depth analysis (5.4.2).

5.4.1. Quantitative analysis

Table 8 shows the results for the two social groups[9]. The relative frequency of all three non-standard features is much higher for the lower class participants than for their higher class peers. These differences are all highly significant, and for emoticons and non-standard Dutch words, the correlations are quite strong too. The effect size (expressed as odds ratio), finally, compares the odds of a feature occurring in the two groups. The odds of an emoticon occurring, for instance, are 2.42 times higher for the lower than for the higher class participants. When controlling for age and gender influences, the same tendencies were observed, with the same levels of significance.

	Tokens	Abbreviations	Emoticons	Non-standard Dutch words
'Higher' social background	217 717	0.78%	4.74%	12.70%
'Lower' social background	30 567	1.82%	10.77%	21.94%
Significance of correlation (p)		**p < 0.0001**	**p < 0.0001**	**p < 0.0001**
Chisq.		324.240	1879.366	1916.853
Strength of correlation (Cramer's V)		3.61%	**8.70%**	**8.79%**
Effect size (odds ratio)		2.37	2.42	1.93

Table 8: Relative counts for all features per social cluster and results chi-square analyses

The lower class adolescents' more frequent use of non-standard Dutch words has multiple possible explanations. It could indicate a lower proficiency in standard writing or in standard Dutch in general, either related to the absence of Dutch in the home context or to lower proficiency levels aimed for at school. Another possible explanation concerns attitudes rather than skills: different linguistic varieties could appeal differently to the two social groups, as suggested in Section 2. Lower class adolescents could simply show a stronger preference for non-standard features than their higher class peers. We will come back to these different hypotheses when discussing the results of the in-depth analysis in Section 5.4.2. The differences concerning the expressive feature of emoticon use could indicate that lower class youngsters' writing is more strongly focused on the expression of emotional involvement. We will investigate this hypothesis in the next section as well. It could also, just like the (small) difference in abbreviation use, be symptomatic of a difference in attitudes towards popular internet culture: the typical chatspeak features could have less (contemporary) prestige (i.e. be perceived as less 'cool', or even as ridiculous) for higher class teenagers.

[9] The rather large difference in number of tokens for the two groups is related to the difference in number of participants. Out of the 1384 original informants, 62 have a distinct higher social background according to our cluster of criteria, but only 8 have a distinct lower social background if we use the same cluster of criteria. These groups really represent the extreme poles of the social class continuum, based on a cluster of variables (see text). However, adding the language restriction in the cluster may have had a too limiting effect, especially on the lower class group. In follow-up research, it might be wise to drop the language criterion from the social class discussion and analyze the effect of the home language on its own. We also note that many students provided insufficient or ambiguous information about their parents' profession, which is why many participants were filtered out for this particular analysis.

5.4.2. Group-bound preferences

Since each of the dependent variables encompasses a range of diverse features, the general quantitative analyses need to be supplemented by a more detailed analysis of the subtypes of features that are favored by the different groups.

For chatspeak abbreviations, similar tendencies can be found among youngsters with different backgrounds. Both lower and higher class adolescents prefer shortened word forms over acronyms (76% - 24% and 73% - 27% resp.). English acronyms, however, are very popular among both groups. Some of the most popular Dutch abbreviations, regardless of the participants' class, are *gwn* (*gewoon*, 'simply/normal') and *idd* (*inderdaad*, 'indeed'). The most popular English acronyms in both groups are *lol* ('laughing out loud'), *wtf* ('what the fuck') and *omg* ('oh my god'). We can conclude that social background has a rather small impact on this brevity-related feature: only small quantitative and qualitative differences emerge. A possible explanation is that brevity is a very pragmatic and functional (rather than expressive or personal) principle in chatspeak, allowing for less personal or socio-demographic variation. This corresponds with the results presented by De Decker and Vandekerckhove (2017), who did not find significant age or gender correlations for the use of acronyms and abbreviations in CMC. They conclude that "[t]hey seem to be the most stable markers of the genre: […] they are not features to show off with, but useful and efficient CMC-tools" (278).

Concerning emoticon use, the two social groups prefer different types. We make a distinction between faces (emoticons representing facial expressions, such as the traditional 'smiley'), hearts (all kinds of hearts as well as faces or lips throwing kisses) and pictograms (all remaining emoji: a party hat, the Facebook 'like'-thumb, a pint of beer, a palm tree, etc.). The higher class adolescents show a very strong preference for the traditional face-emoticons (85.80%). Their share of pictograms and hearts is much smaller (11.60% and 2.60% respectively). While the lower class teenagers also show a preference for faces, it is much less outspoken (only 60%), as they use pictograms and hearts much more frequently than their higher class peers (29% and 11% respectively). These differences can already be observed in the top emoticons per group (decreasing in frequency from left to right):

Top emoticons lower class: 😂 😘 ❤ 😭 🤷 🎉 😍

Top emoticons higher class: 🙂 😉 😛 😄 😂 😐 xp

For the higher class, all top emoticons are traditional smileys, such as the smiling and the winking face. Furthermore, all of them can be manually 'composed' with letters and punctuation marks. In the top list of the lower class, however, fewer faces appear, and many of their favorites cannot be manually composed. Their top list is more varied: it contains faces as well as hearts and pictograms. These observations lead us to adjust our previous hypothesis which was based on the emoticon category in its entirety and which suggested

that the lower class group's writing might be more emotionally expressive. In fact, besides hearts and kisses, which are (as a group) the least popular type among all participants, faces are the most emotionally expressive emoticons (as opposed to pictograms, which mostly represent objects). Consequently, the observed tendencies suggest that higher class youngsters, although using *fewer* emoticons, use them in a more expressive way, i.e. to add emotional content to their text messages. Their lower class peers seem to use them more frequently for creative and playful purposes. In conclusion, this expressive feature appears to be strongly correlated with the participants' social background, both in terms of its overall frequency and in terms of preferences for specific features and their pragmatic functions.

Finally, we examine the youngsters' use of non-standard Dutch words. The most popular lexemes for both groups are the function words listed below:

da	standard Dutch: *dat* ('that')
ni	standard Dutch: *niet* ('not')
ma	standard Dutch: *maar* ('but')
gij	standard Dutch: *jij* ('you')
wa	standard Dutch: *wat* ('what')

While the pronoun *gij* is one of the most prototypical markers of non-standard Flemish Dutch, the other words represent phonological deviations from standard Dutch (in most cases through word final t-deletion which is typical of colloquial Flemish Dutch). However, as mentioned in Section 3, the output for this feature is quite heterogeneous, containing different kinds of deviations from the written standard. We distinguish four important categories. The first one concerns the use of Dutch vernacular words (i.e. regiolect/dialect or colloquial words), like the function words listed above. The second category consists of standard Dutch words containing (deliberate) chatspeak spelling deviations (rather than genuine errors). A typical phenomenon is cluster reduction, like in *egt* (standard Dutch: *echt,* 'real/really'), in which the consonant cluster *ch* (representing the fricative /χ/) is replaced by one grapheme, *g*. Also included are unconventionalized and low frequency shortenings of words, e.g. by deleting all vowels so that only the 'consonantal skeleton' remains[10] (Androutsopoulos 2011, 152; Vandekerckhove, Cuvelier & De Decker 2015, 355), like in *nrml*

[10] Some of these shortened spelling forms have become highly popular and conventionalized abbreviations (detected as such by the software), whereas others are more individual spelling variations, made up on the spot by the chatters. The first ones have been categorized as chatspeak abbreviations, while the latter have been included here, in the category of non-standard lexemes. Although this categorization is partly triggered by methodological issues (i.e. because of the large variation, it is not feasible to detect all abbreviated forms with a predefined list), it is definitely supported by the actual occurrences in the corpus. We can illustrate this with the abbreviated forms *gwn* (for *gewoon*, 'simply/normal') and *nrml* (for *normaal*, 'normal'). *Gwn* is detected as an idiomatic abbreviation with our predefined list. *Nrml* was not in this list and is therefore detected as non-standard Dutch (chatspeak spelling). While these two forms are highly similar (both consonantal skeletons), their frequencies in the corpus reveal an important difference in popularity and status: *gwn* occurs 4774 times (0.17% of all tokens in the corpus) and *nrml* 91 times (or 0.003% of all tokens). Note that this difference cannot just be explained by a higher popularity of the lexeme *gewoon* versus *normaal*, as in their full form, *gewoon* is only 3 times more frequent than *normaal*, but in abbreviated form, *gwn* is no less than 52 times more frequent than *nrml*.

(standard Dutch: *normaal,* 'normal'). The third category consists of standard Dutch or English words containing genuine typing or spelling mistakes, like *vr**ine**d* instead of *vr**ien**d* ('friend' – typing error) or *a**bb**o**n**ement* instead of *a**b**o**nn**ement* ('subscription' – spelling error). A fourth category contains words in a language other than Dutch or English. Finally, some words were labeled 'non-standard Dutch' incorrectly by the software, such as specific named entities that were not recognized as such.

For both groups, the 350 most frequent types were manually annotated and classified into one of the subcategories. Strikingly, different tendencies can be found among youngsters with different backgrounds. When the higher class teenagers use non-standard Dutch words, they primarily opt for 'real' vernacular (67%). They do not frequently make spelling 'errors' (10% of their non-standard words), nor do they often use (deliberate) chatspeak spelling (7%). No foreign language words occurred in their top 350. A different pattern can be found for the lower class adolescents. While they also show a preference for vernacular words, it is less outspoken (40%). They frequently opt for typical chatspeak spelling (27%). The share of typographic and spelling errors is larger (15%) than in the data for the higher class teenagers. Finally, some foreign language words (mostly Arabic) occur (5%). For both groups, the remaining share of the top 350 non-standard Dutch tokens contains words that were either misclassified by the software or that were unclear to the annotator (and for which the automatic classification could thus not be evaluated): 16% (57 out of 350) for the higher class teenagers and 13% (44 out of 350) for their lower class peers.

These results supplement and nuance the general quantitative finding that lower class teenagers use more non-standard Dutch lexemes. Whereas higher class adolescents seem to be attracted more strongly to 'old vernacular' (i.e. traditional non-standard language use, like colloquial or regional speech), their lower class peers show a strong preference not only for old vernacular but for 'new vernacular' as well, such as creative and economic chatspeak spelling. This suggests once again that typical chatspeak features possess more contemporary prestige (i.e. seem 'cooler') for lower class than for higher class adolescents. The larger share of spelling and typographic 'errors' for the lower social group, finally, could suggest a lower proficiency of the written standard or more carelessness regarding orthography.

6. Conclusion

The analyses of the CMC-data produced by Flemish youngsters revealed that three different determinants of adolescents' social class (level of education, home language and profession of the parents) each significantly impact on their non-standard writing practices. When these three subfactors were combined, we got a more distinct representation of the complex and multidimensional phenomenon that is social class. We observed a clear linguistic distinction between the two 'poles' of the social continuum, i.e. 'higher' class teenagers and their 'lower' class peers. The non-standard features were used much more frequently (and significantly so) by the lower class, and correlations were especially strong for emoticon use and non-standard

Dutch words. While the deliberate use of non-standard Dutch clearly was attractive to both lower and higher class teenagers, the more frequent use of non-standard Dutch words and especially the larger share of spelling and typing 'errors' in the CMC-data of the lower class adolescents could be symptomatic of a lower proficiency in the written standard. However, the lower social class adolescents certainly did not demonstrate less chat dexterity or chat linguistic skills, on the contrary: the larger proportion of deliberate chat spelling as well as the more frequent and more creative use of emoticons suggests that typical chatspeak features enjoy higher prestige amongst lower class teenagers than amongst their higher class peers. The latter wrote in a more standardized way, and when they deviated from the standard, they did so in more traditional ways, by rendering vernacular colloquial speech or using traditional (expressive) smileys. In other words, while at first sight the impact of social class seemed unidirectional, with lower social class adolescents producing more non-standard writing, detailed analyses showed more varied and subtle patterns which enforce more nuanced interpretations in terms of skills and the exploitation of the chat repertoire.

In the next phase of our research, we would like to examine the language practices of social groups that fall outside the scope of this study, i.e. teenagers who do not belong in one of the two opposing social clusters (upper class or working class adolescents), but are somewhere 'in between'. It would be interesting to verify if their language use holds an intermediate position as well, or else, if the opposite is true, and their language use is more dynamic and open to change, as lower middle class and upper working class people have often been found to be the trendsetters of linguistic change (Aitchison, 2013, 69). Furthermore, we want to enhance our understanding of the potential explanatory factors (skills versus attitudes) for the observed linguistic differences, and include more linguistic features as dependent variables, in order to improve the representation of (the different aspects of) non-standardness.

Acknowledgments

We thank the two anonymous reviewers for their helpful and constructive feedback on the previous version of this paper.

References

Aitchison, Jean. (2013). *Language change: Progress or decay?* Cambridge: Cambridge University Press.

Androutsopoulos, Jannis. (2011). Language change and digital media: A review of conceptions and evidence. In Tore Kristiansen, & Nikolas Coupland (Eds), *Standard languages and language standards in a changing Europe* (pp. 145-161), Oslo: Novus.

Braham, Peter. (2013). *Key concepts in sociology.* Los Angeles / London / New Delhi / Singapore / Washington DC: SAGE.

Coates, Jennifer. (1993). Quantitative studies. In Jennifer Coates, *Women, men and language*. *A sociolinguistic account of gender differences in language* (pp. 61-86), London / New York: Longman.

Crompton, Rosemary. (2010). The rise, fall and rise of social class. In Anthony Giddens, & Philip W. Sutton (Eds), *Sociology: Introductory readings. 3rd edition* (pp. 154-160), Cambridge / Malden: Polity.

De Decker, Benny, & Reinhild Vandekerckhove. (2012). De mythe van dialectrevival. In Saskia Kindt, Patrick Dendale, & Anne Vanderheyden (Eds), *La langue mise en contexte : essais en l'honneur d'Alex Vanneste* (pp. 27-46), Maastricht: Shaker.

De Decker, Benny, & Reinhild Vandekerckhove. (2017). Global features of online communication in local Flemish: Social and medium-related determinants. *Folia Linguistica* 51(1), 253-281.

de Jager, Hugo, Albert Louis Mok, & G. Sipkema. (2009). *Grondbeginselen der sociologie*. Groningen / Houten: Noordhoff.

Eckert, Penelope. (2000). *Linguistic variation as social practice.* Malden / Oxford: Blackwell.

Erikson, Robert, John H. Goldthorpe, & Lucienne Portocarero. (1979). Intergenerational class mobility in three Western European societies: England, France and Sweden. *The British Journal of Sociology* 30(4), 415-441.

Erikson, Robert, & John H. Goldthorpe. (1992). The constant flux: A study of class mobility in industrial societies. Oxford: Clarendon.

[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* (pp. 8-18), Brussels: Department of Education and Training.

Goldthorpe, John H., & Richard Breen. (2007). Explaining educational differentials. Towards a formal rational action theory. In John H. Goldthorpe (Ed.), *On Sociology. Second edition. Volume two: Illustration and retrospect* (pp. 45-72), Stanford: Stanford University Press.

Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans. (2016). Expressiveness in Flemish online teenage talk: A corpus-based analysis of social and medium-related linguistic variation. In Darja Fišer, & Michael Beißwenger (Eds), *Proceedings of the 4th conference on CMC and social media corpora for the humanities, Ljubljana, Slovenia, 27-28 September 2016* (pp. 30-33), Ljubljana: Znanstvena zalozba Filozofske fakultete.

Holmes, Janet. (1992). *An introduction to sociolinguistics*. London / New York: Longman.

Kucukyilmaz, Tayfun, B. Barla Cambazogly, Cevdet Aykanat, & Fazli Can. (2006). Chat mining for gender prediction. In Tatyana Yakhno, & Erich Neuhold (Eds), *Advances in Information Systems. ADVIS 2006. Lecture Notes in Computer Science* (pp. 274-283), Berlin: Springer.

Macionis, John J. (2011). *Society. The basics.* Upper Saddle River: Pearson Education.

Marsh, Ian (Ed.). (2000). *Sociology. Making sense of society.* Harlow: Prentice Hall.

Nguyen, Dong, A. Seza Doğruöz, Carolyn P. Rosé, & Franciska de Jong. (2016). Computational sociolinguistics: A survey. *Computational Linguistics* 42(3), 537-593.

Parkins, Róisín. (2012). Gender and emotional expressiveness: An analysis of prosodic features in emotional expression. *Griffith working papers in pragmatics and intercultural communication* 5(1), 46-54.

Peersman, Claudia, Walter Daelemans, Reinhild Vandekerckhove, Bram Vandekerckhove, & Leona Van Vaerenbergh. (2016). The effects of age, gender and region on non-standard linguistic variation in online social networks. *Arxiv,* 26 January 2016, retrieved on 20 September 2016, from http://arxiv.org/abs/1601.02431

Trudgill, Peter. (1983a). Social identity and linguistic sex differentiation. Explanations and pseudo-explanations for differences between women's and men's speech. In Peter Trudgill, *On dialect. Social and geographical perspectives* (pp. 161-168), Oxford: Blackwell.

Trudgill, Peter. (1983b). Sex and covert prestige. Linguistic change in the urban dialect of Norwich. In Peter Trudgill, *On dialect. Social and geographical perspectives* (pp. 169-185), Oxford: Blackwell.

Vandekerckhove, Reinhild, Pol Cuvelier, & Benny De Decker. (2015). The integration of English in Flemish versus African online peer group language: A comparative approach. *Language Matters* 46(3), 344-363.

Vranken, Jan, Geert Van Hootegem, Erik Henderickx, & Luc Vanmarcke. (2017). *Het speelveld, de spelregels en de spelers? Handboek sociologie.* Leuven / The Hague: Acco.

Wolf, Alecia. (2000). Emotional expression online: Gender differences in emoticon use. *Cyberpsychology & Behavior* 3(5), 827-833.

CHAPTER 4

This chapter was published as a journal article. Reference:

Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans. (2018). 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.

Social media writing and social class:

A correlational analysis of adolescent CMC and social background

Abstract

In a large social media corpus (2.9 million tokens), we analyze Flemish adolescents' non-standard writing practices and look for correlations with the teenagers' social class. Three different aspects of adolescents' social background are included: educational track, parental profession, and home language. Since the data reveal that these parameters are highly correlated, we combine them into one social class label. The different linguistic practices emerging from the analyses demonstrate the crucial impact of social class on adolescent online writing practices. Furthermore, our results nuance classical findings on working class adherence to 'old vernacular' by also highlighting working class youth's strong connection to the online writing culture, or 'new vernacular'. Finally, we point out the complexity of the social class variable by demonstrating interactions with gender and age, and by examining groups of teenagers whose social background is ambiguous and therefore hard to operationalize.

Keywords: computer-mediated communication, adolescents, social class, social media, non-standard writing

1. Introduction

In previous studies on informal *computer-mediated communication* (CMC), gender and age have been popular independent variables (e.g., Baron 2004; Hilte, Vandekerckhove, & Daelemans 2018b; Wolf 2000). The authors' social class, however, is rarely taken into account. Moreover, certain groups of people systematically tend to be overrepresented in CMC research, as participants are very often middle class people, in most cases middle class youngsters. Consequently, the chat practice of working class teenagers has hardly been studied. Therefore, the present study includes youngsters with this profile and compares their linguistic behavior to that of other social groups.

The study focuses on the correlation between Flemish adolescents' online non-standard writing practices (including typical chatspeak phenomena) and their social background. The paper is structured as follows: First, we discuss the theoretical framework (Section 2) and the methodology (Section 3). Section 4 presents the results of the analyses, and Section 5 is devoted to the discussion of these results and some concluding remarks.

We note that we already conducted a pilot study on this topic (see Hilte, Vandekerckhove, & Daelemans 2018a). The differences with the present study relate to the operationalization of

non-standard writing and the overall methodological focus. The present study includes eight more markers of non-standard writing, the combination of different social subfactors has been optimized, and two new sections were added focusing on methodological challenges related to classification of participants with hybrid social profiles and to interactions between social class, age, and gender.

2. Theoretical framework

In order to obtain a feasible yet accurate operationalization of adolescents' social class, we included criteria from both academic research and Belgian government studies. Taking into account the complexity and multidimensionality of social class, we treat it as a variable consisting of three subvariables (representing different aspects of class, e.g., cultural, financial, and economic): the teenagers' educational track, their home language, and the profession of their parents. Potential correlations between the three social subfactors will be addressed in Section 4.1.2.

For educational track, we include the three main types of Belgian secondary education (Flemish Ministry of Education and Training – from now on FMET, 2017):

- General Secondary Education: theory-oriented educational track that prepares students for higher education.
- Technical Secondary Education: educational track with a strong practical and theoretical (technical) focus. After graduation, students can either enter higher education or start working.
- Vocational Secondary Education: practice-oriented educational track where students are taught a specific (often manual) profession. Students (can) start working right after graduation. This degree excludes direct access to higher education.

Adolescents' educational track strongly impacts their current and future (adult) social networks and future professional career (de Jager, Mok, & Sipkema 2009). As today's western societies have evolved toward meritocracies – i.e., "social stratification based on personal merit" (Macionis 2011, 206) – with a strong emphasis on knowledge and skills, education and obtained degrees have become increasingly important determinants of social status and position (de Jager et al. 2009).

Concerning the participants' home language, it is important to note that Dutch is the only official language in Flanders and the only medium of instruction in Flemish education. For the present study, three home language contexts are distinguished:

- The adolescent only speaks Dutch at home.
- The adolescent speaks Dutch and one (or multiple) other language(s) at home.
- The adolescent does not speak Dutch at home, but one (or multiple) other language(s).

In most cases, the "other" language listed by the teenagers appears to be a language which suggests a recent migration background (e.g. Arabic). Thus home language can be considered an important socio-cultural factor. Furthermore, home language may have an indirect impact on the adolescents' school experience and performance, as it might indicate the presence/absence of a parent who can easily connect with the school context.

The final determinant of minors' social background included in this study is parental profession, as it often has a large impact on the overall family situation (e.g., in financial, economic, and cultural terms). For the classification, we applied the well-known sociological EGP-scheme, which ranks professions based on different criteria, such as degree of autonomy and supervision, and required level of education or skills (Erikson, Goldthorpe, & Portocarero 1979; Vranken, Van Hootegem, Henderickx, & Vanmarcke 2017). The requirement of a university degree was added as an extra criterion for distinguishing between upper and middle class professions, so as to fit the current Flemish social landscape more adequately, and the original seven EGP-categories were regrouped into three clusters:

- *Upper class* professions: Non-manual professions for which a university degree is required (e.g., doctor, civil engineer).
- *Middle class* professions: Professions for which a degree of higher education is required, encompassing both non-manual professions for which a non-university degree is required (e.g., secretary, nurse), and manual work for which specific technical degrees are required (e.g., electrician) and that entails a certain degree of autonomy.
- *Working class* professions: unskilled manual professions (e.g., truck driver, cashier).

Whenever the profession of both parents was known, the one that ranked highest served for classification, since the highest ranked profession may have a major impact on the general family situation, e.g., in terms of financial resources and consumption of cultural goods. Finally, we note that we were unable to classify certain social positions which fall outside the scope of the scheme, such as unemployed people or housewives/-men (Marsh 2000).

In previous research, distinct age and gender patterns were observed in CMC. With respect to gender, women appear to show stronger preferences for expressive markers, such as emoticons (see Section 3.2.1) (e.g., Baron 2004; Hilte et al. 2018b; Parkins 2012; Varnhagen et al. 2010), which corresponds to older sociolinguistic findings on the strong emotionally and socially connective dimension in women's discourse (e.g., Tannen 1990).

Concerning age patterns, previous research showed that adolescents tend to use more stylistic chatspeak features than adult chatters (e.g., Argamon, Koppel, Pennebaker, & Schler 2009; Schwartz et al. 2013). Especially young adolescents appear to favor typical chatspeak features (both expressive markers and unconventional spelling forms) in online interaction (De Decker & Vandekerckhove 2017; Hilte et al. 2018b; Tagliamonte & Denis 2008; Verheijen 2015). These age patterns seem indicative of changing linguistic attitudes as adolescents grow older (Verheijen 2015).

Social class, however, has – to our knowledge – not been operationalized as a linguistic determinant in (adolescent) CMC, and neither have the three social subfactors included in the present study. First of all, parental profession has never been operationalized in CMC research. Educational track and CMC have actually been linked to each other, though from a completely different perspective. Some studies discuss the educational use of CMC (e.g., Heemskerk, Brink, Volman, & Ten Dam 2005; Yates 2001). The same holds for home language. Its impact on CMC writing has not been tested, but there has been research on the application of CMC in foreign language teaching (e.g., Warner 2004) and on the use of English CMC (in a business context) by non-native speakers (Zummo 2018). Furthermore, some studies examine the impact of CMC on students' writing performance in school contexts (e.g., Vandekerckhove & Sandra 2016). The latter study points to educational track as a determining variable. Students in Vocational Education seem to have more trouble avoiding chatspeak interferences in formal school writing than their peers in more theory-oriented educational tracks. Still, the question whether home language or educational track actually influences online writing style remains unanswered.

Although social class has not yet been examined systematically in variationist research on informal CMC, several studies have addressed the visibility of social structures and inequality in the genre. In the early days, digital communication was assumed to be free of inequality, because of the lack of (social) face-to-face cues. However, Yates (2001) concluded that this so-called democratic theory/model of CMC does not hold, because the technology does not "strip away existing social structures" (32), and because "CMC suffers, like all communications media, from the intrusion of existing social relations, including those that are based upon inequalities of access and power" (32-33).

An important non-linguistic class difference that has been addressed in previous CMC research concerns the access to technology and familiarity with digital writing (Heemskerk et al. 2005; Yates 2001). Heemskerk et al. (2005) conclude that the use of ICT-tools might actually "increase inequality in education" (1), because of a "digital divide […] that follows the traditional lines of race and social class" (1-2). This approach falls outside the scope of the present paper, but obviously all teenagers in our corpus have access to the technology and at least some CMC-literacy, since they donated personal CMC-data (see below).

3. Methodology

Below, we discuss the corpus (Section 3.1) and the procedure of the data processing and feature extraction (Section 3.2).

3.1. Corpus

The corpus consists of over 2.8 million tokens (488K posts) produced by 1384 Flemish teenagers in an informal interactive CMC-context. Table 1 shows the distribution of the social

variables. All participants' age, gender, and educational track is known, and for almost all of them, home language could be included too. Parental profession was hard to get access to, as many participants either left this field blank or produced answers which were too vague for classification (e.g., a company name without a job description).

Variable	Variable levels	Tokens	Participants
Education	General Secondary Education	920 114 (32%)	596 (43%)
	Technical Secondary Education	1 213 483 (42%)	395 (29%)
	Vocational Secondary Education	751 487 (26%)	393 (28%)
	Unknown	0 (0%)	0 (0%)
Home language	Dutch only	2 563 096 (89%)	1 154 (83%)
	Dutch + other language	216 558 (8%)	87 (6%)
	Other language only	93 978 (3%)	105 (8%)
	Unknown	11 452 (0.4%)	38 (3%)
Parental profession	'Upper class' professions	415 965 (14%)	99 (7%)
	'Middle class' professions	743 952 (26%)	214 (15%)
	'Working class' professions	392 215 (14%)	87 (6%)
	Unknown	1 332 952 (46%)	984 (71%)
Total		2 885 084	1 384

Table 1: Distributions in the corpus

As the corpus contains an imbalance for gender (66% of the tokens were produced by girls and 34% by boys) and a slight imbalance for age (younger teenagers (aged 13-16): 55%, older teenagers or young adults (aged 17-20): 45%), we will control for gender and age influences in the linguistic analyses. There is no need to control for other factors such as dialect region or medium, as these are highly constant in the corpus. Almost all tokens (96%) were collected from participants living in the same dialect region in the center of Flanders, Antwerp-Brabant, which makes region a (quasi-)constant. The same holds for medium and year. Almost all tokens (99%) were extracted from instant messages on Facebook/Messenger or WhatsApp, and the vast majority of the tokens (87%) were produced in 2015-2016 (compared to 10% in 2013-2014 and 2% in 2011-2012).

The data were collected in a school context, but the conversations delivered by the students were produced outside of school and before the time of collection. The participants were instructed to submit conversations with Dutch as the main language. Entire conversations in a language other than Dutch were excluded from this study, but data with some code switching were not. Students were free to participate and donate their chat conversations. Photos were deleted automatically and all data were anonymized in order to guarantee the privacy of the participants.

3.2. Procedure

3.2.1. The operationalization of non-standard writing

We operationalize adolescents' online non-standard writing as a combination of eleven kinds of deviations from the formal writing standard. These deviations relate to the three "maxims" of informal CMC, i.e. three largely implicit but widely applied rules of linguistic conduct in CMC-contexts: those of orality, brevity (also economy/speed), and expressive compensation (Androutsopoulos 2011; De Decker & Vandekerckhove 2017). Below, we discuss the feature sets, define the underlying principles, and provide examples from the corpus.

The largest set consists of *expressive* features: (mostly typographic) linguistic markers which add or enhance the expression of emotional or social involvement in a chat message. They are related to the chatspeak principle of *expressive compensation*, which implies that all kinds of strategies are used to compensate for the absence of certain expressive cues in online communication, such as volume and facial expressions. Seven types of expressive markers were included in the analysis. The selection of these markers is based on related research (Androutsopoulos 2011; Parkins 2012; Varnhagen et al. 2010; Verheijen 2015; Wolf 2000).

1. Emoticons and emoji: stylized facial expressions and hearts (manually composed with characters or selected as a pictogram from the platform's keyboard interface) and pictograms (representing various objects)

e.g. *zie u graag !* 😍😊☺️😘💋❤️ ("love you!")

2. Allcaps: the capitalization of entire words or sentences to convey a feeling (anger, excitement, etc.), to mimic shouting, or to emphasize a particular word

e.g. *IK BEN ECHT BOOS* ("I AM REALLY ANGRY")

e.g. *Dan zijn we om 1u* ***ZEKER*** *thuis* ("Then we will be home by 1 o'clock **FOR SURE**")

3. Deliberate letter repetition: written representation of the oral phenomenon of lengthening a sound to stress a word

e.g. *Een s**uuuuuuuuuuuu**pergelukkige verjaardag* ("A s**uuuuuuuuuuuu**per happy birthday")

4. Deliberate repetition of question or exclamation marks: to increase their expressive function

e.g. *Ja**!!!*** ("Yes**!!!**")

5. Combinations of question and exclamation marks: often used to convey disbelief or shock

e.g. *Serieus?!?!* ("Seriously**?!?!**")

6. The onomatopoeic rendition of laughter

e.g. *Hahahahahahahaha*

7. Kisses: the rendition of kisses/hugs through combinations of the letters "x" and "xo"

e.g. *Ik spreek u morgen* ***xxx*** ("I will talk to you tomorrow **xxx**")

The second set of deviations from the formal written standard consists of features related to the *write like you speak* principle. This *orality* principle implies that in spite of the written character of the digital medium, the register in informal CMC is often closer to oral than to written communication. We included:

8. non-standard Dutch lexemes: dialect words, slang, or written representations of non-standard phonological phenomena (like the deletion of the final "t" in short function words, as shown in the last example below)
e.g. ***gij*** *hebt niks te vertellen* (std. Dutch "**jij** hebt niks te vertellen", "**you** have got nothing to say")
e.g. *ik ook* ***ni*** (std. Dutch "ik ook **niet**", "me **neither**")

9. English lexemes (in a Dutch conversation)
e.g. *Die zijn echt heel* ***nice*** ("They are really very **nice/cool**")

The inclusion of English in the orality category may at first sight seem surprising. However, the (abundant) use of English marks Flemish adolescent speech, and most of the English lexemes and utterances have not been integrated into standard Dutch (yet).

The third set of features concerns the principle of *brevity* and covers all kinds of strategies to compress words or utterances and thus maximize typing speed and minimize typing effort. They enable chatters to mimic, to a certain extent, the *flow* of a face-to-face conversation. We included the following cluster of features:

10. typical chatspeak abbreviations and acronyms (none of them standard Dutch)
e.g. ***omg*** *das geweldig* (full version: "**oh my god** das geweldig", "**oh my god** that is awesome")
e.g. *ja* ***idd*** (full version: "ja **inderdaad**", "yes **indeed**")

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:

11. Discourse markers: # ("hashtag", to indicate a topic or express a feeling about it) and @ ("at", to address one person directly in a group conversation)
e.g. ***#****verslaafd* ("**#**addicted")
e.g. ***@****robin*

This collection of deviations from the formal written standard consists of both "old vernaculars" and "new vernaculars" (Androutsopoulos 2011, 146), or old and new types of non-standard writing. The typographic expressive features, the prototypical non-standard chatspeak abbreviations and the discourse markers can be considered new vernacular: they cover new ways of deviating from formal written standards that are bound to digital culture (Androutsopoulos 2011). The non-standard Dutch lexemes can be considered traditional vernaculars: they represent "locally bound ways of speaking" (Androutsopoulos 2011, 146), or in this context, regional and slang linguistic variants that have marked colloquial speech for ages. The only feature that cannot be classified in terms of old or new vernaculars unambiguously is the use of English lexemes in Dutch chat conversations. As it generally

reflects offline colloquial speech practices, it resembles some of the old vernacular features. However, the term "old" is largely inappropriate here, since the increasing impact of "global" English is a relatively recent phenomenon. Moreover, some English practices do not reflect adolescent speech but cover specific terms, acronyms, and memes related to international chat culture.

3.2.2. Feature extraction

Occurrences of the features were extracted and counted automatically with Python scripts. For a test set of 200 randomly selected posts (1257 tokens), the software's output was compared to human annotations and judged to be reliable. The average precision score (i.e. the percentage of detected occurrences of a feature that are indeed valid occurrences of that feature) for all eleven features was 0.92. The average recall score (i.e. the percentage of all occurrences of a feature present in the corpus that are detected as such) was 0.88. We note that in the present study, both measures are (equally) important, as we want our software to be precise in its detections without missing relevant occurrences. The average scores as well as the scores for the individual features indicate that the overall feature detection is reliable.

4. Results

This section discusses the impact of (aspects of) adolescents' social class on their online non-standard writing. First, we analyze the correlation between educational track, home language, and parental profession and evaluate their combined impact (Section 4.1). Next, we broaden up the scope on social class by examining adolescents with hybrid social profiles (Section 4.2) as well as possible interactions between social class, age, and gender (Section 4.3).

4.1. The impact of social class on non-standard writing practices

We start with a brief discussion of the individual impact of educational track, home language, and parental profession on adolescents' online writing practices (Section 4.1.1). Next, we show how these social subfactors are actually correlated (Section 4.1.2). Finally, we operationalize social class as a combination of educational track and parental profession and examine their combined linguistic impact (Section 4.1.3).

4.1.1. Individual impact of educational track, home language, and parental profession

Educational track, home language, and parental profession all significantly correlate with the use of non-standard features ($p < 0.0001$ for the three chi-square tests). All of the social patterns remained valid (and equally strong) after correction for age and gender imbalances in the dataset. Students in theory-oriented educational tracks score lower for non-standard features than their peers in practice-oriented tracks, and so do participants with higher class

parents compared to their peers with a lower class family background. Finally, teenagers who only speak Dutch at home produce fewer non-standard markers than their peers with a – combined or exclusive – "other language" profile. Interestingly, the "other language" groups' higher rate of non-standardness does not seem to be related to a more frequent use of other languages (e.g., Arabic) in Dutch chat conversations, but instead appears to indicate a stronger preference for typographic expressive markers (e.g., emoticons).

4.1.2. Correlations between educational track, home language, and parental profession

We start by examining the potential correlation between the teenagers' educational track and the profession of their parents. The analysis is performed on the profiles (and not on the chat conversations, as no linguistic variable is included here) of participants whose parents' profession is known (400 or 29% out of 1384 participants). Information on the educational track is available for all participants. The data reveal a significant and strikingly strong correlation between educational track and parental profession (chisq. = 99.638, $p < 0.0001$, Cramer's $V = 0.35$). The mosaic plot (Figure 1) shows that most youngsters of parents with an upper class profession are in General Secondary Education: a theory-oriented educational track in which students are prepared for higher education, through which they may obtain an upper class profession themselves. The majority of adolescents of parents with a working class profession are in the Vocational system: a practice-oriented education type where a specific (often manual) profession is taught and which generally prepares for a working class career. For children of middle class parents, the three education types are balanced. Their educational track seems much less affected by their social family background.

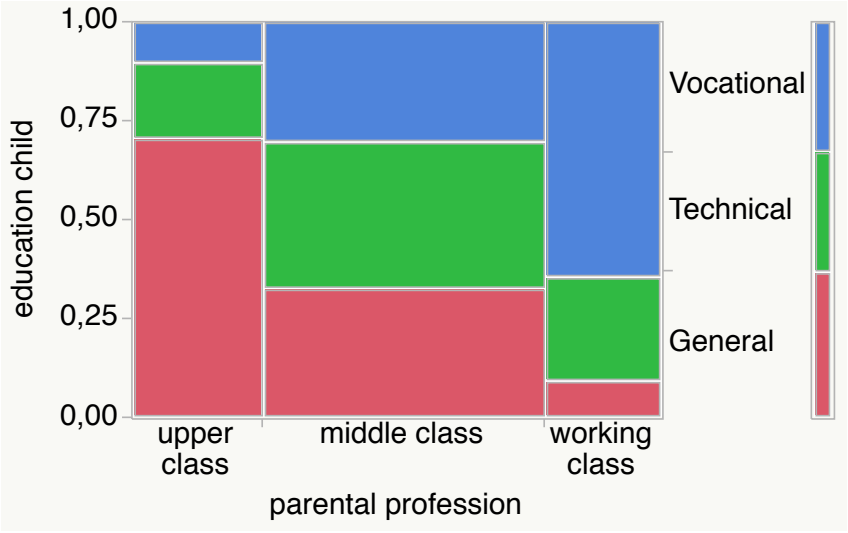


Figure 1: Educational track by parental profession (see also Hilte et al. 2018a)

The correlational analysis between adolescents' educational track and their home language was performed for participants whose home language is known (1346 or 97% out of 1384 participants). A significant but not very strong correlation was found (chisq. = 23.249, $p < 0.0001$, Cramer's $V = 0.09$). The results suggest that it is harder for children from non-Dutch

speaking families to get access to more theoretical education systems (see Figure 2). Adolescents with Dutch as their only home language are more likely to attend the theoretical General Education than adolescents who speak another language at home, as 45% of the former category attend General Education compared to 32% (Dutch + other language) versus 34% (only other language) of the latter group. The data for the Vocational track are even more striking. Only 26% of the students with Dutch as their only home language attend Vocational Education compared to 46% of the students with a combined "Dutch + other language" profile and 39% of the students with an exclusive "other language" profile. The orientation toward Technical Education is comparable for all language groups: 29% of the "Dutch only" teenagers, 22% of the "Dutch + other language" teenagers and 27% of the "other language only" teenagers are students in the Technical track.

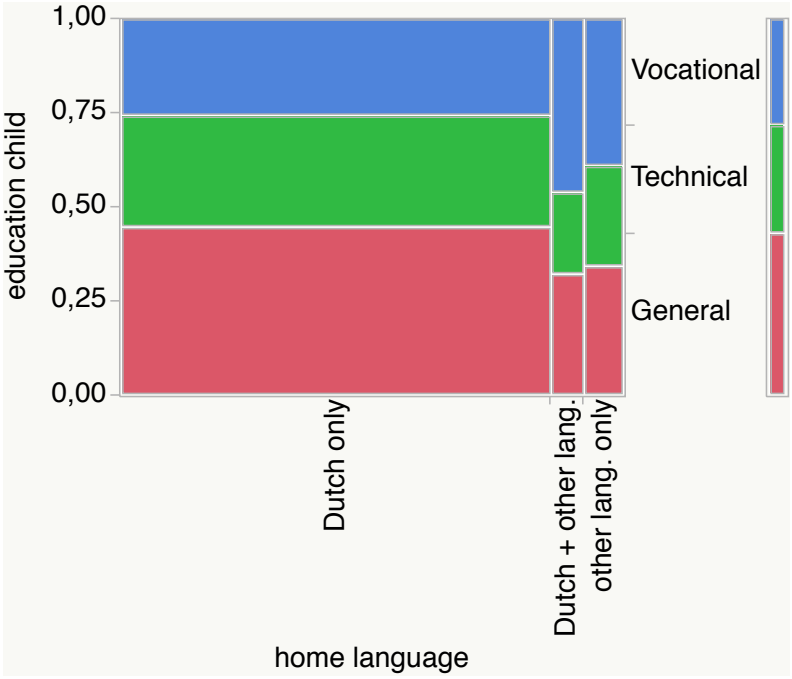


Figure 2: Educational track by home language

The final correlational analysis was performed for participants for whom both parental profession and home language are known (398 or 29% out of 1384 participants). Home language significantly and strongly correlates with parental profession (chisq. = 16.138, p = 0.0028, Cramer's V = 0.14). The following pattern emerges (see Figure 3): working class professions seem more common and upper class professions less common in families in which Dutch is not the only home language or is not a home language at all. Most parents in a "Dutch only" home context have a middle class profession (55%), followed by upper class (27%) and working class (18%) professions. In the families where both Dutch and another language are spoken, middle class professions are still the most common category (52%), but working class professions are far more prominent than in the families where Dutch is the only home language (31%), and upper class professions are less well represented (17%). Finally, in the families where only a language other than Dutch is spoken, half of the parents have a working class profession (50%), followed by middle class (36%), and upper class professions (14%).

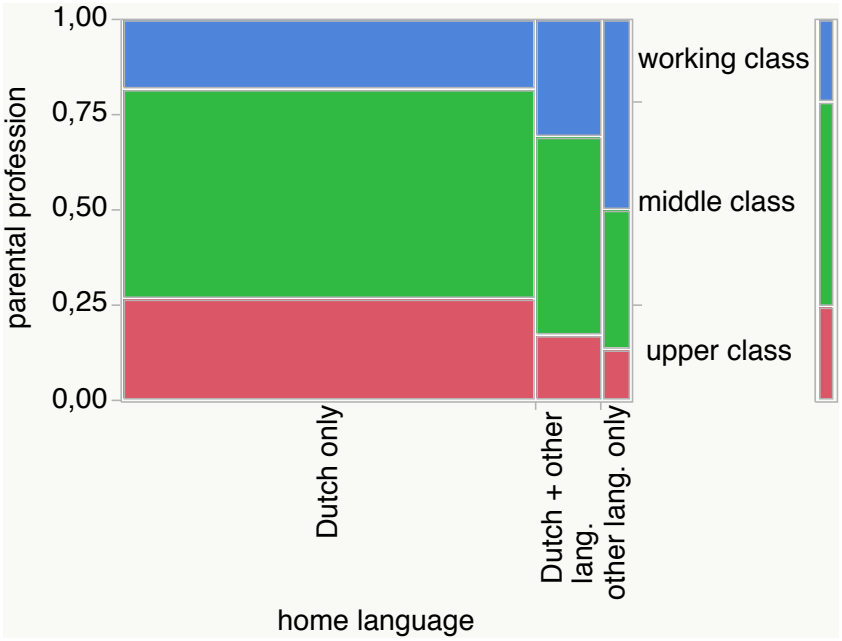


Figure 3: Parental profession by home language

The tendencies visualized in the plots do not only have implications for the processing of the linguistic data (see Section 4.1.3), they clearly have a more general sociological relevance. First of all, Figure 1 shows that both upward and downward social mobility amongst the youngsters is fairly limited (social mobility and status congruence theory will be discussed in Section 4.2). Moreover, while Figure 2 suggests that youngsters with a migration background are relatively overrepresented in the Vocational track, Figure 3 reveals that their parents are overrepresented in working class professions.

4.1.3. Combined linguistic impact of educational track and parental profession

The results of the correlational studies (Section 4.1.2.) suggest that the social subfactors representing different aspects of adolescents' social class should not only be examined in isolation, but also in combination. However, the inclusion of home language in the combined analysis had some undesirable consequences (see below). Therefore, three groups of teenagers were distinguished based on the combination of two of the three socio-cultural criteria discussed above, i.e. educational track and parental profession. They were labeled as upper class, middle class, and working class. The upper class group consists of adolescents in General Secondary Education whose parents have an upper class profession. The middle class group contains teenagers in Technical Education whose parents have a middle class profession. Finally, the working class youngsters are adolescents in Vocational Education whose parents have a working class profession. Table 2 shows an overview of the groups. For two reasons, home language was not included as a criterion for categorization. First, the analyses in the pilot study (Hilte et al. 2018a), in which social clusters were created based on all three social subfactors, suggested that home language was too restrictive as a criterion because the dataset for working class youngsters (operationalized in the pilot study as "other language only" students in Vocational Education, with working class parents) became too

small. As the large majority of participants speak Dutch at home (either exclusively, or combined with another language), only 8 participants met the three criteria for the working class profile. Additionally, although home language is an important socio-cultural and linguistic factor (see Sections 2 and 4.1.1), including it as a criterion implies restricting the analyses to the comparison of the linguistic behavior of "autochthonous" upper class adolescents to that of working class adolescents with a migration background. This implies a questionable simplification of social reality. Obviously many working class families in Flanders are "autochthonous", and needless to say, there are also non-Dutch speaking higher class families, either with or without a recent migration background.

	educational track	parental profession	participants	tokens
'working class' teenagers	Vocational	working class	56	218 676
'middle class' teenagers	Technical	middle class	79	387 363
'upper class' teenagers	General	upper class	70	221 917

Table 2: Three prototypical social groups

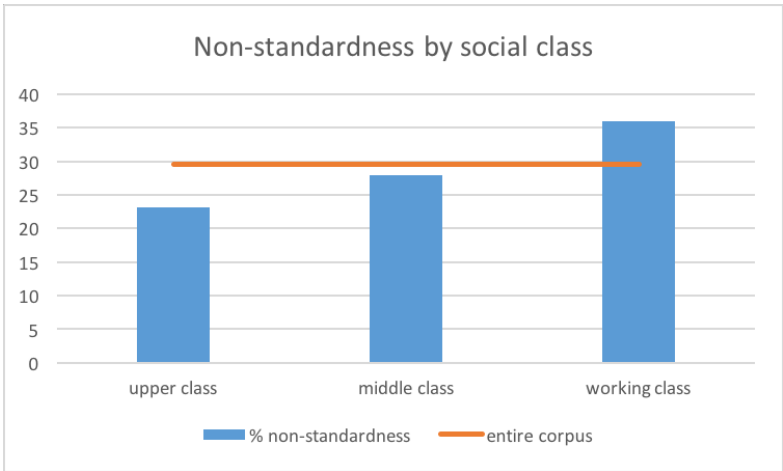


Figure 4: Non-standard writing by social class

Figure 4 shows a gradual pattern for the linguistic variable, with less non-standard writing for adolescents in "higher" social layers. For upper class teenagers, the proportion of non-standard features amounts to 23%, but it rises to 28% and 36% for their middle class and working class peers respectively. The correlation between this construct of social class and non-standardness is statistically significant and also quite strong (chisq. = 9054.840, p < 0.0001, Cramer's V = 0.10, performed on 827956 tokens or 29% out of 2885084). After correcting for age and gender imbalances, the same pattern remains, and the correlation is equally significant and strong.

The differences between the two groups holding extreme positions on the social continuum, i.e. upper and working class youngsters, are very consistent for the different features. Higher frequencies can be found in the working class corpus for eight of the eleven features – for the remaining three (infrequent) features, there are no significant differences. The position of

middle class youngsters is quite variable. They hold a middle position for some features (e.g., repetition of punctuation marks), but for other features they have either the lowest frequency scores (e.g., emoticons) or the highest (e.g., kisses). In other words, when it comes to online language practices, middle class adolescents do not just hold an intermediate position, they have a distinct sociolinguistic profile.

All three social groups deviate from formal writing practices mainly for the sake of orality and expressiveness. Interestingly, the distributions in terms of types of markers also show a gradual difference. The middle class teenagers are strongly oriented toward orality (68% of their non-standard markers are oral features), and much less toward expressiveness (28%). Upper class teenagers show a similar – but less outspoken – preference pattern, with 60% oral features versus 37% expressive markers. For working class adolescents, however, the distribution between expressive and oral features is much more balanced: 53% of their non-standard features serve the purpose of orality, and 44% are used for expressive purposes. In all three groups, chatspeak abbreviations and acronyms score much lower than the other sets of features. They represent 3 to 4% of all non-standard markers.

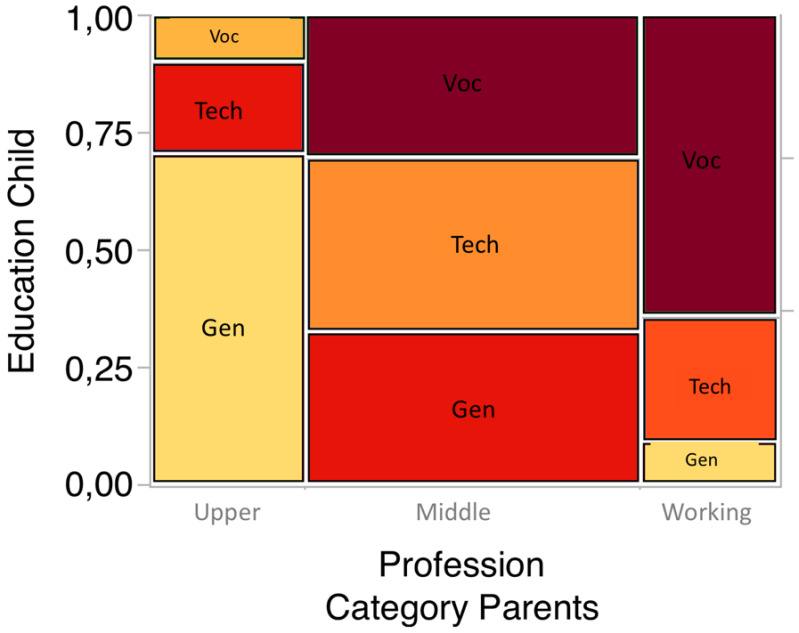
As working class youngsters use both expressive and oral features significantly more often than their upper class peers, we can conclude that they seem to be attracted more to both "old" vernacular (e.g., dialect words) and "new" vernacular (e.g., typographic chatspeak features such as emoticons). We note that the more frequent use of oral features and of non-standard Dutch lexemes in particular might also point to a lower proficiency in formal written standard Dutch and/or more carelessness toward standard language norms, which in turn might both be related to a minor focus on standard Dutch proficiency and a stronger focus on skills in practice-oriented education types. The more frequent use of expressive markers, finally, suggests more (typographically) expressive writing by these youngsters.

For brevity-related features, we found no differences between the different groups when dealing with the variables educational track, parental profession, and home language individually and this holds for the combined social profiles. De Decker and Vandekerckhove (2017) already signaled that no gender and hardly any age differences could be attested for the use of acronyms and abbreviations in Flemish CMC, and concluded that these features are the most stable markers of the genre. So, apparently, these features are so useful and functional that they are appreciated by all groups to more or less the same extent.

4.2. Non-prototypical social profiles

The operationalization of adolescents' social class presented in Table 2 leads to three prototypical social groups which we labeled as working class, middle class, and upper class. However, many participants do not fit into one of these categories, but have a more "hybrid" social profile: e.g., teenagers in General Secondary Education whose parents are unskilled manual workers (i.e. working class profession). The online language use of these participants with a hybrid social profile will be examined in this section.

In order to visualize the linguistic behavior for all potential combinations of educational track and parental profession, we adapted the mosaic plot from Figure 1. In Figure 5, the color of the blocks reflects the relative proportion of non-standard features: dark blocks represent higher frequency scores than the paler ones. In every group or block, the participants' profiles in terms of age and gender were checked, and none of the groups were too skewed. Nevertheless, these results should be interpreted with caution, as some of the smaller blocks consist of few participants. In the bottom left and upper right corners are two of the prototypical groups from Table 2, holding extreme positions on the social continuum. These two groups are youngsters from the upper class and the working class. These two groups' significantly diverging frequency scores for non-standard markers (discussed in the previous section) are now visualized in Figure 5 by extreme color contrasts. The middle block represents the typical middle class youngsters: the orange color shows that their overall frequency score for non-standardness is somewhere in between that of their upper class (pale yellow) and working class (dark red / maroon) peers. The remaining blocks represent youngsters with "hybrid" social profiles. The groups in the upper left and bottom right corner seem to be strikingly deviant concerning their use of non-standard features, as their color stands out. The block in the upper left corner represents adolescents in Vocational Education whose parents have an upper class profession. The pale orange color indicates a relatively low frequency score for non-standard markers. In other words, their language use is fairly standard-oriented. Interestingly, it is more similar to the linguistic profile of their peers with a similar (upper class) family background than to that of their peers in Vocational Education. The opposite pattern can be found for the group in the bottom right corner, which represents adolescents in General Education whose parents have a manual working class profession. The pale yellow reveals that these youngsters produce a relatively small amount of non-standard markers, just like their peers from the same (General) education system and unlike their peers with a similar (working class) family background. Interestingly, the linguistic behavior of these two groups reveals a stronger orientation toward standard writing norms than that of the hybrid groups of Technical students with an upper class family background and General students with a middle class family background. This might point to a tendency of sociolinguistic hypercorrection (see below) amongst youngsters with a strong clash between social family background and educational track.



Legend: relative number of non-standard features

Figure 5: Visualization of non-standardness for different groups of adolescents

The, in some respects, 'deviant' linguistic practices of particular hybrid groups suggest that some determining factors are still missing in the current operationalization of minors' social class. The operationalization might be optimized by including attitudinal factors, such as social ambition: Do the youngsters aspire upward social mobility or not? We interpret the adolescents' social mobility in terms of educational track (assuming this is a reliable predictor for their future professional career) and the professions of their parents. In sociological literature, this type of mobility is called *intergenerational* mobility, as it concerns changes in profession type/class between multiple generations (Vranken et al. 2017). Figure 1, which visualized the number of participants per combination of the different profession and education categories, shows that half of the participants "stagnate" (i.e. no social mobility) (51%): their educational track corresponds to their parents' profession type. A quarter of the participants move "down" (24%) and a quarter move "up" (25%) the social ladder, since their level of education is likely to lead to a "lower" versus "higher" profession type than that of their parents. We note that these percentages largely correspond to the proportions reported by Vranken et al. (2017) for father-son intergenerational social mobility in the Netherlands in the 1970s. They report 54% immobility versus 26% upwards and 20% downwards mobility. (Follow-up studies showed a decrease in social immobility in the Netherlands to 45% in the early 2000's, versus an increase in upwards mobility to 35% and a stagnation of downward mobility, 20%).

In our data, stagnation is clearly most frequent for upper class and working class professions (followed by slight downward or upward mobility respectively), whereas for the middle class professions, the three possibilities (stagnation, upward and downward mobility) are more balanced. The tendencies with respect to social stagnation can be explained by the sociological status congruence theory. *Status congruence* implies that different components of one's social status are "congruent" or reconcilable, whereas *status incongruence* indicates an imbalance between these components (Vranken et al. 2017). The theory states that status congruence facilitates social interaction and is therefore generally positively reinforced (Vranken et al. 2017). This theory offers a frame for the finding that parents tend to send their children to an education type corresponding to their own status. It predicts that a lower class background counteracts upward social mobility, while a higher class background counteracts downward mobility. Vranken et al. (2017) therefore conclude that the larger the potential status incongruence, the more mobility will be impeded.

The two groups in the upper left and bottom right corner of Figure 5, whose online language use is most deviant, represent "extreme" social mobility (i.e. they experience the strongest incongruence between family background and future professional career). We see "downward" social mobility for the students in the Vocational track with upper class parents and "upward" social mobility for the students in the General track with working class parents. This type of extreme social mobility appears to be highly infrequent, which seems to confirm the status congruence theory.

Social mobility might affect the teenagers' language use, making it more dynamic and open to change. While Aitchison (2013) states that lower middle class and upper working class people (i.e. people on the 'boundaries' between different social groups) often act as the trendsetters of linguistic change, Labov (1966) already found that the unclear and insecure position of the lower middle class and its aspirations for upward social mobility favor sociolinguistic hypercorrect behavior (see also Labov 2006). Thus, the dynamic social position of these teenagers might explain the less predictable patterns of non-standard writing practices in their data.

4.3. Interactions between social class, age, and gender

We focused on how the social class parameters interact, but we did not yet discuss possible interactions between adolescents' social class and other aspects of their socio-demographic profile, such as their age and gender. These interactions will be examined in this section.

For all linguistic analyses described in this paper, additional 'weighted' tests were carried out to correct for possible age and gender imbalances, since both age and gender have proven to impact adolescents' online writing (e.g., Baron 2004; De Decker & Vandekerckhove 2017; Hilte, Vandekerckhove, & Daelemans 2017; Hilte et al. 2018b; Schwartz et al. 2013; Verheijen 2015). Moreover, the analyses of the CMC-data for the present case study reveal that age and gender actually interact with social class. In other words, social class does not have the same

impact on the online writing practices of boys versus girls, or on those of younger adolescents (aged 13-16) versus older adolescents/young adults (aged 17-20).

The three-way interaction between gender, age, and social class is visualized in Figures 6a, 6b, and 6c. Each figure shows the 'age*gender'-interaction for one of the three social groups (upper class, middle class, and working class youngsters). In all three plots, the relative number of non-standard features is shown on the y-axis (i.e. the absolute number of features divided by all tokens). The two age categories are shown on the x-axis, and the gender groups are represented by the orange solid lines (girls) and blue dashed lines (boys). Strikingly, different 'age*gender'-patterns emerge depending on the adolescents' social class.

For upper class teenagers, a clear interaction can be observed (see the cross pattern in Figure 6a). Age has a different effect on the language use of upper class girls versus boys. Whereas boys tend to use marginally more non-standard markers as they grow older, girls do not, on the contrary: non-standard features decrease as they age. In related research, girls were found to converge more strongly toward the adult standard as they grow older than boys (see Eisikovits 2006 for adolescents with a working class family background). Eisikovits (2006) ascribed these distinct age patterns to a difference between (working class) boys' and girls' attitudes toward society when they graduate from high school; while accepting the responsibilities of adulthood, girls converge toward mainstream societal norms, whereas boys insist on their autonomy more strongly.

Interestingly, and contrary to Eisikovits' (2006) findings for working class teenagers, we can only find this pattern for the upper class participants. However, the study of Eisikovits (2006) is not perfectly comparable to ours, since she studied spoken language and focused on 'old vernacular'. For middle class adolescents (Figure 6b), no real interaction seems to emerge between age and gender. Although the figure suggests a marginal increase for boys and a marginal decrease for girls, the difference between both gender groups essentially stagnates as they grow older.

For working class adolescents, however, Figure 6c does reveal an interaction, but the pattern strongly deviates from that of the upper class group. While girls more or less stagnate, boys clearly use more non-standard markers as they grow older. Strikingly, the girls' frequency scores for non-standard markers consistently exceed those of the boys. Once again, it should be noted that this need not be due to a stronger preference of old vernacular, since the non-standard features include a wide range of typographic expressive markers and girls tend to use these (much) more frequently than boys do (see Baron 2004; Hilte et al. 2018b; Parkins 2012; Varnhagen et al. 2010).

Finally, the three plots indicate that gender differences are most outspoken (in both age categories) for working class adolescents. In the middle class group, gender and especially age differences are very small, whereas in the upper class group gender differences are small in early adolescence, but increase toward late adolescence. Summarizing, different patterns of age and gender dynamics emerge depending on the adolescents' social background.

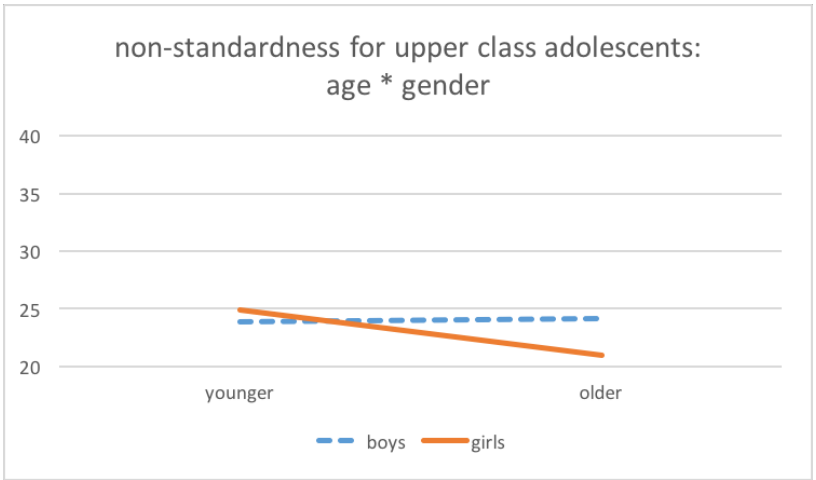


Figure 6a: 'Age*gender'-interaction for upper class teenagers

Figure 6b: 'Age*gender'-interaction for middle class teenagers

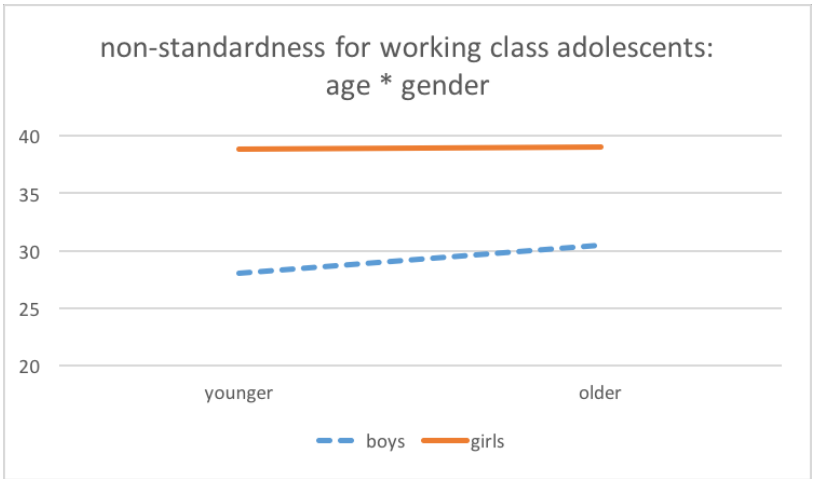


Figure 6c: 'Age*gender'-interaction for working class teenagers

5. Discussion

The present study was devoted to the impact of Flemish adolescents' social class on their informal online writing practices. More specifically, it focused on the occurrence of both old (i.e. traditional regional and slang) and new (i.e. bound to the digital writing culture) vernacular features which generally are no part of formal standard writing and therefore were clustered into a general non-standardness index. The adolescents' social class was operationalized in terms of educational track, home language, and parental profession.

While each of these variables had a significant impact on non-standard writing practices, it was demonstrated that they were correlated rather than independent. For educational track and parents professions, this correlation corroborates previous sociological findings. Therefore, these two factors were clustered so as to create more prototypical social class groups: a working class, middle class, and upper class group. This "clustered" approach revealed more distinct sociolinguistic patterns. Especially upper class and working class youngsters appeared to diverge strongly, with the working class youngsters using much more non-standard markers. The language use of middle class youngsters held an intermediate position when all non-standard features were clustered, but showed a more varied pattern for the individual non-standard markers.

While the distinct online linguistic behavior of the upper versus working class adolescents may at first sight seem to corroborate classic sociolinguistic findings, the distinction between old and new vernacular features actually changes the perspective to some extent. Ever since Labov (1972), working class people, and especially working class men, have been found to be attracted to the toughness of vernacular speech. The same holds for youngsters (see e.g., Eisikovits 2006; Trudgill 1983, and many more). The informal CMC-context offers adolescents a medium for the integration of oral vernacular features in writing and apparently they eagerly exploit this opportunity. In view of previous findings, it is hardly surprising that we attest significantly more of this old vernacular in the CMC-data of working class youngsters. However, they also score much higher for new vernacular features, e.g., they use much more typographic expressive markers that are typical of informal CMC. So these working class adolescents strongly connect to the digital culture too and demonstrate a high chat linguistic dexterity (see Deumert & Lexander 2013). In other words, by including several sets of features in the category of non-standard markers, it could be demonstrated that working class youngsters certainly do not exclusively exploit classic ways of divergence from standard language norms.

For all groups, the oral vernacular features and the expressive markers largely outnumbered the brevity-related features that are also typical of informal CMC. Interestingly, no social correlations could be found for the latter. This confirms that these features have become stable markers of the genre. Said features are so functional for all social groups that hardly any social variation emerges (see De Decker & Vandekerckhove 2017).

An unwanted side-effect of clustering social variables was that more hybrid social groups could no longer be incorporated in the research design. Therefore, we examined the language use of adolescents with non-prototypical social profiles. The, in some respects, "deviant" linguistic behavior of certain groups suggested that more subtle social factors such as aspirations toward social mobility should definitely be included in the operationalization of class in future research on adolescents' (online or offline) linguistic practices. Furthermore, the operationalization of social class also benefits from including age and gender information, as social class background appeared to interact with both gender and age: different age and gender dynamics were found depending on the youngsters' social background.

To our knowledge, social class has not been operationalized systematically in variationist sociolinguistic research on youngsters' informal CMC – and neither have the different aspects of class included in this study. The present paper illustrates both the relevance of the social class variable for this type of CMC-research and the challenges related to the operationalization of such a complex and multidimensional concept which includes several aspects of people's socio-demographic profile and even of their personality, if we take into account social ambition.

Acknowledgments

This work was supported by the FWO (Research Foundation Flanders) under grant G041115N. We thank the anonymous reviewers for their pertinent feedback.

References

Aitchison, Jean. (2013). *Language change: Progress or decay?* Cambridge: Cambridge University Press.

Androutsopoulos, Jannis. (2011). Language change and digital media: A review of conceptions and evidence. In Tore Kristiansen, & Nikolas Coupland (Eds), *Standard languages and language standards in a changing Europe* (pp. 145-161), Oslo: Novus.

Argamon, Shlomo, Moshe Koppel, James W. Pennebaker, & Jonathan Schler. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM. Inspiring Women in Computing* 52(2), 119-123.

Baron, Naomi S. (2004). See you online: Gender issues in college student use of instant messaging. *Journal of Language and Social Psychology* 23(4), 397-423.

De Decker, Benny, & Reinhild Vandekerckhove. (2017). Global features of online communication in local Flemish: Social and medium-related determinants. *Folia Linguistica* 51(1), 253-281.

de Jager, Hugo, Albert Louis Mok, & G. Sipkema. (2009). *Grondbeginselen der sociologie*. Groningen / Houten: Noordhoff.

Deumert, Ana, & Kristin Vold Lexander. (2013). Texting Africa: Writing as performance. *Journal of Sociolinguistics* 17(4), 522-546.

Eisikovits, Edina. (2006). Girl-talk/boy-talk: Sex differences in adolescent speech. In Jennifer Coates (Ed.), *Language and gender: A reader* (pp. 42-54), Oxford: Blackwell.

Erikson, Robert, John H. Goldthorpe, & Lucienne Portocarero. (1979). Intergenerational class mobility in three Western European societies: England, France and Sweden. *The British Journal of Sociology* 30(4), 415-441.

[FMET] Flemish Ministry of Education and Training. (2017). *Statistisch jaarboek van het Vlaams onderwijs. Schooljaar 2015-2016.* Brussels: Department of Education and Training.

Heemskerk, Irma, Anouk Brink, Monique Volman, & Geert Ten Dam. (2005). Inclusiveness and ICT in education: A focus on gender, ethnicity and social class. *Journal of Computer Assisted Learning* 21(1), 1-16.

Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans. (2017). Modeling non-standard language use in adolescents' CMC: The impact and interaction of age, gender and education. In Egon W. Stemle, & Ciara R. Wigham (Eds), *Proceedings of the 5th conference on CMC and social media corpora for the humanities* (pp. 11-15), Bolzano.

Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans. (2018a). Adolescents' social background and non-standard writing in online communication. *Dutch Journal of Applied Linguistics* 7(1), 2-25.

Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans. (2018b). Expressive markers in online teenage talk: A correlational analysis. *Nederlandse Taalkunde* 23(3), 293-323.

Labov, William. (1966). Hypercorrection by the lower middle class as a factor in linguistic change. In William Bright (Ed.), *Sociolinguistics* (pp. 84-113), The Hague: Mouton.

Labov, William. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Labov, William. (2006). *The social stratification of English in New York City.* New York: Cambridge University Press.

Macionis, John J. (2011). *Society: The basics.* Upper Saddle River: Pearson Education.

Marsh, Ian. (Ed.). (2000). *Sociology: Making sense of society.* Harlow: Prentice Hall.

Parkins, Róisín. (2012). Gender and emotional expressiveness: An analysis of prosodic features in emotional expression. *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. *PLoS ONE* 8(9), e73791.

Tagliamonte, Sali A., & Denis, Derek. (2008). Linguistic ruin? LOL! Instant messaging and teen language. *American Speech* 83(1), 3-34.

Tannen, Deborah. (1990). *You just don't understand. Women and men in conversation.* New York: Ballantine Books.

Trudgill, Peter. (1983). *On dialect. Social and geographical perspectives.* Oxford: Blackwell.

Vandekerckhove, Reinhild, & Dominiek Sandra. (2016). De potentiële impact van informele online communicatie op de spellingpraktijk van Vlaamse tieners in schoolcontext. *Tijdschrift voor Taalbeheersing* 38(3), 201-234.

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. *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 Verónica González Temer, Jelena Horvatic, David O'Reilly, & Aiqing Wang (Eds), *Proceedings of the second postgraduate and academic researchers in linguistics at York (PARLAY 2014) conference* (pp. 127-142).

Vranken, Jan, Geert Van Hootegem, Erik Henderickx, & Luc Vanmarcke. (2017). *Het speelveld, de spelregels en de spelers? Handboek sociologie.* Leuven / The Hague: Acco.

Warner, Chantelle N. (2004). It's just a game, right? Types of play in foreign language CMC. *Language Learning & Technology* 8(2), 69-87.

Wolf, Alecia. (2000). Emotional expression online: Gender differences in emoticon use. *Cyberpsychology & Behavior* 3(5), 827-833.

Yates, Simeon J. (2001). Gender, language and CMC for education. *Learning and Instruction* 11(1), 21-34.

Zummo, Marianna Lyan. (2018). The effect of CMC in business emails in lingua franca: Discourse features and misunderstandings. *International Journal of Society, Culture & Language* 6(1), 47-59.

CHAPTER 5

This chapter has been accepted to be published as a journal article, with minor revisions. The revised version is included in the dissertation. Reference:

Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans. (forthcoming). Modeling adolescents' online writing practices: The sociolectometry of non-standard writing on social media. *Zeitschrift für Dialektologie und Linguistik*.

Modeling adolescents' online writing practices:

The sociolectometry of non-standard writing on social media

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 which 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'.

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 e.g. Androutsopoulos, i.e. the principles of expressive compensation, orality and brevity (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, e.g. 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 e.g. 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 (e.g. 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, e.g. 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 e.g. 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 et al. 2016 note that digitalization (including the emergence of online communication) has led to a "new social and linguistic reality" (2016, 143) in which language norms are pluralized (130) and new types of linguistic superiority criteria have become increasingly important, such as "dynamism", "media cool" or "modern media prestige" (132; see also Kristiansen et al. 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 communication, 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 1384 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

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

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 (e.g. operationalizing gender as a continuum[2]) was infeasible with the profile information 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.

[2] See e.g. Bamman, Eisenstein and Schnoebelen (2014), who (linguistically) approach gender as consisting of multiple gender-oriented (language) clusters, and Killermann (2014) for a conceptualization 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 (1384), 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; e.g. 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 (e.g. 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:
e.g. *zie u graag !* 😍😊☺️😘💋❤️ ('love you !')

2. Allcaps, i.e. entire words or utterances in capital letters:
e.g. *DIT MAAKT MIJ KWAAD* ('THIS MAKES ME ANGRY')

3. Deliberate letter repetition (letter 'flooding'):
e.g. *W**oooooo**w goed gedaan* ('W**oooooo**w good job')

4. Deliberate repetition of punctuation marks (punctuation 'flooding'):
e.g. *Proficiat**!!!!!!*** ('Congratulations**!!!!!!**')

5. Combinations of question and exclamation marks:
e.g. *Wat**?!*** ('What**?!**')

6. The onomatopoeic rendering of laughter:
e.g. *Hahahahah*

7. The typographic rendering of kisses and/or hugs through combinations of the letters 'x' and 'xo':
e.g. *Dankje **xxx*** ('Thanks **xxx**')
e.g. *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 (e.g. regional varieties) or informal speech:

8. Non-standard Dutch lexemes (i.e. dialect, regiolect, colloquial or slang lexemes, or representations of non-standard pronunciation):
e.g. *ik was **efkes** in de war* (std. Dutch 'ik was **even** in de war', 'I was confused **for a while**')
e.g. ***gij** ook* (std. Dutch '**jij** ook', '**you** too')
e.g. ***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:
e.g. *echt **awesome*** ('really **awesome**')

We note that each token in the corpus is classified as either a non-word element (e.g. 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 (e.g. *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 and Vandekerckhove 2012, 2013 and Verheijen, de Weger and 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)
> e.g. ***omg*** *hahaha* (full version: '**oh my god** hahaha')
> e.g. ***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)
> e.g. ***#****crisis*
> e.g. ***@****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 phrases in Flemish

[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' (@).

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 (e.g. 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 et al. 2016, 132) and 'dynamism' (Grondelaers et al. 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 et al. 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 (1257 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, e.g. 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.

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

[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 (e.g. 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 (e.g. a frequency cutoff was applied on the name lists) and updated (e.g. 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.

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 (e.g. 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, Maechler, Bolker & Walker 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 et al. explain that the Poisson distribution is the "simplest distribution for modeling count data" (2008, 5) – for the mathematical background on why this distribution can adequately capture the properties of count data, we refer to Coxe, west and aiken (2009, 123). However, a common problem with 'naive' 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 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).

[7] For different causes of overdispersion, see Harrison (2014, 2) and Tarpey (2012, 23).

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). 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

[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 and Vandekerckhove (2018). Simultaneous inspection of the different dependent variables (i.e. expressive, oral, and brevity-related non-standard markers) and their potential correlations, e.g. 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.

[10] Coxe et al. 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" (2009, 134). 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 et al. 2009, 134).

depend on sample size (Tarpey 2012, 24-25). 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.

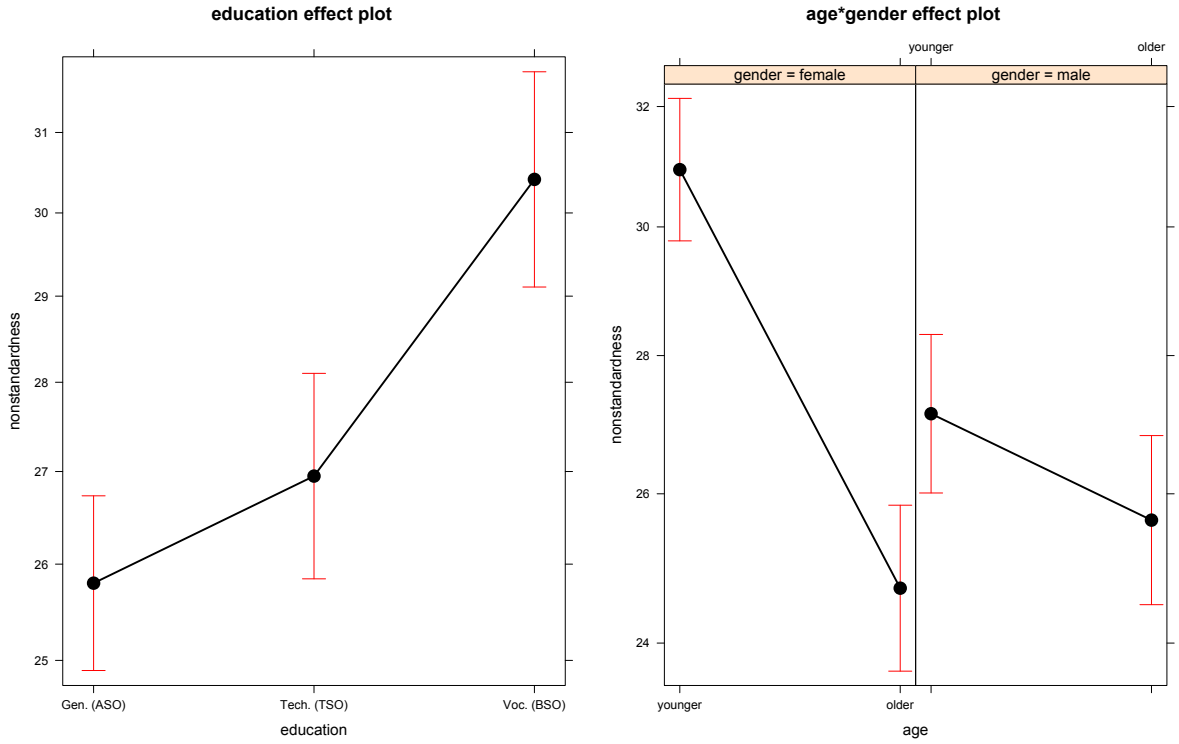


Figure 1: Non-standardness model: Effect plot (predicted counts of non-standard features per 100 tokens)

	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.

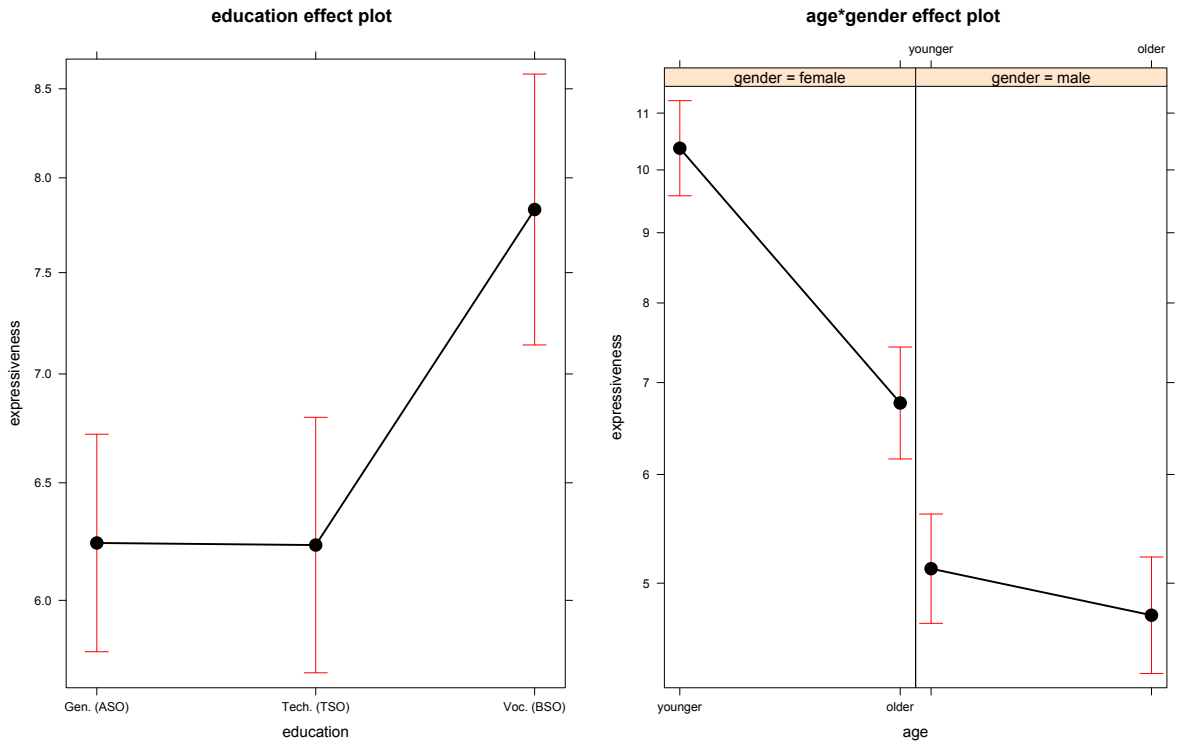


Figure 2: Expressiveness model: Effect plot (predicted counts per 100 tokens)

	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.

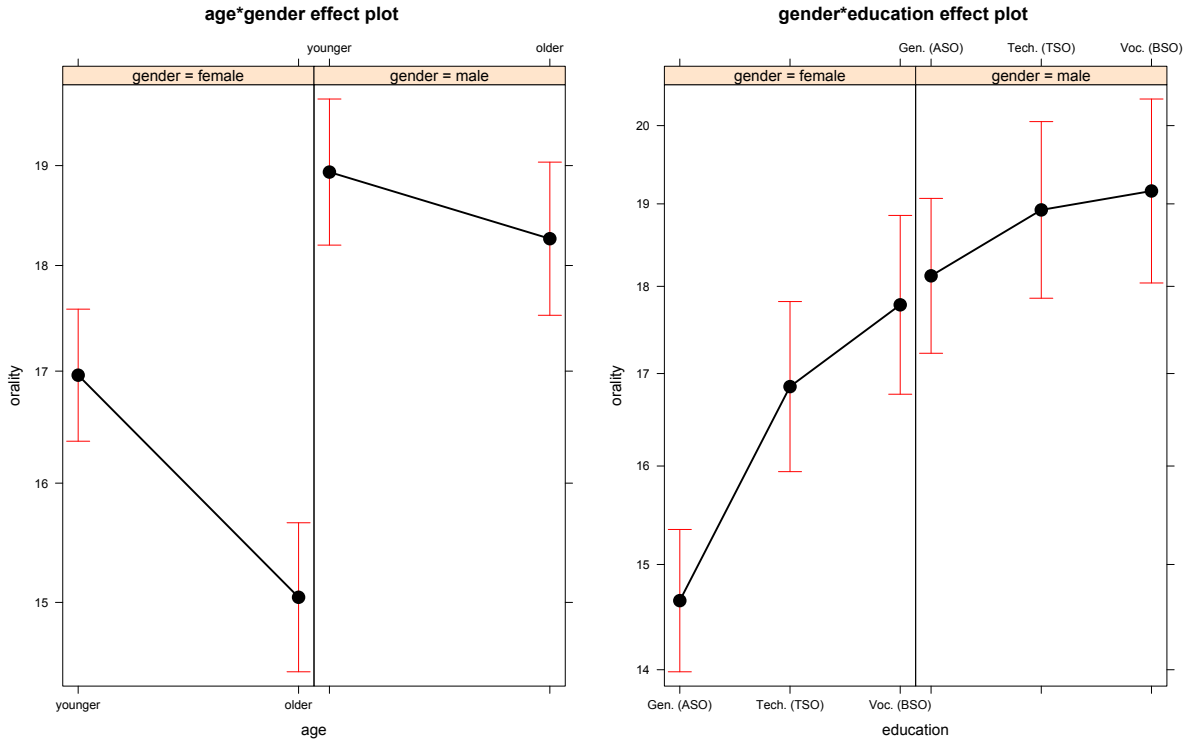


Figure 3: Orality model: Effect plot (predicted counts per 100 tokens)

	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.

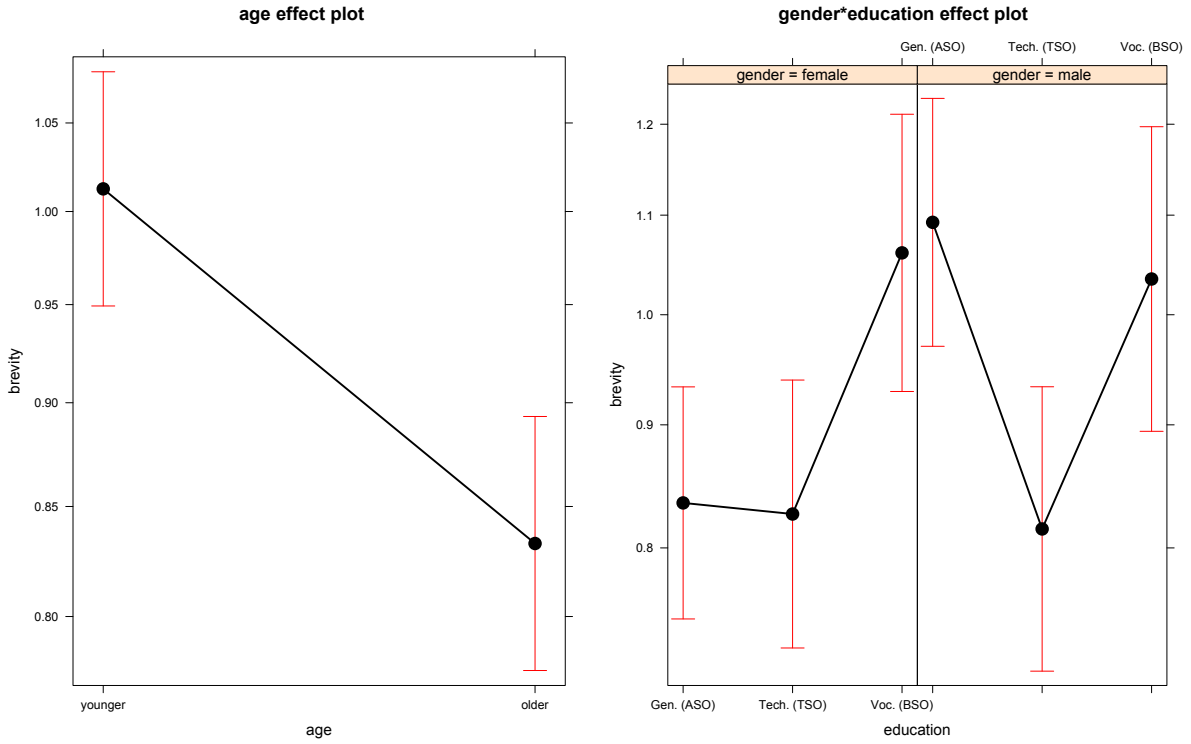


Figure 4: Brevity model: Effect plot (predicted counts per 100 tokens)

	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 et al. call the combination of standard language components and "socially meaningful non-standard features" a "linguistic tool for modern self-portrayal" (2016, 130). 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 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 *both* sexes (1983, 182-183). 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 which 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 e.g. 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, Koppel, Pennebaker, & Schler 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 (e.g. class groups) are often reflected in online networks, e.g. 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-

[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.

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 more narrow 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 educational 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. E.g.: 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 considered 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: e.g. the insertion of an emoticon can be seen as less non-standard than the use of a dialect word. Some features which 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 et al. 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" (2016, 131). 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 (e.g. 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 *perception* of computer-mediated communication will complement our previous and current work on the *production* of this varied, fascinating linguistic register, as we will not only try to answer the question of *how* teenagers write on social media, but also *why* they appear to favor certain linguistic markers or styles.

Acknowledgments

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

Publications

Allison, Paul D. (1999). *Logistic regression using SAS: Theory and application*. Cary: SAS Institute Inc.

Androutsopoulos, Jannis. (2011). Language change and digital media: A review of conceptions and evidence. In Tore Kristiansen, & Nikolas Coupland (Eds), *Standard languages and language standards in a changing Europe* (pp. 145-161), Oslo: Novus.

Argamon, Shlomo, Moshe Koppel, James W. Pennebaker, & Jonathan Schler. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM. Inspiring Women in Computing* 52, 119-123.

Baron, Naomi S. (2004). See you online: Gender issues in college student use of instant messaging. *Journal of Language and Social Psychology* 23(4), 397-423.

Bamman, David, Jacob Eisenstein, & Tyler Schnoebelen. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18(2), 135-160.

Bates, Douglas, Martin Maechler, Ben Bolker, & Steven Walker. (2017). Package 'lme4'. Url: https://cran.r- project.org/web/packages/lme4/lme4.pdf

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. *Journal of Personality Assessment* 91(2), 121-136.

De Decker, Benny, & Reinhild Vandekerckhove. (2012). English in Flemish adolescents' computer-mediated discourse: A corpus-based study. *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. *Nederlandse Taalkunde* 18(1), 2-34.

De Decker, Benny. (2014). *De chattaal van Vlaamse tieners. Een taalgeografische analyse van Vlaamse (sub)standaardiseringsprocessen tegen de achtergrond van de internationale chatcultuur*. Antwerp: University of Antwerp (doctoral thesis).

De Decker, Benny, & Reinhild Vandekerckhove. (2017). Global features of online communication in local Flemish: Social and medium-related determinants. *Folia Linguistica* 51, 253-281.

Eisikovits, Edina. (2006). Girl-talk/boy-talk: Sex differences in adolescent speech. In Jennifer Coates (Ed.), *Language and gender. A reader* (pp. 42-54), Oxford: Blackwell.

[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* (pp. 8-18). Brussels: Department of Education and Training.

Grondelaers, Stefan, & Dirk Speelman. (2013). Can speaker evaluation return private attitudes towards stigmatised varieties? Evidence from emergent standardisation in Belgian Dutch. In Tore Kristiansen, & Stefan Grondelaers (Eds), *Language (de)standardisation in late modern Europe: Experimental studies* (pp. 171-192), Oslo: Novus.

Grondelaers, Stefan, Roeland van Hout, & Paul van Gent. (2016). Destandardization is not destandardization. *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. *Lingua* 172, 62-71.

Harrison, Xavier A. (2014). Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ* 2, e616.

Hilbe, Joseph M. (2011). Modeling count data. In Miodrag Lovric (Ed.), *International encyclopedia of statistical science* (pp. 836-839), Berlin: Springer.

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* (pp. 328-334), Association for Computational Linguistics.

Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans. (2018a). Adolescents' social background and non-standard writing in online communication. *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. *International Journal of Society, Culture & Language* 6(2), 73-89.

Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans. (2018c). Expressive markers in online teenage talk: A correlational analysis. *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. *Casualty actuarial society forum. Citeseer*, 103-158.

Killermann, Sam. (2014). Breaking through the binary: Gender as a continuum. *Issues* 107, 9-12.

Kristiansen, Tore. (2001). Two standards: One for the media and one for the school. *Language Awareness* 10, 9-24.

Kristiansen, Tore, Peter Garrett, & Nikolas Coupland. (2005). Introducing subjectivities in language variation and change. *Acta Linguistica Hafniensia* 37, 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* (pp. 274-283), Berlin: Springer.

Labov, William. (1972). *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Labov, William. (2001). *Principles of linguistic change. Volume 2: Social factors*. Maiden: Wiley-Blackwell.

Milroy, Lesley, & Carmen Llamas. (2013). Social networks. In Jack K. Chambers, & Natalie Schilling (Eds), *The handbook of language variation and change. Second edition* (pp. 409-427), Oxford: Blackwell.

Parkins, Róisín. (2012). Gender and emotional expressiveness: An analysis of prosodic features in emotional expression. *Griffith Working Papers in Pragmatics and Intercultural Communication* 5(1), 46-54.

Pennebaker, James W. (2011). *The secret life of pronouns: What our words say about us.* New York: Bloomsbury Press.

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. *PLoS ONE* 8.

Tarpey, Thaddeus. (2012). *Generalized linear models (GLM).* Course notes obtained from: http://www.wright.edu/~thaddeus.tarpey/ES714glm.pdf

Trudgill, Peter. (1983). Social identity and linguistic sex differentiation / Sex and covert prestige. In Peter Trudgill, *On dialect. Social and geographical perspectives* (pp. 161-185), Oxford: Blackwell.

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. *Reading and Writing* 23, 719-733.

Verheijen, Lieke. (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).

Verheijen, Lieke, Laura de Weger, & Roeland van Hout. (2018). Code-mixing with English in Dutch youths' online language: OMG SUPERNICE LOL! In Reinhild Vandekerckhove, Darja Fišer, & Lisa Hilte (Eds), *Proceedings of the 6th conference on computer-mediated communication (CMC) and social media corpora* (pp. 63-67), Antwerp: University of Antwerp.

Wolf, Alecia. (2000). Emotional expression online: Gender differences in emoticon use. *Cyberpsychology & Behavior* 3, 827-833.

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

Zeileis, Achim, Christian Kleiber, & Simon Jackman. (2008). Regression models for count data in R. *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 Winthrop Nelson Francis and Henry Kučera. Brown University. Providence, Rhode Island, USA.

Algemeen Nederlands Woordenboek (ANW). http://anw.inl.nl/

Corpus of Contemporary American English (COCA). https://corpus.byu.edu/coca/
Dutch Parallel Corpus (DPC). https://www.kuleuven-kulak.be/DPC
Named Entity Recognition (NER) Datasets. CLiPS research center, University of Antwerp. http://www.cnts.ua.ac.be/conll2003/ner.tgz
Roularta Consortium (2011): Roularta corpus.
Stevin Nederlandstalig Referentiecorpus (SoNaR). https://ivdnt.org/downloads/tstc-sonar-corpus

CHAPTER 6

This chapter was published as a research paper in conference proceedings. Reference:

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, 328-334.

Predicting adolescents' educational track from chat messages on Dutch social media

Abstract
We aim to predict Flemish adolescents' educational track based on their Dutch social media writing. We distinguish between the three main types of Belgian secondary education: General (theory-oriented), Vocational (practice-oriented), and Technical Secondary Education (hybrid). The best results are obtained with a Naive Bayes model, i.e. an F-score of 0.68 (std. dev. 0.05) in 10-fold cross-validation experiments on the training data and an F-score of 0.60 on unseen data. Many of the most informative features are character n-grams containing specific occurrences of chatspeak phenomena such as emoticons. While the detection of the most theory- and practice-oriented educational tracks seems to be a relatively easy task, the hybrid Technical level appears to be much harder to capture based on online writing style, as expected.

1. Introduction

While some social variables, such as gender and age, have often been studied in author profiling (see e.g. the overview paper by Reddy et al. 2016), educational track remains largely unexplored in this respect. The goal of this paper is twofold: we aim to develop a model that accurately predicts adolescents' educational track based on their language use in social media writing, and gain more insight in the linguistic characteristics of youngsters' educational background through inspection of the most informative features for this classification task.

The paper is structured as follows: we start by discussing related research (Section 2). Next, we describe the corpus, as well as the three main types of Belgian secondary education, i.e. the three class labels in the classification experiments (Section 3). Finally, we discuss our methodology (Section 4) and present the results (Section 5).

2. Related research

Related work on this topic is scarce; only some studies in education profiling can be found, and they examine the impact of tertiary (and not secondary) education, on text genres other than social media writing. Furthermore, Dutch is never the language of interest. Estival et al. (2007), for instance, approached tertiary education profiling as a binary classification task (none versus some tertiary education) for a corpus of English emails. They obtained promising results with an ensemble learner (Bagging algorithm) using character-based, lexical and structural text features while explicitly excluding function words. Pennebaker et al. (2014), however, stressed the importance of function words in a related task: they linked students'

writing in college admission essays to their later performance in college. Obtaining higher or lower grades appeared to be associated with the use of certain function words, belonging to either 'categorical' or 'dynamic' writing styles. In previous work on language and social status, Pennebaker (2011) had already pointed out the importance of pronouns: he described a more frequent use of you- and we-words as more typical of high status, as well as a less frequent use of I-words.

When we expand the scope of previous research from profiling studies to other related linguistic fields, we again conclude that this specific topic is underresearched. There are many studies on the characteristics of (youngsters') computer-mediated communication (CMC) (see e.g. Varnhagen et al. 2010; Tagliamonte & Denis 2008; and many more) and even some on the interaction between CMC and education (see e.g. Vandekerckhove & Sandra 2016 for the impact of CMC on school writing). However, the impact of educational track on adolescents' online writing is not addressed. For this specific topic, we can – to our knowledge – only refer to our previous sociolinguistic work focusing on youngsters with distinct secondary education profiles, in which we have shown that teenagers in practice-oriented tracks tend to deviate more from formal standard writing on social media, by using more typographical chatspeak features (e.g. emoji), more non-standard lexemes (e.g. dialect words) and more non-standard abbreviations (Hilte et al. 2018a, 2018b). While for all examined linguistic features, these differences were very consistent between the two 'poles' of the continuum between theory and practice, i.e. General and Vocational students, the Technical students did not always hold an intermediate position, but their chat messages showed a rather unpredictable linguistic pattern (Hilte et al. 2018a, 2018b). We investigate in this paper whether these sociolinguistic results are confirmed in machine learning experiments.

3. Data collection

Our corpus consists of Flemish[1] adolescents' private chat messages, written in Dutch on the social media platforms Facebook Messenger and WhatsApp. The data were collected through school visits during which the students were informed about the research, and could voluntarily donate chat messages. We asked for the students' (and for minors, their parents') consent to store and analyze their anonymized texts.

The final corpus contains 434 537 chat messages (2 531 354 tokens) by 1384 authors. All authors are Flemish high school students, aged 13-20, attending one of the three main types of Belgian secondary education: the theory-oriented General Secondary Education (which prepares for higher education), the practice-oriented Vocational Education (which prepares for a specific manual profession) and the hybrid Technical Education, which has both a strong

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

theoretical and practical focus (Flemish ministry of education and training 2017). An overview of the distributions in the corpus can be found in Table 1.

We note that the Belgian secondary school system is similar to that of several other countries. The distinction between a vocational and an academic training is quite common (e.g. in Denmark, Finland, Croatia, France, Paraguay, China, etc.). The division between three main tracks (offering a more general, technical and vocational program respectively) is made in several countries as well (e.g. Czech Republic, Italy, Turkey, etc.)[2]. Consequently, the present classification task transcends the Belgian context and may be relevant in different countries and cultures, too.

Educational track	Participants	Posts	Tokens
General Secondary Education	596 (43%)	120 839 (28%)	739 831 (29%)
Technical Secondary Education	393 (28%)	197 534 (45%)	1 151 684 (46%)
Vocational Secondary Education	395 (29%)	116 164 (27%)	639 839 (25%)
Total	1 384	434 537	2 531 354

Table 1: Distributions in the corpus

4. Methodology

In this section, we describe the preprocessing of the data and the feature design (resp. Sections 4.1 and 4.2) as well as the experimental setup (Section 4.3).

4.1. Preprocessing

Since we will predict educational track on a participant-level, we must ensure to have sufficient data (and thus a fairly representative sample of online writing) for each participant. For this purpose, we deleted the participants who donated fewer than 50 chat messages. Next, we divided the remaining corpus in a training set (70% of the participants), and a test set (15%). A second test set (15%) was put aside for future experiments. This division was random but stratified, i.e. every subset contained the same proportion of participants per educational track.

4.2. Feature design

The features used in the classification experiments consist of general textual features and features representing the frequency of typical chatspeak phenomena.

The general features include frequencies for token n-grams (uni-, bi- and trigrams) and character n-grams (bi-, tri- and tetragrams). In addition, average token and post length and vocabulary richness (type/token ratio) are taken into account as well. Finally, we use the

[2] en.wikipedia.org/wiki/List_of_secondary_education_systems_by_country

dictionary-based computational tool LIWC (Pennebaker et al. 2001) in an adaptation for Dutch by Zijlstra et al. (2004) to count word frequencies for semantic and grammatical categories. While counts for individual words are already captured by the token unigrams, these counts per category can allow for broader generalizations for words which are semantically or functionally related. However, we note that the accuracy of this feature might not be optimal, as the social media texts are very noisy (and contain many non-standard elements, e.g. in terms of orthography or lexicon), whereas LIWC is based on standard Dutch word lists.

The set of chatspeak features contains counts for occurrences of several typographic phenomena. It includes the number of character repetitions (e.g. *suuuuuper nice!!!*) and combinations of question and exclamation marks (e.g. *what?!*). The number of unconventionally capitalized tokens is added as well (alternating, inverse or all caps, e.g. *AWESOME*). The final typographic features are emoticons and emoji (e.g. *:)*, *<3*), the rendition of kisses and hugs (e.g. *xoxoxo*), hashtags for topic indication (e.g. *#addicted*) and 'mentions' for addressing a specific person in a group conversation (e.g. *@sarah*). We also add an onomatopoeic variable, i.e. the number of renditions of laughter (e.g. *hahahahah*). Another typical element of chatspeak are non-standard abbreviations and acronyms (e.g. *brb* for 'be right back'). The final feature concerns language or register choice per token, in order to explicitly take into account the authors' use of words in a different language or linguistic variety than standard Dutch. We count the number of standard Dutch, English, and non-standard Dutch (e.g. dialect) lexemes. While the other chatspeak features are detected with regular expressions (typographic and onomatopoeic markers) or predefined lists (abbreviations), this lexical feature is extracted using a dictionary-based pipeline approach. For each token, we first checked if it was an actual word (and not e.g. an emoticon). Next, we checked if it occurred in a list of standard Dutch words and named entities. If not, we checked its presence in a standard English word list. Finally, if the token was absent again, it was placed in the 'non-standard Dutch' category. Figure 1 shows a sample of authentic chat messages from the corpus, illustrating the use of several chatspeak features.

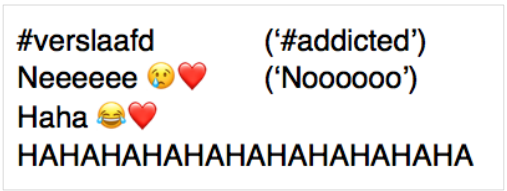


Figure 1: Example messages from the corpus

For each participant, an individual feature vector was created containing the counts for all of these features. We proceeded with relative counts (to normalize for submission size) by dividing the absolute counts by the author's total number of tokens (e.g. for token unigrams, emoji, …) or n-grams (for n-gram frequencies). For initial dimensionality reduction, we applied a frequency cutoff, only taking features into account that are used at least 10 times in the corpus, by at least 5 different participants.

4.3. Experimental setup

We compared different models to predict Flemish adolescents' educational track based on their social media messages. The classification algorithms we tested were: Support Vector Machines, Naive Bayes (Multinomial, Gaussian and Bernoulli), Decision Trees, Random Forest, and Linear Regression. For all classifiers, we used the Scikit-learn implementation (Pedregosa et al. 2011). For each model, we searched for the optimal parameter settings through a randomized cross-validation search on the training data. We searched for optimal values for classifier-bound parameters (e.g. kernel for SVM), as well as an optimal feature scaler (no scaling, MinMax scaling or binarization) and an optimal percentile for univariate (chi-square based) feature selection, chosen from a continuous distribution. We compared the models' performance in 10-fold cross-validation experiments on the training data.

5. Results

In Section 5.1, we discuss the best model resulting from the 10-fold cross-validation experiments on the training data and compare it to different baseline models. In addition, we inspect the most informative features for the task. In Section 5.2, we discuss additional experiments which provide further insight in the classification problem.

5.1. Model performance and feature inspection

Class levels	Precision	Recall	F-score
General	0.67	0.78	0.72
Technical	0.70	0.54	0.61
Vocational	0.68	0.71	0.70
Avg/total	0.68	0.68	0.68

Table 2: Classification report (in cross-validation)

		Predicted class		
		Gen.	Tech.	Voc.
Actual class	Gen.	153	22	22
	Tech.	49	89	27
	Voc.	25	17	105

Table 3: Confusion matrix (in cross-validation)

The best performing model in CV-setting on the training data is a Multinomial Naive Bayes classifier, with optimized parameters: the value for the smoothing parameter alpha is 0.98, and the model uses the 12.50% best features (according to chi-square tests). The features were binarized. The classification report (Table 2) indicates that the performance is good, with a value of 0.68 for (prevalence-weighted macro-average) precision, recall and F-score (std. dev. 0.05). While precision is very similar for the three educational levels, recall is good for General Education, but slightly worse for the Vocational and much worse for the Technical

level. Consequently, the model seems to miss many Technical profiles, confusing them with the other educational tracks. The confusion matrix (Table 3) shows that most (64%) misclassified Technical profiles were incorrectly labeled as the more theory-oriented General track, rather than as the more practice-oriented Vocational track (36%).

As Table 5 summarizes, the model strongly outperforms a probabilistic baseline (0.34) in cross-validation, as well as a simple bag-of-words model (which only uses token unigrams as features) without any parameter tuning, scaling or feature selection (F-score = 0.22). However, when parameter tuning, scaling and feature selection are introduced, the BoW-model obtains almost identical scores in cross-validation: it yields an overall precision, recall and F-score of 0.67 (std. dev. 0.03). There is, however, a difference in how well both models generalize to unseen data. While the first model reaches an average F-score of 0.60 (see Table 4 for the detailed classification report), the BoW-model achieves a lower score of 0.55, and particularly underperforms in the detection of Technical profiles, with an F-score of 0.38 (vs 0.50 for the full model).

Class levels	Precision	Recall	F-score
General	0.64	0.69	0.67
Technical	0.57	0.44	0.50
Vocational	0.58	0.68	0.63
Avg/total	0.60	0.61	0.60

Table 4: Classification report (on unseen data)

	Cross-validation			Unseen data		
Model	Precision	Recall	F-score	Precision	Recall	F-score
Best model	**0.68**	**0.68**	**0.68**	**0.60**	**0.61**	**0.60**
BoW (non-finetuned)	0.15	0.39	0.22	0.15	0.39	0.21
BoW (finetuned)	0.67	0.67	0.67	0.55	0.55	0.55
Stylistic	0.65	0.64	0.64	0.59	0.60	0.59
Prob. baseline	0.34	0.34	0.34	0.34	0.34	0.34

Table 5: Comparison of the different models and baselines

In order to better understand the differences and similarities between both models, we compared their feature sets (after feature selection was applied) and inspected the 1000 most informative ones, using information gain as ranking criterion. While we expected that the most informative features for the BoW-model would be lexical and the ones for the full model stylistic, this analysis suggests that in both models, many of the most informative selected features are specific occurrences of chatspeak markers. For the BoW-model, which uses only token unigrams as features, many of the most informative tokens contain one or more chatspeak features (e.g. colloquial register, a spelling manipulation, an emoticon, character repetition, etc.). Some other informative tokens seem to be more content- than style-related, revealing topics such as hobbies, specific locations, friends and school. Strikingly, although the full model contains abstraction of chatspeak phenomena (e.g. total count for emoticons), specific occurrences of these genre markers are still most informative.

The 1000 most informative features are all character n-grams: only some reveal topics (e.g. school), but many more indicate the use of chatspeak features, and particularly combinations of emoji/emoticons. Other n-grams indicate the use of English and Arabic words, of colloquial terms, of chatspeak spelling, abbreviations and character repetition. As opposed to the BoW-model's token unigrams, these character n-grams allow the model to capture stylistic features on a sub-token level (e.g. the n-gram *sss* captures repetition of the letter 's' in different words). We can illustrate a clear advantage by the Arabic word *wallah* (meaning 'I swear on God's name'), which is often used by our participants with Arabic roots, who spell it in many different ways. Because of these alternative spellings, *wallah* does not appear among the most informative tokens in the BoW-model. However, for the full model, several related character n-grams (e.g. *wlh*, *wll*) do.

Next, we compared the full model to a stylistic model using only chatspeak features (both abstractions and specific occurrences), and no token or character n-grams. This stylistic model performs slightly worse on both the training set (F-score = 0.64, std. dev. 0.04) and unseen data (F-score = 0.59) (see Table 5). However, inspection of the most informative features in this feature set provides further insight in the education profiling task. Many of the most informative features are again specific occurrences of stylistic phenomena (e.g. specific emoticons, specific lexemes containing letter repetition). Some abstract representations of online writing style characteristics appear among the top-1000 features too (such as the total use of character repetition, of onomatopoeic laughter, acronyms, English words, mentions and hashtags, and emoticons), but much less prominently. These findings suggest that even in a purely stylistic model, abstract representation of certain style features is not informative enough for education profiling, and appears to be less important than the use of these features within specific tokens or contexts.

5.2. Additional experiments

Additional experiments indicate that the task becomes much easier when the hybrid Technical Education level is not included. Performance for this binary classification task (distinguishing between General and Vocational students only) is much higher (F-score = 0.81 with std. dev. 0.04 in cross-validation, and 0.78 on unseen data; see Tables 6 and 7 for the classification reports), showing that Vocational and General students are not often linguistically confused by the model. Strikingly, in this setting, the purely stylistic model performs similarly on the training data (F-score = 0.81, std. dev. 0.08), and even better on the unseen data (F-score = 0.82) than the full model. This suggests that stylistic differences are more outspoken and consistent between General and Vocational students, and might be sufficient for classification.

Class levels	Precision	Recall	F-score
General	0.86	0.80	0.83
Vocational	0.75	0.83	0.79
Avg/total	0.82	0.81	0.81

Table 6: Classification report for binary task (in cross-validation)

Class levels	Precision	Recall	F-score
General	0.82	0.79	0.80
Vocational	0.73	0.77	0.75
Avg/total	0.78	0.78	0.78

Table 7: Classification report for binary task (on unseen data)

Finally, first experiments with separate classifiers for girls and for boys, and for younger versus older teenagers, suggest interesting distinctions (see Table 8). It appears to be easier to correctly predict educational track for girls (F-score = 0.67 with std. dev. 0.07 in cross-validation; and 0.69 on unseen data) than for boys (F-score = 0.60 with std. dev. 0.09 in cross-validation; and 0.66 on unseen data). This suggests that more education-based linguistic variation can be found among girls than among boys. Similarly, better predictions could be made on unseen data for older teenagers, aged 17-20 (F-score = 0.62 in cross-validation, std. dev. 0.07; and 0.63 on unseen data), than for younger adolescents, aged 13-16 (F-score = 0.69 in cross-validation, std. dev. 0.09; and 0.55 on unseen data). This might be due to the fact that the older teenagers have been together in the same peer networks and class groups for a longer time, and might write more similarly on social media. Furthermore, some of the younger students might actually still change educational track.

	Cross-validation			Unseen data		
Model	Precision	Recall	F-score	Precision	Recall	F-score
Girls	0.67	0.67	0.67	**0.69**	**0.69**	**0.69**
Boys	0.61	0.61	0.60	0.67	0.67	0.66
Younger	**0.69**	**0.69**	**0.69**	0.55	0.55	0.55
Older	0.62	0.62	0.62	0.63	0.63	0.63

Table 8: Comparison of the models for separate groups

6. Conclusion

We conducted classification experiments to predict educational track for Flemish adolescents, based on their social media writing. These first results are promising and indicate that the task is doable. However, although the best model strongly outperforms a probabilistic baseline, its performance is similar to that of a simple BoW-model. This might give the impression that lexical features are still very important; however, inspection of the most informative features revealed that many of the most informative tokens contain stylistic features typical of the informal online genre. The most informative features for the full model suggest that abstraction of these stylistic chatspeak features (or at least, the current implementation) is still of lesser importance than specific occurrences.

While the distinction between General and Vocational high school students appears to be relatively easy to make, the detection of students in the intermediate Technical track is much harder. This could indicate that these students are truly a hybrid class with subsets of students that are simply not that different from their peers in more theory- or more practice-oriented tracks, respectively. In addition, related research shows that these students' online writing is rather unpredictable and does not follow a clear pattern (Hilte et al. 2018a, 2018b).

In future work, we want to experiment with additional algorithms, such as ensemble methods, and with a post-level rather than a participant-level approach (in order to have more data samples at our disposal). We also want to improve the current feature design and particularly the abstract representation of style features, because as van der Goot et al. (2018) write, abstract features may increase generalizability to other corpora (and even genres and languages) in author profiling tasks, compared to lexical models. Finally, we want to further investigate the creation of different classifiers for different subgroups of participants (e.g. boys versus girls).

Finally, we stress that this profiling task is not only relevant in a Belgian context, since the educational tracks serving as class labels correspond to several countries' secondary education programs. Furthermore, the inclusion of stylistic features – i.e. chatspeak phenomena occurring in *any* language – adds to this generalizability. While specific lexemes or specific realizations of chatspeak markers may not always be relevant in other languages or corpora, the abstract stylistic features are more universal on social media. We argue that these models for education profiling, when further improved, could be used in different languages and applications. For instance, the addition of an educational compound can increase existing profiling tools' performance, which can be important in different tasks (e.g. the detection of fake accounts on social media, and many more).

7. Supplementary materials

Because of the decision of our university's ethical committee, in line with European regulations to ensure the adolescents' privacy, we cannot make the dataset publicly available. The code will be made available.

Acknowledgments

We thank Stéphan Tulkens for his advice on the setup and analyses. We are also grateful towards the two anonymous reviewers for their feedback on a previous version of this paper.

References

Estival, Dominique, Tanja Gaustad, Son Bao Pham, Will Radford, & Ben Hutchinson. (2007). Author profiling for English emails. In *Proceedings of the 10th conference of the Pacific association for computational linguistics* (pp. 263-272).

Flemish Ministry of Education and Training. (2017). *Statistisch jaarboek van het Vlaams onderwijs. Schooljaar 2015-2016.* Brussels: Department of Education and Training.

Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans. (2018a). Adolescents' social background and non-standard writing in online communication. *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. *International Journal of Society, Culture & Language* 6(2), 73-89.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, 2825-2830.

Pennebaker, James W. (2011). *The secret life of pronouns. What our words say about us.* New York: Bloomsbury Press.

Pennebaker, James W., Cindy K. Chung, Joey Frazee, Gary M. Lavergne, & David I. Beaver. (2014). When small words foretell academic success: The case of college admissions essays. *PLoS ONE* 9(12), e115844.

Pennebaker, James W., Martha E. Francis, & Roger J. Booth. (2001). *Linguistic inquiry and word count: LIWC 2001.* Mahway: Lawrence Erlbaum Associates.

Reddy, T. Taghunadha, B. Vishnu Vardhan, & P. Vijayapal Reddy. (2016). A survey on authorship profiling techniques. *International Journal of Applied Engineering Research* 11(5), 3092-3102.

Tagliamonte, Sali A., & Derek Denis. (2008). Linguistic ruin? LOL! Instant messaging and teen language. *American Speech* 83(1), 3-34.

van der Goot, Rob, Nikola Ljubešić, Ian Matroos, Malvina Nissim, & Barbara Plank. (2018). Bleaching text: Abstract features for cross-lingual gender prediction. In *Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 2: Short papers)* (pp. 383-389).

Vandekerckhove, Reinhild, & Dominiek Sandra. (2016). De potentiële impact van informele online communicatie op de spellingpraktijk van Vlaamse tieners in schoolcontext. *Tijdschrift voor Taalbeheersing* 38(3), 201-234.

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. *Reading and Writing* 23(6), 719-733.

Zijlstra, Hanna, Tanja Van Meerveld, Henriët Van Middendorp, James W. Pennebaker, & Rinie Geenen. (2004). De Nederlandse versie van de 'linguistic inquiry and word count'(LIWC). *Gedrag Gezond* 32, 271-281.

CHAPTER 7

This chapter has been submitted as a journal article. Reference:

Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans. (submitted). Lexical patterns in adolescents' online writing: The impact of age, gender and education.

Lexical patterns in adolescents' online writing:

The impact of age, gender and education

Abstract
The present paper examines the impact of adolescents' socio-demographic profile (i.e. their age, gender and educational track) on lexical aspects of their online discourse. A variety of lexical features and related parameters is examined, such as lexical richness, top favorite words and word length. The analyses reveal a strong common ground among the adolescents with respect to some features (e.g. conversation topic) but divergent writing practices by different groups of teenagers with regards to other parameters (e.g. lexical richness). Furthermore, this study combines a traditional focus (examining standardized versions of social media messages) and a new media focus (examining the original utterances, including non-standard chatspeak markers). Strikingly, different results emerge with respect to adolescents' exploitation of more traditional versus digital literacy skills, in particular with respect to the expression of sentiment (verbal versus typographic/pictorial).

Keywords: Social media, teenagers, post length, word length, lexical richness, sentiment, topic analysis

1. Introduction

Informal online writing tends to deviate from formal 'standard' writing practices in various ways, e.g. with respect to spelling or typography. Several of these deviations from standard writing can be considered prototypical markers of the genre (e.g. emoji). While many previous studies discuss the incorporation of such *chatspeak markers* in online discourse, more traditional linguistic variables and patterns – related to general language proficiency rather than to the specificities of online communication – are less prominent in research on computer-mediated communication (CMC). The present study will analyze the latter type of features in youths' online writing, with a specific focus on lexical patterns. This lexical focus is motivated by the fact that many typographic chatspeak markers can take over the function of lexical items to a certain extent: e.g. while emotional involvement can be expressed lexically, emoticons may serve the same purpose in online writing. So teenagers have both a 'traditional' (lexical) and 'digital media' (typographic/pictorial) repertoire at their disposal for informal online communication. However, it has hardly been investigated to which extent they use both repertoires and whether their preferences in this respect are socially determined. Since previous research indicates that teenagers' production of chatspeak markers is significantly impacted by multiple aspects of their socio-demographic profile, we will 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).

The paper is structured as follows: Section 2 presents an overview of related research. Next, in Section 3, the corpus and participants are described. Section 4 introduces the linguistic variables along with the methodology for feature extraction and the linguistic and statistical analyses. We note that some methodological challenges emerged because of the 'noisy' (i.e. non-standard) nature of the online text genre; consequently, (some degree of) normalization of the original social media messages was required. The applied normalization strategy is discussed and evaluated in Section 4 too. Finally, In Sections 5 and 6, the results of the analyses are presented and discussed, respectively.

2. Related research

The linguistic characteristics of informal computer-mediated communication (or CMC) have been widely investigated. Quite a lot of studies with a sociolinguistic orientation demonstrate how people with distinct socio-demographic profiles (e.g. in terms of age or gender) appear to favor certain prototypical chatspeak markers (e.g. typographic features such as emoji) to different extents (De Decker 2014; Hilte et al. forthcoming, 2018c; Varnhagen et al. 2010; Verheijen 2015; Wolf 2000; and many more). However, the more traditional linguistic features and writing patterns that are of interest in the present paper are much less prominent in CMC-research. Furthermore, no consensus emerges regarding the attested sociolinguistic patterns.

First of all, with respect to sentence length, Lin reports that adult male authors produce longer sentences in chat conversations than their female peers (2007, 20-21). However, for *adolescent* authors, she observes the opposite tendency (2007, 20-21). This tendency of girls/women producing longer sentences is confirmed in other research on both spoken and written language (Newman et al. 2008, 213, and references therein), although occasionally men have also been found to produce longer phrases in a spoken conversational context (Singh 2001, 260). Some other related results concern text length rather than sentence length: in a formal (non-conversational) writing task, girls and higher educated youths produced longer texts (Verheijen & Spooren 2017, 9). As for average word length, consistent gender findings are reported, with males producing more longer words in both spoken and written (CMC) conversations (Lin 2007, 21, 25; Mehl & Pennebaker 2003, 865; Newman et al. 2008, 213-214, 223, and references therein).

With respect to lexical richness, however, conflicting patterns of gender-related variation are attested. A larger vocabulary range is reported for male adolescent online writing (Lin 2007, 21, 25). Some research confirms this pattern for spoken conversations (Singh 2001, 260), but other studies reveal no significant gender differences for spoken or (offline) written discourse (Yu 2009, 253). In formal non-conversational writing tasks, both higher educated and older youths have been found to produce lexically richer texts than their lower educated and younger peers, respectively (Verheijen & Spooren 2017, 9). Finally, Yu reports an important relationship between lexical richness and (the evaluation of) language proficiency: in a corpus

containing speaking and writing tasks, significant positive correlations are attested between the lexical richness in the tasks and the candidates' general language proficiency, and between the lexical richness in the tasks and raters' judgement of the overall quality of the candidates' writing and speaking performances (2009, 236). With respect to the diverging results discussed in this paragraph, we note that 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, 241). Therefore, different measures are discussed in Sections 4.1 and 7.

Two major points of reference for the analysis of authors' top favorite words (and the associated topics) in CMC are the studies on English blog posts conducted by Schwartz et al. (2013) and Argamon et al. (2009). With respect to gender, their results reveal that many of female bloggers' most prominent words relate to personal life and relationships (e.g. *boyfriend*, *mom*, *bestie*) (Argamon et al. 2009, 121; Schwartz et al. 2013, 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, 121; Schwartz et al. 2013, 8). A female preference for positive emotion words is also reported by Mehl and Pennebaker (2003, 866). Finally, some prominent lexemes used by women reveal more (stereo)typical female topics (e.g. *chocolate*, *shopping*, *my hair*) (Schwartz et al. 2013, 8). Many of the male bloggers' most prominent words concern politics (e.g. *government*, *democracy*) and fighting (e.g. *fight*, *battle*) (Schwartz et al. 2013, 8). Swear words frequently occur among the top male lexemes too (e.g. *fuck*, *shit*) (Schwartz et al. 2013, 8, but also Mehl & Pennebaker 2003, 866; Newman et al. 2008, 213-214, and references therein). Furthermore, negative emotion words (specifically those related to anger) also appear to be more frequent in male online writing (Mehl & Pennebaker 2003, 866). Finally, some prominent lexemes used by men reveal more (stereo)typical male topics, such as technology (e.g. *system*, *software*), gaming (e.g. *xbox*, *ps3*) and football (e.g. *football*, *team*) (Argamon et al. 2009, 121; Schwartz et al. 2013, 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 twenties (Argamon et al. 2009, 121-122; Schwartz et al. 2013, 10).

Finally, with respect to the expression of sentiment or emotion in CMC, very consistent age and gender patterns are reported in previous work. Girls/women and younger people appear to use more emotionally expressive language than their male and older peers. These tendencies hold both for the use of emotion words, i.e. *lexical* expressiveness, and for *typographic* expressiveness (see Section 6), in offline as well as online communication (Baron 2008, 51; Hilte et al. 2018c; Newman et al. 2008, 223, 229; Schwartz et al. 2013, 9; Wolf 2000,

831). As for educational variation, youths in practice-oriented tracks appear to use more typographic emotional markers in CMC, too (see Hilte et al. forthcoming, 2018a, 2018b).

While related research reveals interesting tendencies concerning the lexical patterns and related parameters included in this paper, these variables are seldom included in research on social media writing. Furthermore, the use of lexical expression, in other words classical verbal expression, is seldom set off against the exploitation of typographic means of expression that mark online communication. The present paper aims to fill that gap.

3. Data

The corpus consists of 434,537 social media posts (over 2.5 million tokens) written by 1384 Flemish[1] secondary 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. Dialect region is a quasi-constant, as 96% of the teenagers live in the central Flemish province of Antwerp.

Three aspects of the adolescents' socio-demographic profile are included in the research design as independent variables (and this information is available for *all* participants in the corpus): their age, gender and educational track (see Table 1 for an overview of the distributions in the corpus). 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 adolescents' non-standard language use does not evolve linearly, but 'peaks' mid-puberty (around the age of 16) – a sociolinguistic phenomenon that is 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, with a distinction between girls and boys, since a non-binary approach (e.g. conceptualizing gender as a continuum) was infeasible given the available profile information.

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, with a practical and theoretical orientation, and a focus on technical courses (FMET 2018, 10). An educational difference that might be of particular importance in the present study concerns the focus on standard Dutch proficiency and formal writing, which is stronger in more theoretical tracks. While correct and formal standard Dutch writing is a major objective in theoretical school systems, it is much less of a priority in practice-oriented

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

tracks. A weaker familiarity with and possibly also a more limited proficiency in the formal written standard might also influence these students' writing practices on social media, even though the genre is essentially different.

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[2]
	Older teenagers / young adults (17-20)	1 170 456 (46%)	897
Total		2 531 354	1 384

Table 1: Distributions in the corpus

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 before our visits. The latter condition was meant to exclude the observer's paradox. We asked the students' (and for minors also their parents') consent to store and linguistically analyze their anonymized texts.

4. Linguistic variables and methodology

Below, the linguistic variables included in the research design are presented (Section 4.1). Sections 4.2 and 4.3 discuss the challenges with regards to the 'noisy' (non-standard) nature of the social media texts, and the applied normalization procedure, respectively. Finally, Section 4.4 presents the methodology and statistical models used in this study.

4.1. Linguistic variables

In order to obtain a nuanced view on lexical variation and related matters, a variety of features is examined. First of all, each author's (productive) *lexical richness*[3] is measured, which "summarizes the range of vocabulary and the avoidance of repetition in the sample" (Malvern & Richards 2012, 1). We operationalize lexical richness as the Guiraud correction of

[2] The number of younger and older participants does not add up to the total number of participants, but to a higher number (which is why 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 subject (participant) as a random effect in the statistical models (see Section 4.3).

[3] In related work, this variable is referred to by a variety of names, such as *lexical diversity*, *lexical density*, *lexical variation*, *vocabulary richness* and *vocabulary size*.

the type/token-ratio. Type/token-ratio (TTR), i.e. 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, 66). However, it is heavily criticized, as its outcome may be unreliable when samples of different lengths are compared (van Hout & Vermeer 2007, 121; Yu 2009, 239). The measure is "notorious for being sensitive to sample size" (Yu 2009, 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, 2; Vermeer 2000, 68; Yu 2009, 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* or *index of Guiraud*) is considered a more adequate measure of lexical diversity, holding a more constant value for increasing sample sizes (Vermeer 2000, 68), and a "happy medium between doing nothing to the number of tokens (TTR) and applying a too strong a transformation […] that levels out all relevant differences" (van Hout & Vermeer 2007, 136). For an overview and evaluation of different approaches (including other adjustments of TTR as well as more complex measures), see e.g. Malvern and Richards (2012) and van Hout and Vermeer (2007). Finally, we note that in the present study, lexical diversity is calculated for the non-lemmatized tokens (e.g. *loop* 'run' and *liep* 'ran' are counted as two different types, and not as two occurrences of the same lemma – i.e. the canonical/dictionary form or 'base form' –, i.e. the lemma *lopen* 'to run').

The next variable concerns the authors' top favorite words or lexemes. For this analysis, we automatically extract the 500 most frequently used words per subgroup of participants (e.g. girls versus boys) and manually inspect these in order to discover the associated topics. However, as "direct association of word types with high-level dimensions remains problematic" (Bamman, Eisenstein & Schnoebelen 2014, 145), the topics that will be assigned to the lexemes should be interpreted as suggestive rather than absolute labels.

The third and fourth dependent variables are the authors' average token[4] and post length, expressed in number of characters and number of tokens, respectively. We note that after normalization of the corpus, a token always represents a word (and not e.g. an emoticon – see Section 4.3 for more information on the normalization procedure). The production of longer tokens thus equals the production of longer words, and might be indicative of a stronger command of more complex words and thus potentially a stronger traditional literacy. 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 token-level feature. For each post in the corpus, the lexical expression of sentiment is measured, using the 'sentiment' function in the Pattern package for Python (De Smedt & Daelemans 2012a). This function assigns two scores to a text string. The *polarity score* expresses how negative or

[4] A token is a visual unit separated by whitespace from the preceding visual unit.

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 chatspeak markers (e.g. emoticons) by comparing adolescents' exploitation of a traditional (verbal) and a digital (typographic) repertoire with respect to the expression of emotional or social involvement in their online writing.

4.2. 'Noisy' text: Issues and challenges

The feature extraction from the corpus and statistical analysis is complicated by the 'noisy' nature of the social media texts: many messages contain various deviations from standard writing, mainly in terms of spelling (i.e. words are rendered in different, non-standard, ways) or typography (e.g. deliberate letter repetition, the use of emoji). As illustrated below, this generates distorted results with regards to the measurement of lexical richness and the analysis of the lexical expression of sentiment. Therefore, a reliable analysis requires normalization or standardization of the corpus, i.e. a conversion of the original utterances into their standard Dutch equivalent.

Starting with lexical richness: utterance (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 type/token-ratio 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 'non-word' elements, i.e. tokens that are not words, but e.g. 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 non-standard spelling or morphology can distort the results with regards 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 non-standard spelling alternative) rather than lexical in nature. In example (4), the non-standard 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 (5), 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.

(3) ***egt*** *vervelend* / *ben het* ***echt*** *beu* ('really annoying' / 'I'm really sick of it')
(4) ***kga*** *eten* / ***ik ga*** *eten* ('I'm going to eat')
(5) ***OMG*** *geweldig* / ***oh my god*** *geweldig* ('oh my god, awesome')

Other issues emerge concerning the analysis of lexically expressed sentiment, since the automated tool used for this examination is based on a (sentiment) 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 non-standard text.

Nr.	Utterance	Polarity [-1, 1]	Subjectivity [0, 1]
(6)	Ik ben blij ('I am happy')	0.55	0.95
(7)	Ik ben zeer blij ('I am very happy')	0.61	1.00
(8)	Ik ben zeer blij! ('I am very happy!')	0.76	1.00
(9)	**Ik ben zeer bly! ('I am very happy!')**	**-0.63**	**0.90**
(10)	**kben echt meeeeega bly!! :D ('I'm really suuuuuper happy!! :D')**	**0.66**	**0.70**

Table 2: Illustration of sentiment analysis

The first three posts in Table 2 are standard Dutch sentences. For these examples, the sentiment function performs well: compared to message (6), the polarity score increases when the intensifying adverb *zeer* ('very') is added to the positive adjective *blij* ('happy') in message (7), and it increases even more when an exclamation mark is added in message (8). Consequently, an increasingly positive sentiment appears to be expressed in messages (6) to (8). The subjectivity scores follow a similar pattern. However, when non-standard elements are added to the utterances, the output of the sentiment function becomes less reliable. The sole deviation from standard Dutch in message (9) is the non-standard spelling of the adjective *blij* ('happy') as *bly* – this cluster reduction (*ij* to *y*) is a common spelling manipulation in Dutch CMC. Table 2 demonstrates how this small orthographic adaptation causes the polarity score to drop under zero, which indicates that the utterance is considered to express a negative rather than a positive sentiment. In addition, the subjectivity score slightly decreases too. Message (10), finally, can be considered as a more enthusiastic and also a very non-standard version of the original message (6), containing multiple common chatspeak markers (contraction of *ik ben* 'I am' to *kben*, expressive lengthening of the vowel in the intensifier *mega* ('super'), chatspeak spelling of *bly,* expressive repetition of the exclamation mark, and a smiley face emoticon). Even though intuitively message (10) 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 the table. This demonstrates the unreliability of the tool's outcome when applied to noisy social media data, and thus confirms the need for normalization of the corpus prior to linguistic analysis.

4.3. 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[5], we decided to develop our own normalization procedure in order to improve the results. For alternative approaches of normalizing social media data, see e.g. De Clercq, Schulz, Desmet, Lefever and Hoste (2013) and Han, Cook and Baldwin (2013).

The applied normalization procedure is token-based and consists of four steps (summarized and illustrated in Table 3). In step one, non-word tokens (e.g. emoji) are deleted. In step two, the remaining tokens' typography is normalized (e.g. expressive character repetition is reduced). These first two steps were carried out automatically using regular expressions. In step three, common non-standard abbreviations and acronyms are replaced by their full versions, and in the fourth and final step, common non-standard renderings of Dutch words or contractions of (multiple) words are replaced by their standard equivalents. For these last two steps, predefined handcrafted lists were used, containing both non-standard forms and their standard Dutch equivalent.

Step	Example: original post	Example: post after normalization step
1. Delete non-words	*dammn we look so hot* 😍🔥💋❤	*dammn 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 non-standard 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]')*

Table 3: Normalization procedure

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 (summarized in Table 4). In the first two scenarios, the original token was already rendered conform the standard (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

[5] 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.

concerned non-standard 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 fairly accurate and therefore reliable for further linguistic analysis.

Scenario	Before	After	Nr. of tokens
1	**standard**	**standard (unchanged)**	**406 (69%)**
2	standard	incorrectly changed	0
3	non-standard	non-standard (unchanged)	48 (8%)
4	non-standard	incorrectly changed	0
5	**non-standard**	**standard (changed)**	**137 (23%)**

Table 4: Error analysis of the normalization procedure

4.4. Methodology

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 frequently used tokens were inspected manually.

All other linguistic analyses were carried out using linear mixed models (LMMs), as implemented in the 'lme4' package for R (Bates et al. 2017). Per linguistic feature (e.g. average post length), a separate model is trained, with that particular feature serving as response variable. The models enable simultaneous inspection of the impact of the different social variables (serving as *fixed* effects or predictors) included in the research design, i.e. the authors' age, gender and educational track. Each predictor's main effect on the linguistic variable is examined as well as its impact in interaction[6] with the other predictors. For each linguistic variable, the optimal model (and its optimal subset of predictors) is experimentally determined using a backward stepwise procedure in which fixed effects with a non-significant impact are removed. In addition to the fixed effects, a random effect for participant is always added, which enables the models to take into account the impact of individual chatters, and to deal with repeated observations (i.e. the teenagers can thus occur in the corpus at both a younger and an older age). For more detailed information on these models, we refer to Hilte et al. (forthcoming).

Finally, we note that all LMM-analyses are carried out on the participant-level (rather than a post- or token-level) – e.g. average sentiment scores are 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.

[6] In the results section (Section 5), graphs will only be inserted to visualize interactions. If the optimal model does not include any significant interactions, the (much less complex) patterns are only described in the text.

5. Results

This section presents the results per linguistic variable. We recall that the analyses are carried out on the normalized versions of the social media texts in the corpus. Additional examinations of the *original* texts (including non-standard elements) are discussed in Section 6.

5.1. Average post length

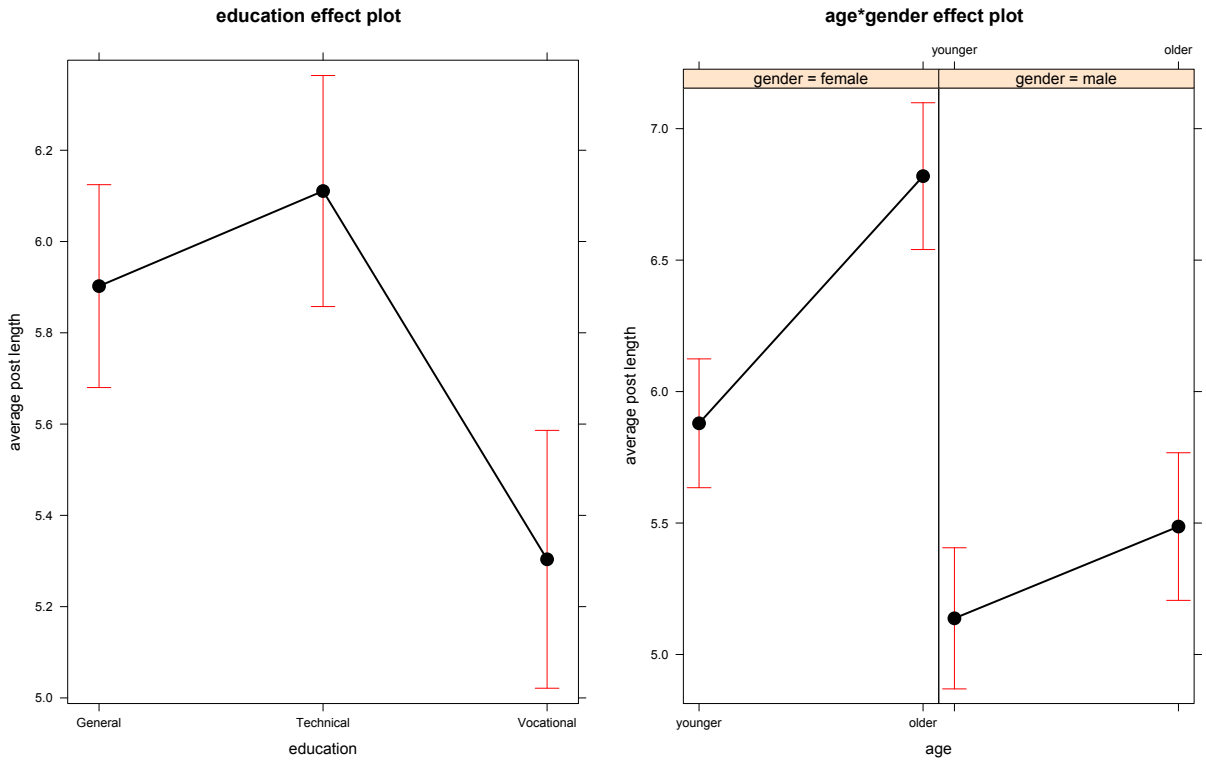


Figure 1: Average post length: Effect plot

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 utterance 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, and thus simultaneously influence the linguistic variable. 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.

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 non-word elements such as emoji, a longer post length implies more *lexical* expression (which of course does not exclude typographic 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 standard Dutch 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 regards 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, 9). However, they experimented with formal writing tasks. The observed gender difference – i.e. 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, 20-21; Newman et al. 2008, 213, and references therein) and average text length (Verheijen & Spooren 2017, 9).

	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 5: Average post length: Fixed effects (reference category: younger girls in General Education)

	Chisq	Df	Pr(>Chisq)	Signif.
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	***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 6: Average post length: Anova

5.2. 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. This finding corresponds to previous results (Lin 2007, 21; Mehl & Pennebaker 2003, 865; Newman et al. 2008, 213-214, 223, and references therein). 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. Interestingly, the combination of this result and the findings on utterance length (Section 5.1) 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.

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

Table 7: Average token length: Fixed effects (reference category: girls)

	Chisq	Df	Pr(>Chisq)	
gender	6.2038	1	0.01275	*
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 8: Average token length: Anova

5.3. Lexical richness

With respect to lexical richness, expressed as Guiraud type/token-ratio (Guiraud TTR), age and educational track are significant predictors (see Tables 9 and 10 for the fixed effects and the Anova). Older teenagers produce lexically richer texts than younger adolescents (sig.), which confirms the assumption (and previously attested pattern) that people's vocabulary expands with age (see Sankoff & Lessard 1975, 689, for results on spoken language). This age pattern with respect to active vocabulary production 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, 9). For informal speech, Sankoff and Lessard (1975) conducted a similar linear regression analysis with lexical richness (TTR) as response variable. Although their operationalization is somewhat different from ours (i.e. no random effects are included, uncorrected TTR is used, age is treated as a continuous variable and education as blocks of 4 years of education), it is interesting to compare results. 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, 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.

Our results reveal a somewhat surprising educational pattern, with students in General Education producing less lexical variation than students in Technical Education[7]. As the focus on language teaching is stronger 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

[7] 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).

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, 689, and Verheijen & Spooren 2017, 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 Section 2). 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. In addition, it indicates that teenagers in practice-oriented educational tracks, who attend a school system with a weaker focus on formal writing skills, might, in an informal social media setting, actually outperform their more theory-oriented peers with respect to verbal eloquence.

	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 9: Lexical richness: Fixed effects (reference category: younger teenagers in General Education)

	Chisq	Df	Pr(>Chisq)	
age	10.685	1	0.00108	**
education	6.953	2	0.03092	*
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 10. Lexical richness: Anova

5.4. Lexical expression of sentiment

The final linguistic variable that is analyzed quantitatively is the lexical expression of sentiment 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; 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, i.e. the teenagers' age, gender and 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).

	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 11: Average polarity (abs. value): Fixed effects (reference category: younger girls in General Education)

	Chisq	Df	Pr(>Chisq)	
age	13.918	1	0.000191	***
gender	38.833	1	4.617e-10	***
education	29.022	2	4.988e-07	***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 12: Average polarity (abs. value): Anova

Strongly related[8] to the variable of polarity, is the (lexically expressed) subjectivity of a text. Again, all three social predictors significantly influence this linguistic variable (see Tables 13 and 14 for the fixed effects and the Anova). Similar patterns can be found as before, with older teenagers, girls and theoretically educated students producing more lexically subjective messages, once again with a gradual increase from Vocational to Technical to General Education (with all pairs significantly differing from one another).

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, 51; Newman et al. 2008, 223, 229; Schwartz et al. 2013, 9; Wolf 2000, 831). We refer to Section 6 for a discussion of *typographic* rather than lexical expressiveness.

	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 13: Subjectivity: Fixed effects (reference category: younger girls in General Education)

	Chisq	Df	Pr(>Chisq)	
age	26.672	1	2.411e-07	***
gender	38.818	1	4.652e-10	***
education	29.742	2	3.479e-07	***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 14: Subjectivity: Anova

[8] Though 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, 3569) make a distinction between *sick* meaning 'sadistic', and *sick* meaning 'ill'. While the former is both very negative and very subjective, the latter is fairly objective, but is still connected to a negative sentiment.

5.5. 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 e.g. emoji. Below, we discuss the words and associated topics that appear to be popular among all groups of teenagers. Next, we present more detailed findings per social group.

Manual inspection of different groups of youths' top-500 words reveal that the most prominent topics discussed on social media are nearly identical for teenagers with different socio-demographic 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, for all groups of participants, 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, 121-122; Schwartz et al. 2013, 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 words for younger and 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 frequently used words by boys and girls reveals very similar tendencies too, with largely the same topic preferences (i.e. school, family and friends, and social media and communication). However, in spite of a strong common ground, 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 social conflict, which are prominent for girls only. E.g.: *ruzie* ('quarrel'), *praten* ('to talk'), *wenen* ('to cry'), *mis* ('miss'), and *lachen* ('to laugh'). Taboo words and 'tough' words, on the other hand, are only favored by boys. Examples are *fucking, shit* and *gast* ('dude'). 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 in previous studies. This holds for instance for the male preference for swear words (Mehl & Pennebaker 2003, 866; Newman et al. 2008, 213-214, 223, and references therein; Schwartz et al. 2013, 8) and for lexemes related to football and gaming (Argamon et al. 2009, 121; Schwartz et al. 2013, 8) versus the female preference for words referring to social or psychological processes (Newman et al. 2008, 223). However, related research suggests that 'family and friends' is a prominent topic in female discourse only (Argamon et al. 2009, 121; Schwartz et al. 2013, 8), whereas our data do not reveal a distinction between both genders in this respect.

As for educational variation, finally, the analysis once again reveals more similarities than divergence among the different groups of teenagers, i.e. students in distinct educational tracks. Obviously, the same topics prevail: school, family and friends, and social media and

communication. 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 'tough' 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 'tough' words is present in the top-500 lexemes for the two more practice-oriented tracks. This difference might indicate an attitudinal difference, i.e. this particular vocabulary seems to hold a higher (dynamic/modern) prestige in the eyes of these students compared to their peers in General Education. Strikingly, in addition to these 'tougher' 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 social words in general (e.g. *samen* 'together', *praten* 'to talk', *mis* 'miss', *helpen* 'to help', *ruzie* 'quarrel', *voel* 'feel', *pijn* 'pain', *kwaad* 'angry'). 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.

6. Normalized versus non-normalized 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 – i.e. familiarity with the characteristics and conventions of informal online communication, and the inclusion of non-lexical realizations of the above mentioned phenomena. In this section, we will compare both types of literacy in the adolescents' instant messages.

Example (11), for instance, is a social media message consisting solely of emoticons and emoji. It clearly is a highly expressive utterance: it is subjective rather than neutral, and conveys 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. In addition, the author of (11) demonstrates his or her familiarity with a whole range of emoji, which may indicate a strong exploitation of digital literacy. It would be relevant to compare the adolescents' reliance on their digital repertoire to their reliance on more traditional literacy in a social media setting.

(11) 😍😊☺️😘💋❤️

Therefore, for each of the five variables discussed above, we compared the normalized social media posts, i.e. the "standardized" Dutch texts, to the original ones, i.e. the authentic texts that include non-standard features and chatspeak 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. 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, i.e. girls producing more diverse messages than boys. This could suggest that girls either use more alternative (non-standard) spellings for the same word (e.g. spelling errors, but also deliberate, expressive, typographic manipulations, such as vowel repetition), or that they use more non-word elements (such as emoticons) – both of these assumptions are confirmed in previous research, see e.g. Hilte et al. (forthcoming, 2018c). 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.

The most striking differences, however, concern the expression of sentiment. The teenagers' *lexical* expression was compared to their *typographic* expression. Some illustrations of typographic chatspeak markers that express emotion can be found below. In example (12), the use of capital letters (which mimics shouting) and the repetition of the exclamation mark both intensify the expression of anger. In example (13), the lengthening of the vowel (which mimics oral stress) increases the expression of enthusiasm.

(12) *IK BEN ECHT BOOS!!!!* 'I AM REALLY ANGRY!!!!'
(13) *suuuuuuper leuk!* 'suuuuuuper nice!'

In our previous research on the present corpus (Hilte et al. forthcoming), analyses with Poisson regression models 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 paper: the lexical analyses discussed in Section 5 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. (forthcoming) 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 Section 5).

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. forthcoming). So girls definitely invest more in the expression of social an emotional engagement and they do so by all means, i.e. 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 Section 5) 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. forthcoming). In other words, the lexical expression of sentiment is more favored by theory-oriented students, whereas typographic new media markers like emoji are more popular amongst their peers in the Vocational track.

7. Conclusion

The present study on teenagers' online writing practices specifically focused on more general linguistic variables that have received minor attention in research on online language use compared to the prototypical (e.g. typographic) chatspeak 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 aimed at analyzing 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 socio-demographic profiles (in terms of age, gender and educational track) might use these repertoires to a different extent.

We 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 (top favorite words and associated topics) and divergent writing practices for different groups of youths for other features (average word and post length, lexical richness, lexical expression of sentiment). While some subtle nuances could be noted depending on the authors' profile, prominent topics in all adolescents' instant messages were school, family, friends, communication and social media. This significant overlap in top favorite words suggests that all teenagers, regardless of their specific age, gender or educational background, largely share the same interests and preoccupations, and discuss these in their online conversations.

The authors' profile did appear to have a significant impact on average word and post length and on the lexical richness of social media posts (analyzed in the normalized version of the corpus). Higher scores for these three variables – i.e. the production of longer words, longer utterances and more lexically diverse utterances – may be indicative of a stronger traditional

literacy or a stronger verbal orientation. While such traditional literacy might be expected to increase with age, and potentially be stronger for students in more theory-oriented tracks (where the focus on language teaching is stronger, and correct *formal* writing is a more prominent learning objective), our findings indicate that these expectations are 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 stronger traditional literacy skills for older adolescents 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 stronger 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 stronger 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 e.g. a higher rate of alternative spellings (including genuine mistakes as well as deliberate manipulations) 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 non-standard Flemish renderings of general Dutch words, actual dialect lexemes that are onomasiological alternatives for their Dutch equivalents 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 non-standard words, spelling and typography in their online writing (Hilte et al. forthcoming, 2018a). So to some extent, more heavy reliance on a non-standard lexical repertoire might be an explanatory factor.

This paper combined two perspectives on online writing practices: a more traditional, largely lexical, focus (examining normalized versions of the social media texts) and a digital media focus (examining the original texts, including non-standard digital chatspeak markers). The analyses revealed a clear interaction between traditional verbal expression and eloquence and the exploitation of the new media repertoire with its typographic features in teenagers' online writing practices, 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 suggests that the former groups are more verbally-oriented, whereas the latter ones are more

inclined to express themselves using digital media-specific 'tools'. So this discrepancy reveals a notable difference with respect to the expression of emotional and social involvement in online writing between specific adolescent 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 finetuning of certain lexical features. It might be relevant to examine alternative operationalizations of lexical richness, as they may help us understand the linguistic phenomenon from different – complementary – perspectives. A recommended strategy consists in taking frequency distributions into account, e.g. by adding a distinction between more common and more difficult/sophisticated vocabulary, between function and context words, or by adding a general frequency measure (Malvern & Richards 2012, 1; Read 2000; van Hout & Vermeer 2007; Vermeer 2000, 79). On a content-level, it is advised to control for conversation topic (and potential topic changes within a conversation): 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, or whether the interlocutors are familiar with it or not) may significantly impact lexical diversity (van Hout and Vermeer 2007, 129; Vermeer 2000, 77; Yu 2009, 254). However, the influence of topic on the dataset discussed in the present paper most probably is quite insignificant, since our findings reveal that all teenagers largely discuss 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, Speelman & Geeraerts 2014, 95). Finally, correcting for phenomena such as homonymy and polysemy might add more nuance too (see e.g. De Hertog, Heylen & Speelman 2014).

Finally, with respect to the expression of sentiment on social media, it would be highly relevant to expand existing tools by incorporating typographic chatspeak markers that serve an expressive purpose (e.g. emoji, character repetition). 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.

References

Androutsopoulos, Jannis. (2011). Language change and digital media: A review of conceptions and evidence. In Tore Kristiansen, & Nikolas Coupland (Eds), *Standard languages and language standards in a changing Europe* (pp. 145-161), Oslo: Novus.

Argamon, Shlomo, Moshe Koppel, James W. Pennebaker, & Jonathan Schler. (2009). Automatically profiling the author of an anonymous text. *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. *Journal of Sociolinguistics* 18(2), 135-160.

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

Bates, Douglas, Martin Maechler, Ben Bolker, & Steven Walker. (2017). Package 'lme4'. Retrieved from: https://cran.r- project.org/web/packages/lme4/lme4.pdf

Coates, Jennifer. (1993). *Women, men and language. A sociolinguistic account of sex differences in language.* London: Longman.

[FMET] Flemish Ministry of Education and Training. (2018). *Statistisch jaarboek van het Vlaams onderwijs. Schooljaar 2016-2017.* Brussels: Department of Education and Training.

De Clercq, Orphée, Sarah Schulz, Bart Desmet, Els Lefever, & Véronique Hoste. (2013). Normalization of Dutch user-generated content. In Galia Angelova, Kalina Bontcheva, & Ruslan Mitkov (Eds), *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013* (pp. 179-188), Hissar: Incoma.

De Decker, Benny. (2014). *De chattaal van Vlaamse tieners: Een taalgeografische analyse van Vlaamse (sub)standaardiseringsprocessen tegen de achtergrond van de internationale chatcultuur*. Antwerp: University of Antwerp (doctoral thesis).

De Decker, Benny, & Reinhild Vandekerckhove. (2017). Global features of online communication in local Flemish: Social and medium-related determinants. *Folia Linguistica* 51, 253-281.

De Hertog, Dirk, Kris Heylen, & Dirk Speelman. (2014). Stable lexical marker analysis: A corpus-based identification of lexical variation. In Augusto Soares da Silva (Ed.), *Pluricentricity: Language variation and sociocognitive dimensions* (pp. 127-142), Berlin / Boston: Walter de Gruyter.

De Smedt, Tom, & Walter Daelemans. (2012a). Pattern for Python. *Journal of Machine Learning Research* 13, 2031-2035.

De Smedt, Tom, & Walter Daelemans. (2012b). "Vreselijk mooi!" (terribly beautiful): A Subjectivity Lexicon for Dutch Adjectives. In *LREC* (pp. 3568-3572).

Han, Bo, Paul Cook, & Timothy Baldwin. (2013). Lexical normalization for social media text. *ACM Transactions on Intelligent Systems and Technology (TIST)* 4(1), 5.

Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans. (forthcoming). Modeling adolescents' online writing practices. The sociolectometry of non-standard writing on social media. *Zeitschrift für Dialektologie und Linguistik.*

Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans. (2018a). Adolescents' social background and non-standard writing in online communication. *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. *International Journal of Society, Culture & Language* 6(2), 73-89.

Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans. (2018c). Expressive markers in online teenage talk: A correlational analysis. *Nederlandse Taalkunde* 23(3), 293-323.

Holmes, Janet. (1992). *An introduction to sociolinguistics.* London / New York: Longman.

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

Malvern, David, & Brian Richards. (2012). Measures of lexical richness. *The encyclopedia of applied linguistics*. Accessed online: https://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal0755

Mehl, Matthias R., & James W. Pennebaker. (2003). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality & Social Psychology* 84, 857-870.

Newman, Matthew L., Carla J. Groom, Lori D. Handelman, & James W. Pennebaker. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes* 45(3), 211-236.

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

Ruette, Tom, Dirk Speelman, & Dirk Geeraerts. (2014). Lexical variation in aggregate perspective. In Augusto Soares da Silva (Ed.), *Pluricentricity: Language variation and sociocognitive dimensions* (pp. 103-126), Berlin / Boston: Walter de Gruyter.

Sankoff, David, & Réjean Lessard. (1975). Vocabulary richness: A sociolinguistic analysis. *Science* 190, 689-690.

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. *PLoS ONE* 8(9), e73791.

Singh, Sameer. (2001). A pilot study on gender differences in conversational speech on lexical richness measures. *Literary and Linguistic Computing* 16(3), 251-264.

van der Goot, Rob, & Gertjan van Noord. (2017). MoNoise: Modeling noise using a modular normalization system. ArXiv preprint. Retrieved from: https://arxiv.org/pdf/1710.03476.pdf

van Hout, Roeland, & Anne Vermeer. (2007). Comparing measures of lexical richness. In Helmut Daller, James Milton, & Jeanine Treffers-Daller (Eds), *Modelling and assessing vocabulary knowledge* (pp. 93-115), Cambridge: Cambridge University Press.
Retrieved from: https://www.researchgate.net/publication/254801850_Comparing_measures_of_lexical_richness (in-text references refer to pagination of online version)

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. *Reading and Writing* 23(6), 719-733.

Verheijen, Lieke, & Wilbert Spooren. (2017). The impact of WhatsApp on Dutch youths' school writing. In Egon W. Stemle, & Ciara R. Wigham (Eds), *Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities (cmccorpora17), 3-4 October 2017, Eurac Research, Italy* (pp. 6-10), Bolzano.

Vermeer, Anne. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing* 17(1), 65-83.

Wolf, Alecia. (2000). Emotional expression online: Gender differences in emoticon use. *Cyberpsychology & Behavior* 3(5), 827-833.

Yu, Guoxing. (2009). Lexical diversity in writing and speaking task performances. *Applied Linguistics* 31(2), 236-259.

CHAPTER 8

This chapter has been submitted as a journal article. Reference:

Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans. (submitted). Adolescents' perceptions of social media writing: Has non-standard become the new standard?

Adolescents' perceptions of social media writing:

Has non-standard become the new standard?

Abstract

The present study examines adolescents' attitudes and perceptions with respect to writing practices on social media. It reports the findings of a survey conducted among 168 Flemish high school students with various socio-demographic profiles. The survey examines linguistic attitudes and awareness of sociolinguistic patterns in computer-mediated communication, as well as relevant language skills. Moreover, the present paper uniquely combines the study of both adolescents' perceptions and their production of informal online writing, as the participants' responses to the survey are compared to their peers' actual online writing practices.

The respondents appear to have a fairly accurate intuition with respect to age and gender patterns in social media writing, but much less so with respect to education-related patterns. Furthermore, while typical chatspeak phenomena are easily identified as such, ordinary spelling mistakes often are not. Strikingly, the teenagers do not claim a high standard language proficiency, although they do state to care about standard language use in formal contexts. Finally, some significant differences were found between participants with distinct socio-demographic profiles, e.g. girls and highly educated teenagers appear to be more sensitive to the potential negative connotations of linguistic features and that sensitivity seems to increase with age.

Keywords: Computer-mediated communication, social media, adolescents, perception, survey

1. Introduction

The genre of informal computer-mediated communication (CMC) is assumed to have led to a "pluralisation of written language norms" (Androutsopoulos 2011, 146; see also Grondelaers et al. 2016, 130). While informal online writing has been characterized in terms of "linguistic whateverism" (Baron 2008, 169) and the impression has often been created that 'anything goes', CMC-researchers seem convinced that the genre "has its own rules rather than that it follows no rules whatsoever" (Verheijen 2013, 584). However, we do not know how these rules are perceived or evaluated by their users. Many studies have laid bare the prototypical features and communicative strategies of informal online writing and the way these are conventionalized, but there is hardly any research with respect to the perception of these conventions or implicit rules. Therefore, the present research wants to find out how the most ardent users of social media, i.e. the adolescent generation, perceive and evaluate informal online writing conventions. It does so by comparing youths' meta-(socio)linguistic awareness and linguistic attitudes with findings on their actual online writing practices. For the latter, we rely on our previous research on Flemish adolescents' informal CMC and on numerous related studies (see below).

The paper is structured as follows: Section 2 summarizes related research. Section 3 deals with the experimental design of the survey and the collection of the corpus that serves as a reference point for the data analysis. In Section 4, the results of the survey are discussed. Finally, Section 5 presents the conclusions.

2. Research context

The present study primarily concerns attitudinal research, focusing on teenagers' attitudes with respect to their peers' online writing practices. When referring to attitudes, we envisage "an evaluative orientation" towards a linguistic variety or phenomenon (Lybaert 2014, 22). This evaluation generally has a cognitive and an affective dimension: people have knowledge and beliefs with respect to language varieties and these evoke (positive, negative or mixed) feelings. We will primarily analyze the identification and appreciation of CMC features and study what kind of features are attributed to which social groups.

The linguistic genre that is the main subject of this paper are adolescents' informal online conversations, which tend to deviate from formal writing practices in several respects. Some of the deviations result from the integration of spoken language features in written CMC, whereas others are more typical of digital media. Most prototypical chatspeak features can be described in terms of three implicit 'rules' of informal CMC captured by e.g. Androutsopoulos: the principles of expressive compensation, orality and brevity (2011, 149). The principle of brevity leads to a maximization of typing speed, e.g. through the use of abbreviations. The orality maxim relates to the fact that the register in many forms of informal CMC is to a large extent conceptually oral, reflecting oral communication rather than classical written communication. Symptomatic in this respect is e.g. the use of regional features. 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 (e.g. emoticons can represent facial expressions). For an extensive overview of the linguistic properties of chatspeak, see e.g. Hilte et al. (2018b), Verheijen (2015) or Varnhagen et al. (2010).

Because of the omnipresence of these deviations from formal writing norms in youths' online writing, many people worry about the effects of CMC on youths' (formal) language skills and those concerns have been widely reflected in negative media attention for the genre (Vandekerckhove & Sandra 2016). Verheijen (2018, 36-44) offers an extensive overview of attitudinal research on the perceived effects of online writing on literacy, and concludes that mostly teachers and young adults seem pessimistic about the impact of CMC on literacy, whereas adolescents tend to have a more neutral opinion on the matter (40-41). While these studies are relevant for the present paper, the research focus is essentially different: they all examine people's attitudes with respect to the effect of CMC on formal writing skills, while we report on attitudes and perceptions with respect to the CMC genre itself. However, the evaluations that predominate in the studies discussed by Verheijen (2018) are most telling

with respect to the appreciation of online writing practices, since the concerns expressed by the respondents implicitly reveal a predominantly negative evaluation of (at least some) characteristics of informal CMC. Moreover, the finding that adolescents tend to report more positive attitudes (with respect to the impact of CMC on traditional literacy) is highly relevant too, since the present study focuses on this age category, but, once again, we want to know how this age group evaluates CMC writing in itself. Nevertheless, we want to add that the results of Verheijen and Spooren (2017, 6) suggest that there is no solid ground for pessimism in terms of effects on literacy, since the experiments revealed no (short-term) effect of WhatsApp on youths' school writing.

A survey into the perception and evaluation of CMC conventions inevitably entails an enquiry into the attitudes with respect to more standard ways of writing too. Therefore, part of the survey focuses on Flemish adolescents' appreciation of and self-estimated proficiency in formal standard Dutch. The concepts of standardization and destandardization have been widely discussed in variational linguistic research in the past decade (see e.g. Kristiansen & Coupland 2011). A major question in the present-day debate is "whether standard languages […] are destandardizing, as is commonly held, or whether it could be the case that the 'classical' standardness criteria […] have become too narrow to fit present-day standard language dynamics." (Grondelaers et al. 2016, 143). Grondelaers et al. (2016) point to a "new social and linguistic reality" (143), marked amongst others by digitalization processes that led to changing linguistic practices which "pluralized language norms and further amplified the importance of identity" (130). The authors revisit classical criteria for standard languages (e.g. Auer 2011), signaling an "internal change in the concept of prestige" (134): they claim that apart from traditional (overt) prestige, new types of superiority criteria have emerged and have become increasingly important, such as "dynamism" and "media cool" or "modern media prestige" (Grondelaers et al. 2016, 119, 132; Kristiansen et al. 2005, 12). This coolness-factor in particular may impact youths' online writing, since adolescents tend to intensively engage in identity construction and since self-profiling is an inherent part of most social media communication. In this respect, all kinds of CMC conventions and chatspeak features are potentially useful "linguistic tool[s] for modern self-portrayal" (Grondelaers et al. 2016, 130). However, different types of features might be indexical of different social connotations. While digital vernacular features that are related to the principles of brevity and expressive compensation (see above) might evoke connotations of informality and trendiness, orality markers reflecting more traditional non-standardness (e.g. dialect) might evoke connotations of localness and toughness (see Hilte et al. forthcoming, for distinct preference patterns for old vs. new vernacular amongst different groups of adolescents). With respect to the indexicality of standard language, we note that while standard language is seldom a neutral variety (although it is often claimed to be so), it certainly is not in informal CMC, where its abundant use might trigger "traditional superiority perceptions which are at odds with the local coolness demands" (Grondelaers et al. 2016, 138). The present study can contribute to the debate on standard language ideologies and the evaluation of (non-)standard language by analyzing youths' opinions with respect to the appropriateness and importance of

standard Dutch in different communicative settings, ranging from informal social media contexts to formal school contexts.

3. Experimental design

This section is devoted to the experimental design of the study. First, the design of the survey is discussed (Section 3.1), and next, the group of participants is described (Section 3.2). Finally, in Section 3.3, the social media corpus is introduced from which the examples in the survey are extracted. Moreover, this corpus will serve as the reference point for the comparison of adolescents' perceptions and sociolinguistic awareness with their actual online linguistic practices.

3.1. Design of the survey

We created a survey to complement our previous research on teenagers' *production* of informal online writing (see Hilte et al. 2018a, 2018b, 2018c, and forthcoming) with findings on their attitudes and *perceptions* with respect to this linguistic register. The respondents were recruited in ten class groups in four high schools. They each had a computer at their disposal to fill in the online survey. Participation was voluntary and anonymous; participants were not asked to enter their name or class group. However, they did have to enter general profile information (e.g. age, gender). For more information on the respondents, see Section 3.2.

The survey consisted of multiple question blocks focusing on linguistic attitudes and perceptions and to a minor extent also on language skills. Below, each question block is described and illustrated. The order in which the blocks were presented to the participants was randomized each time (i.e. all students answered the same questions but in different, random, orders). The figures with screenshots from the survey show the original question in Dutch and an English translation that was added for the purpose of this paper only.

Blocks 1-3: Intuitive author profiling tasks

The first question blocks consisted of three distinct intuitive author profiling tasks, each based on five authentic chat messages extracted from the corpus (described in Section 3.3). In the first task, the participants had to guess the author's gender for each of the five messages. They could check one out of three boxes: 'girl', 'boy', or 'I don't know'. Whenever they checked the 'girl' or 'boy' box, they were free but not obliged to write down their argumentation. The second block contained a similar task concerning age profiling: the participants were asked whether the authors of five chat messages were either 13-16 or 17-20 years old. The third block concerned education profiling. For five chat messages, the participants had to guess which of the three main Belgian secondary education tracks the authors attended: General (theory-oriented), Vocational (practice-oriented) or Technical

Secondary Education (hybrid – see Section 3.2 for a detailed description). Just like for gender profiling, in the age and education profiling tasks 'I don't know' was a valid response too. Similarly, the participants were free but not obliged to explain their reasoning. Figure 1 shows one of the gender profiling questions.

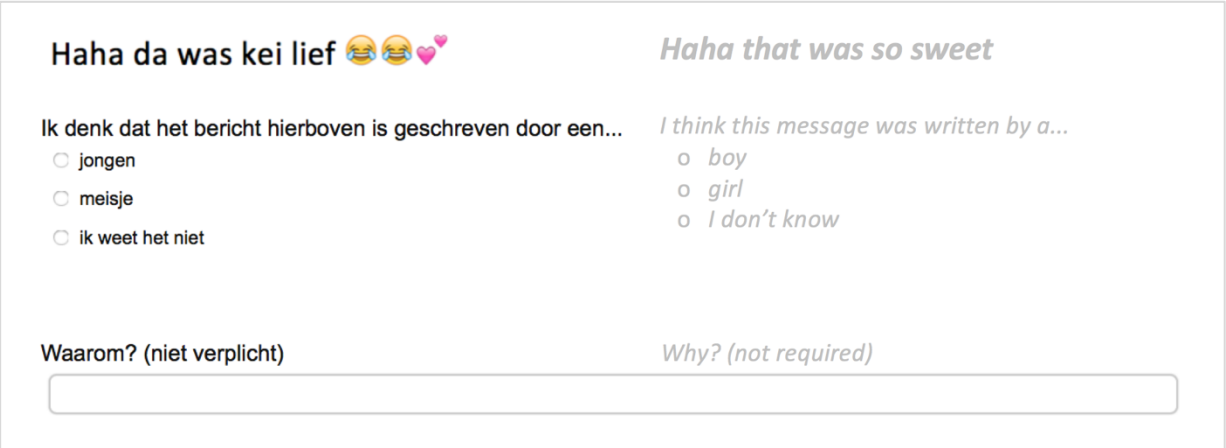


Figure 1: Example from the survey: A gender profiling task

These question blocks served two purposes. The first purpose was to verify whether the participants were able to distinguish between the writing patterns of different socio-demographic groups of teenagers. The second purpose was to obtain insight in the intuitive factors that determined participants' decision-making, and to compare these to sociolinguistic patterns that were attested in the reference corpus or in related research.

Block 4: Statements on author profiling

The fourth question block was related to the tasks described above. It contained statements on potential linguistic differences in chat messages written by adolescents with different social profiles in terms of age, gender and educational track. The participants had to indicate the degree to which they (dis)agreed with the statements on a 5-point Likert scale, ranging from complete disagreement to full agreement. An example is shown in Figure 2.

Figure 2: Example from the survey: Statement on gender profiling

This question block was added to obtain insight in the teenagers' awareness of sociolinguistic variation in social media writing. The participants' replies will be compared to their performance in the actual profiling tasks, as potential correlations or discrepancies may emerge.

Block 5: Correction or 'conversion' task

The longest question block was a language correction or 'conversion' task in which the participants were presented with eight chat messages written by their peers. Each message could contain one deviation from formal standard writing, or none. Some deviations were straightforward linguistic errors (e.g. spelling mistakes), others represented prototypical chatspeak markers that generally are not integrated in formal writing (e.g. a non-standard abbreviation that is common in online writing). The participants first had to decide whether the message corresponded to standard Dutch writing norms or not. It was emphasized that the standardness of the message was to be evaluated regardless of the social media context: the students had to check whether the sentence would be acceptable in e.g. a school exam. In case of a positive answer, they proceeded to the next item that had to be judged. If they answered 'no', they had to indicate on a 5-point Likert scale to which extent the deviation would bother them in a chat message on Facebook or WhatsApp. Finally, they were asked to convert the sentence into its standard equivalent. This allowed us to verify whether they had spotted the actual error and were capable of producing the standard equivalent. An example is shown in Figure 3. The utterance *Jij bent sgattig* ('You are cute') contains a non-standard spelling: *schattig* ('cute') is spelled as *sgattig.* This cluster reduction from *ch* (/X/) to *g* is a common spelling deviation in Flemish online teenage talk.

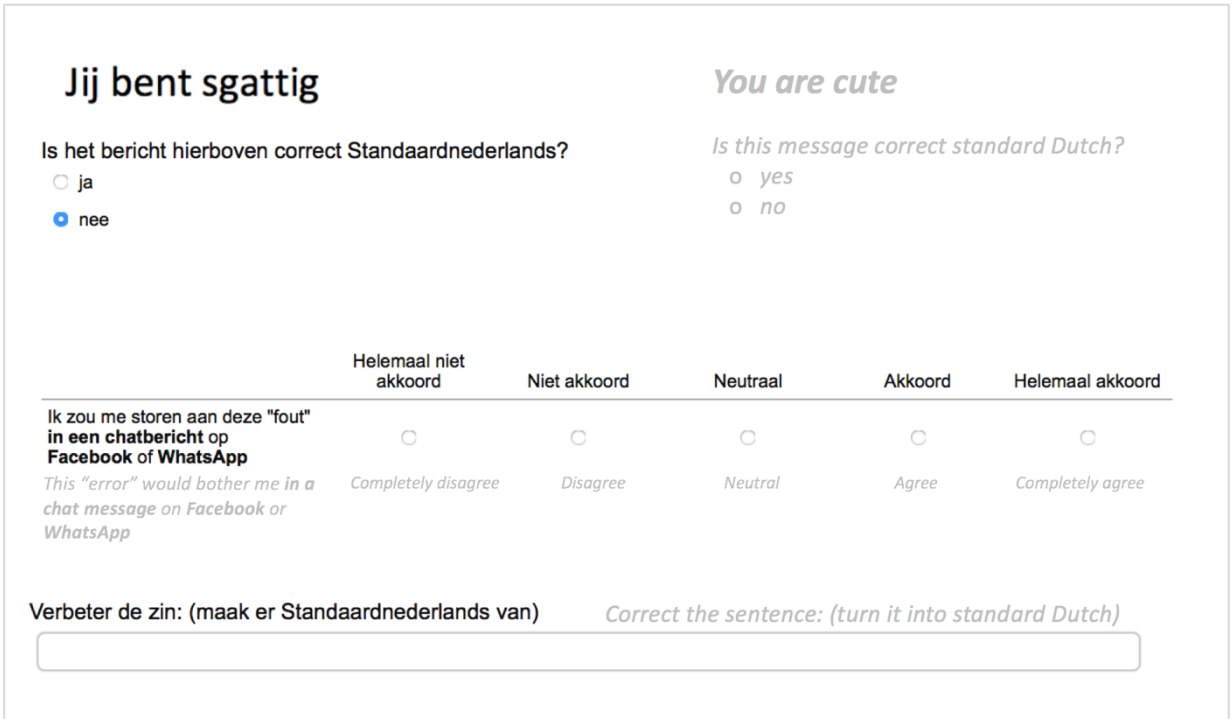


Figure 3: Example from the survey: Correction task[1]

[1] The 'would bother me'-statement and the correction field only appear if the participant answers 'no' to the first question.

This block's purpose was to examine teenagers' detection and perception of non-standard writing practices on social media as well as their proficiency in formal standard Dutch. More specifically, as the individual questions and example messages contain different types of errors and chatspeak markers, we want to verify which deviations from the formal written standard are still perceived as 'incorrect' by adolescents. The evaluative questions can reveal which of these errors – even though they are recognized as incorrect from a formal standard Dutch perspective – are (fully) accepted in social media interactions, and which ones are not.

Block 6: Statements on standard Dutch

In the sixth block, the participants were asked to indicate their (dis)agreement on a 5-point Likert scale with several statements on the importance of standard language (proficiency) in different communicative contexts, ranging from school or professional contexts to the informal setting of peer group communication on social media. An example is shown in Figure 4.

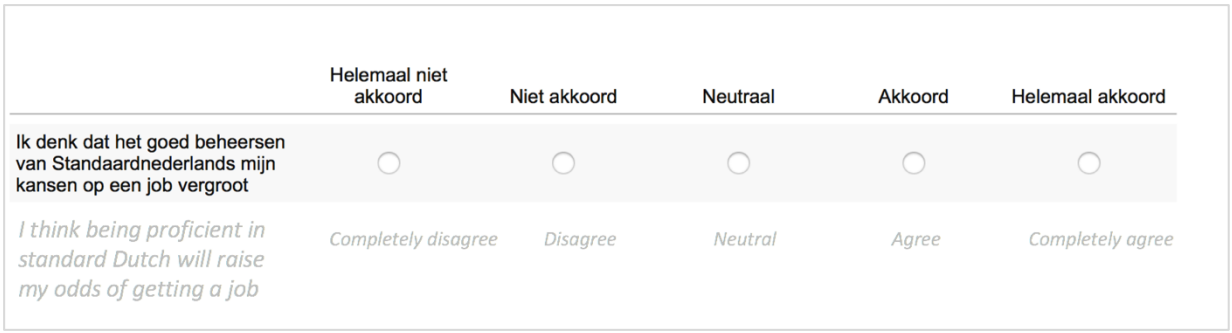


Figure 4: Example from the survey: Statement on the importance of standard Dutch proficiency

The answers to these questions show to what extent adolescents have appropriated mainstream standard language ideologies.

Block 7: The indexicality of linguistic and typographic features

In the seventh block, participants were presented with chat messages written by their peers and had to indicate how friendly or kind the authors sounded. The same utterances re-occurred multiple times, with slight stylistic modifications, as is illustrated in Figure 5.

Figure 5: Example from the survey: Implicit statements on feature indexicality

The answers to these questions reveal the indexicality of particular non-verbal linguistic features and chatspeak markers for the adolescent generation, and may enhance our understanding of the acceptability of these features.

Block 8: Ranking chat messages

In the final block, the participants were asked to rank 13 authentic chat messages in terms of how likely they were to write such utterances themselves. The responses should reveal what kind of chatspeak features or strategies the adolescents identify with and from which features they dissociate themselves. Consequently, they might give us an idea of the type of features that are used for identity construction on social media by particular groups of teenagers. A selection of the messages that were to be ordered can be seen in Figure 6.

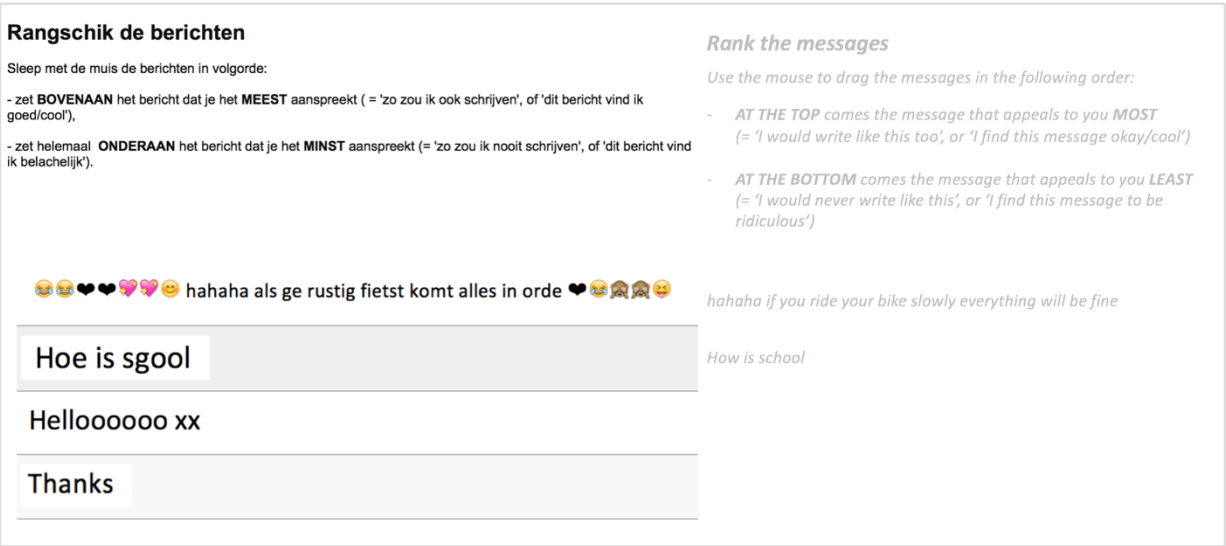


Figure 6: Example from the survey: ranking task

3.2. Participants

The survey was conducted among 168 Flemish[2] teenagers attending four different secondary schools in the central province of Antwerp. The participants were between 15 and 20 years old and were all in the final three years of secondary education when the survey was conducted (i.e. in 2018). They were all students in one of the three main types of Belgian secondary education (FMET 2017, 10):

- General Secondary Education: theory-oriented track that prepares students for higher education.
- Technical Secondary Education: track with a strong theoretical and practical component, and a specific focus on technical courses. Students can either start their professional life after graduating or proceed to higher education.
- Vocational Secondary Education: practice-oriented track that prepares students for a specific (often manual) profession. The focus is on acquiring skills rather than on theoretical knowledge. This degree does not grant direct access to higher education.

Table 1 presents an overview of the participants in terms of their age, gender and educational track. We filtered out data from respondents who did not complete the entire survey, and from one particular student who had made up silly answers for most of the questions. In order to deal with the imbalances with respect to gender and education, we will carry out analyses to examine the impact of these social variables on the teenagers' replies.

		Educational track			
		General	**Technical**	**Vocational**	**Total**
Gender	**Girls**	25	53	24	102 (61%)
	Boys	22	24	20	66 (39%)
	Total	47 (28%)	77 (46%)	44 (26%)	**168**

Table 1: Distribution of the survey participants

3.3. Corpus

The chat messages used in the survey were extracted from a large social media corpus collected by the authors of this paper. The corpus has been extensively described and analyzed in previous work (see e.g. Hilte et al. 2018a, 2018b, and forthcoming). It consists of 434 537 social media posts (>2.5 million tokens) written by 1384 secondary school students in the three educational tracks described in Section 3.2, 13 to 20 years old. Almost all students (96%) live in the central Flemish province of Antwerp. The posts are private instant messages produced in Dutch on Facebook Messenger and WhatsApp. The vast majority of the tokens

[2] I.e. living in Flanders, the Dutch-speaking part of Belgium.

(87%) was produced between 2015 and 2016. Table 2 presents an overview of the distributions in the corpus.

Variable	Variable levels	Tokens
Educational track	General	739 831 (29%)
	Technical	1 151 684 (46%)
	Vocational	639 839 (25%)
Gender	Girls	1 696 517 (67%)
	Boys	834 837 (33%)
Age	Younger teenagers (13-16)	1 360 898 (54%)
	Older teenagers / young adults (17-20)	1 170 456 (46%)
Total		2 531 354

Table 2: Distributions in the corpus

4. Results

In this section, the participants' responses to the survey are discussed and analyzed per question block.

4.1. Block 1-4: Author profiling tasks

The participants were presented with 15 authentic chat messages for which they had to guess the authors' gender, age or educational track. Figure 7 visualizes their performance and sociolinguistic awareness. The former refers to the performance in the profiling tasks (i.e. the percentage of correct responses per subtask for all participants). The latter indicates the extent to which youths are aware of and believe in the existence of these sociolinguistic patterns in social media writing. We recall that responses were to be made on a 5-point Likert-scale ('completely disagree', 'disagree', 'neutral', 'agree' and 'completely agree'). The 'awareness'-graph shows the combined percentage of 'agree'- and 'completely agree'-responses on the existence of gender-, age- and education-related linguistic patterns in chatspeak.

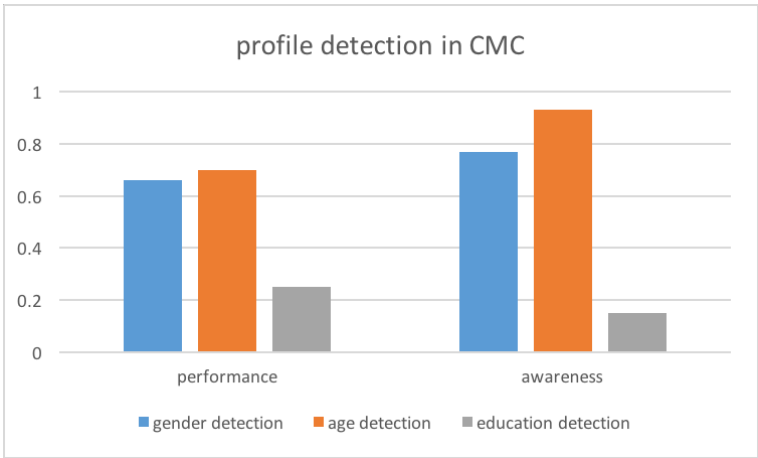


Figure 7: Survey results: Author profiling

As Figure 7 shows, respondents score much lower for education profiling compared to age and gender, in terms of both performance and awareness: the students do not only perform worse in guessing authors' educational track, the awareness or belief with respect to education-related linguistic differences is much weaker too. These differences are highly statistically significant: as for performance, the number of correct answers (versus incorrect and 'don't know'-replies) correlated significantly and strongly with the nature of the task (i.e. age, gender or education detection) ($p < .00001$, chisq. = 402.64, Cramer's V = 0.40). As for the awareness-statements, agreement with (versus disagreement with or a neutral opinion on) the existence of these sociolinguistic differences significantly and very strongly correlated with the nature of the task ($p < .000001$, chisq. = 412.72, Cramer's V = 0.70). Below, we discuss the different tasks in a more detailed way.

4.1.1. Gender profiling

Most participants (77%) agreed on the existence of linguistic gender differences in chat messages. Others (18%) had no opinion – i.e. they checked the 'neutral' box in the middle of the scale –, and only very few disagreed (5%). This rather strong sociolinguistic awareness was reflected in the performance in the detection task: 66% of the gender assignments were correct (versus 12% incorrect and 22% 'don't know'-replies). Additional tests with generalized linear mixed models (GLMMs) revealed no significant impact of participants' age, gender or educational track on their performance in the detection task or on their awareness of linguistic gender differences.

In the detection task, the participants were free to list the cues they used in their decision-making. Table 3 summarizes the arguments. The validity of the arguments rendered in bold and italics could be confirmed by our corpus data: for these features, a statistically significant gender difference was actually found in the reference corpus.

FEMALE	MALE
Chatspeak: ***- more emoticons, esp. hearts*** ***- letter reduplication***	**Chatspeak:** ***- fewer/no emoticons, esp. hearts***
Correctness: ***- correct language,*** incl. punctuation and capitals	**Correctness:** - incorrect language ***- slang, dialect***
Vocabulary: ***- 'omg'***	**Vocabulary:** - some dialect words
Tone of the conversation: - sweet, soft, kind messages - polite	**Tone of the conversation:** - rude, short messages - impolite
Character/nature girls apparent in text: - enthusiastic, overly happy	**Character/nature boys apparent in text:** - short, practical
Content: - gossip - sleepovers	**Content:** /

Table 3: Survey results: Adolescents' intuitions on linguistic gender differences in informal CMC

The participants used both stylistic and content-related features in their decision-making. With regards to content, they considered utterances about sleepovers or gossip as typically female, as well as enthusiastic or overly happy messages, whereas short, practical messages were seen as typically male. In addition, they linked messages that came across as sweet, soft and polite to female authors and rude, short and impolite messages to male authors. While we have not investigated these content- and tone-related dimensions in our corpus, the validity of many of these features as gender markers is confirmed by related quantitative studies. Two studies may serve as main points of reference: both Schwartz et al. (2013) and Argamon et al. (2009) examine corpora of English blog posts and report the most prominent and distinctive lexemes for male and female authors. Many of the female authors' top lexemes express strong enthusiasm (e.g. *excited*, *yay*) or a positive sentiment (e.g. *wonderful, amazing*) (Argamon et al. 2009, 121; Schwartz et al. 2013, 8). A female preference for positive emotion words has been attested in spoken conversations too (Mehl & Pennebaker 2003, 866). Furthermore, intensifiers, which "amplify and emphasize the meaning of an adjective or adverb" (Stenström et al. 2002, 139), were found to be used significantly more frequently by women or girls than by men or boys (Stenström et al. 2002, 142 and references therein). Schwartz et al. (2013, 8) indeed report that *super* and *so* are used abundantly in female blogs. Furthermore, the (reported) 'sweet' nature of female texts has been attested in corpora too, as love- and friendship-related lexemes appear to be typically female (e.g. *sweetheart*, *bestie*) (Argamon et al. 2009, 121; Schwartz et al. 2013, 8). In addition, women generally use more polite linguistic forms (Newman et al. 2008, 213 and references therein). Similarly, the reported harsher character of male texts can be related to a male preference for curse words reported in several studies, or to a male preference for anger-related words (see e.g. Mehl &

Pennebaker 2003, 866; Newman et al. 2008, 213-214 and references therein; Schwartz et al. 2013, 8).

As for stylistic features, the participants interpreted a more frequent use of emoticons and especially hearts as more typical of girls. In previous research, a higher frequency of emoticons has indeed been attested in female utterances (see e.g. Baron 2004, 415; Herring & Martinson 2004, 436; Kucukyilmaz et al. 2006, 282; Parkins 2012, 52; Schwartz et al. 2013, 8). Heart-emoticons in particular appear to be prominent in female CMC (Hilte et al. 2018c; Schwartz et al. 2013, 8). The same tendencies prevail in our corpus: emoticons are used significantly more often by girls than boys ($p < .0001$, chisq. = 7101.96, odds ratio = 1.71), and this tendency is even more outspoken for heart-emoticons ($p < .0001$, chisq. = 3985.79, odds ratio = 2.27). The survey participants also perceived the use of letter repetition (e.g. *s**oooo** nice*) as a typically female preference pattern that is manifest in our corpus too ($p < .0001$, chisq. = 1260.03, odds ratio = 1.73) and has been corroborated by previous research (see Hilte et al. 2018c for findings on older CMC-data; Schwartz et al. 2013, 8). With respect to the dimension standard versus substandard, the respondents considered the use of 'correct' standard language to be typically female, whereas substandard language (e.g. the use of dialect words) was characterized as male. In the corpus, the female chatters indeed use significantly more 'correct' standard Dutch words – although the effect size is not large – ($p < .0001$, chisq. = 410.58, odds ratio = 1.06) and the boys use significantly more non-standard Dutch lexemes (e.g. slang words or words that contain phonological dialect features) ($p < .0001$, chisq. = 1569.18, odds ratio = 1.15). In addition, sociolinguistic studies have reported on a male preference for 'old vernacular' or traditional non-standardness even amongst youths (see e.g. Hilte et al. forthcoming; Labov 1972; Labov 2001). Finally, the participants linked the acronym *omg* ('oh my god') to girls as well. *Omg* is one of the prominent female features reported by Schwartz et al. (2013, 8), and is strongly preferred by girls in our corpus too ($p < .0001$, chisq. = 603.55, odds ratio = 7.24).

4.1.2. Age profiling

Even more so than for gender, the participants showed a strong awareness of linguistic age differences in adolescents' online writing: most of them (93%) confirmed the presence of age patterns, there were hardly any neutral (4%) or negative (3%) responses. This awareness was also reflected in the students' performance in the detection task: 70% of the age assignments were correct, compared to 18% wrong and 12.5% 'don't know'-replies. The participants' profile did not significantly influence their performance.

The cues used by the participants are summarized in Table 4. Again, the relevance of the features rendered in bold and italics was corroborated by our CMC-data. For the features that are struck through however, we found no support in the corpus (e.g. no significant differences could be attested or the opposite pattern was found).

YOUNGER TEENAGERS (13-16)	OLDER TEENAGERS / YOUNG ADULTS (17-20)
Chatspeak: ***- many emoji (+ reduplication)*** ***- laughter ('hahahah')***	**Chatspeak:** ***- fewer/no emoji*** ***- fewer abbreviations***
Correctness: ***- spelling errors****, "ugly/childish" spelling,* often on purpose	**Correctness:** ***- correct, unabbreviated*** ~~- correctly spelled English words~~ - formal
Vocabulary:	**Vocabulary:** ~~- English words~~ - insults/curse words (often not meant negatively)
Character/nature younger teenagers apparent in text: - don't care about correct writing - laziness - desire to be cool	**Character/nature young adults apparent in text:** - think about what to say / how to say it
Content: - party less - care more about school	**Content:** - party more - care less about school

Table 4: Survey results: Adolescents' intuitions on linguistic age differences in informal CMC

Again, the participants used both content- and style-related features in their decision-making. With regards to content, they considered chat messages about partying to be typical of older adolescents, as they claim that younger teenagers "do not go to that many parties / are hardly allowed to go to parties". They also perceived texts in which the author appeared to care about school as more typical of younger adolescents. These features correspond more or less to the prominent age-related words reported by Schwartz et al. (2013) and Argamon et al. (2009), although some caution with respect to the comparability of the studies is needed: while we compare younger (aged 13-16) to older (17-20) high school students, Schwartz et al. compare teenagers (13-18) to college students (19-22), and Argamon et al. teenagers (13-17) to young adults in their twenties (23-27). Yet, in spite of these differences in research design, some interesting parallel tendencies can be noted: in the teenage group, school-related words are indeed more abundant (e.g. *homework*, *math*), and for the older group, more words about partying occur (e.g. *drunk*, *hangover*) (Argamon et al. 2009, 121-122; Schwartz et al. 2013, 10).

As for stylistic features, the survey participants linked a more frequent use of emoticons, onomatopoeic renditions of laughter and chatspeak abbreviations to young adolescents. These intuitions correspond to research findings: from previous studies (Hilte et al. 2018c; Verheijen 2015, 135-136; Verheijen 2016, 283-285) and our current corpus ($p < .0001$, chisq. = 11025.14, odds ratio = 1.82) it appears that younger adolescents show a stronger preference for emoticons than adolescents nearing adulthood. In addition, the younger group in our corpus uses significantly more renditions of laughter – although the effect size is small – ($p < .0001$, chisq. = 81.30, odds ratio = 1.08) as well as more non-standard abbreviations ($p < .0001$, chisq. = 338.55, odds ratio = 1.26). The survey participants also interpreted the occurrence of spelling deviations (both genuine errors and deliberate manipulations) as

typical of younger chatters, and standard writing as typical of older ones. In related research, it is widely accepted that non-standard language use culminates around the age of 15-16 and then decreases as teenagers age – i.e. the 'adolescent peak' (De Decker & Vandekerckhove 2017, 277; Holmes 1992, 184). In our corpus the ratio of words that are spelled conform standard Dutch spelling is higher in older adolescents' CMC than in that of younger teenagers ($p < .0001$, chisq. = 2199.90, odds ratio = 1.15). However, the survey participants' intuitions are not always accurate. For instance, in our data, more English words are produced by younger adolescents and not, as the participants thought, by older ones.

Strikingly, the participants' replies for this task contained much more negative evaluative language compared to their replies for gender detection. The students appeared to have strong judgmental attitudes towards younger teenagers' online writing practices, calling their deviant spelling forms "ugly" and "childish", often assuming that spelling errors were made on purpose. Some participants explicitly noted that younger teenagers do not care about correct writing, that they are lazy, and that they are exclusively focused on being "cool". These questions on linguistic attitudes thus also reveal attitudes on the *people* (in this case young teenagers) associated with certain language varieties or phenomena (cf. Lybaert 2014, 24, and references therein).

4.1.3. Education profiling

The participants did not seem to be aware of or even believe in linguistic differences in the online writing practice of teenagers with different educational backgrounds: 52% explicitly denied the potential existence of such patterns, 33% were neutral and only 15% agreed. This general disbelief was also reflected in the performance in the detection task: only 25% of the answers was correct, versus 35% incorrect and 39% 'don't know'-replies. Once again, the participants' social profile did not significantly influence their performance or their overall awareness of educational differences. However, a difference could be found when splitting up the three awareness questions (i.e. per pair of educational tracks). For two out of these three subquestions, gender was a significant predictor, with girls believing (even) less in the linguistic educational difference than boys. This is a striking result, as *more* educational linguistic variation has actually been found in the online writing of teenage girls compared to boys (Hilte et al., forthcoming).

Table 5 summarizes the cues used by the participants in the detection task. The relevance of the features rendered in bold and italics was corroborated by the reference corpus, i.e. these features were used significantly more or less frequently by students in particular educational tracks. For the features that are struck through however, we found no support in the data.

GENERAL	TECHNICAL	VOCATIONAL
Chatspeak:	**Chatspeak:** ~~- emoticons~~	**Chatspeak**: ***- many emoticons*** ***- repetition of punctuation marks*** ~~- allcaps~~
Correctness: ***- correct standard language / spelling*** - Formal writing	**Correctness**: ***- dialect*** >< ~~- correct standard spelling~~	**Correctness:** *- ("obvious")* ***spelling mistakes*** ***- abbreviated*** ***- no standard Dutch*** - incorrect/"weird" syntactic constructions
Punctuation: - correct (formal) use of punctuation marks	**Punctuation:** /	**Punctuation:** ***- either no punctuation marks at all, or repetition (see chatspeak features)***
Capital letters: - correct capitalization	**Capital letters:** /	**Capital letters:** ~~- either no capital letters or allcaps~~
Character/nature students apparent in text: - inquisitive (school context)	**Character/nature students apparent in text:** - very social	**Character/nature students apparent in text:** - social - do not care about school
Content: - more planning - taking notes in class	**Content:** - cooking courses - asking for notes, checking timetable classes	**Content:** - cooking courses, skills, practice- rather than theory-oriented studying

Table 5: Survey result: Adolescents' intuitions on linguistic educational differences in informal CMC

Once again, the cues are both content- and style-related. With regards to content, the participants linked the topic of the messages to (their idea of) the courses and mindset in the different educational tracks. Students in General Education were thought to plan more and take more notes in class. Students in the less theory-oriented Technical Education were assumed to take notes and ask for notes too, as well as to check timetables for classes with their interlocutors. In addition, the more practice-oriented aspect of their education emerged as well: they were linked to chat messages about cooking, as cooking courses might be a part of their specific educational track. Other participants, however, linked the topic of cooking courses to the practice-oriented Vocational Education, along with chat messages about skills or practice-related issues rather than theory-oriented studying. Furthermore, inquisitiveness was linked to General students, whereas Vocational students tended to be associated with indifference with respect to school. Finally, students in the more practice-oriented tracks (Technical and Vocational) were attributed a greater social involvement.

As for the stylistic features, some typical chatspeak markers were mentioned. The participants considered the use of emoticons and the repetition of punctuation marks to be typical of Vocational students. This tendency is supported by our CMC-corpus data since these features occur more often in conversations by Vocational students (versus those produced by all other students) ($p < .0001$, chisq. = 28119.82, odds ratio = 2.46 for emoji; $p < .0001$, chisq. = 170.37, odds ratio = 1.29 for punctuation repetition). The rendering of entire words or

phrases in capital letters ('allcaps') was also perceived as typical of Vocational students and, once again, this preference pattern is statistically significant in the reference corpus, although the effect size is very small ($p = .0021$, chisq. = 9.45, odds ratio = 1.06). Furthermore, the respondents correctly assumed that General Education students had a greater preference for standard writing ($p < .0001$, chisq. = 7386.24, odds ratio = 1.33 for the use of standard Dutch lexemes), whereas a higher ratio of non-standard lexemes (e.g. dialect words, or words containing spelling mistakes or other non-standard features) and non-standard abbreviated forms were correctly linked to Vocational students ($p < .0001$, chisq. = 351.85, odds ratio = 1.07 for non-standard Dutch lexemes; $p < .0001$ chisq. = 357.23, odds ratio = 1.28 for abbreviations). Concerning Technical students' language use, there was no consensus among the survey participants: while some thought that these students' messages contained more dialect words, others thought they were closer to the linguistic standard. Only the former assumption was supported by the reference corpus ($p < .0001$, chisq. = 6460.98, odds ratio = 1.31). Another assumption that was not supported by the corpus concerned a supposedly higher preference for emoticons amongst Technical students. The less accurate assessments of the characteristics of Technical students' CMC seem to reflect the actual practice of this group: time and again we found that the writing practices of this group, which is in the middle of the educational spectrum, are more varied and unpredictable than those of the other groups (see Hilte et al. 2018a, 2018b).

While some negative evaluative comments could be found among the participants' responses, especially about Vocational students' writing (who were attributed "obvious" or avoidable linguistic errors and an indifferent attitude), other participants explicitly expressed their reluctance with respect to the assumed existence of educational linguistic patterns. Whereas linguistic age and gender patterns are generally accepted, educational differences are not. The subject even appears to be a somewhat sensitive topic. We have to emphasize, however, that this reveals a discrepancy between the perception on online writing practices and the actual production, as our corpus does reveal statistically significant and very consistent linguistic differences between distinct education groups (see also Hilte et al. 2018a, 2018b, forthcoming).

4.2. Block 5: Correction or conversion task

In the next part of the survey, the participants were instructed to detect, 'correct' (i.e. convert into the formal standard equivalent) and give their opinion on different types of deviations from formal standard writing in chat messages written by their peers. We note that we only report deviations as 'detected' when they were also corrected adequately, since in some cases the respondents actually adapted words that in no respect deviated from formal standard Dutch and left the item in question unchanged. Similarly, we only report intolerance scores for participants who succeeded in detecting and correcting the actual deviation. For all participants combined, 62% of the deviations were both detected and corrected adequately. A low intolerance score (i.e. number of 'would bother me on social media'-

responses) of 11% could be noted. The other 89% of these responses were very heterogeneous, containing replies by students who noticed the mistake but felt neutral about it or were not bothered by it, but also replies by students who did not notice the actual deviation.

However, these average scores obfuscate highly diverging results for the distinct types of 'non-standard' features: while most prototypical CMC-deviations from standard Dutch are detected well, classical (not CMC-related) spelling errors are not. The most striking example relates to a highly stigmatized morphological spelling error in Dutch verb conjugation (see e.g. Sandra et al. 2004):

(1) original: *Ja maar de klank verander**d** ook precies*
correction: *Ja maar de klank verander**t** ook precies*
translation: 'Yes, but the sound changes too, it seems'

Strikingly, only 34% of the participants were convinced that the sentence contained a non-standard item and only 10% of all participants saw the actual mistake and adapted it adequately. Consequently, the other 24% of the students who claimed to have spotted the mistake actually hadn't, but instead focused on (and 'corrected') another part of the utterance[3] (which was not incorrect and thus irrelevant in the context of this question).

As opposed to the classical spelling errors, typical chatspeak features (e.g. non-standard abbreviations) were detected and adapted very well: for these deviations, scores of 89% or higher were obtained. These results suggest the existence of register sensitivity among the participants, as the adolescents appear to be very well aware of the non-standard nature of typical CMC-characteristics or at least know that these features can be no part of formal writing (see also Vandekerckhove & Sandra 2016).

Finally, for the attitudinal dimension, the participants were asked to give their opinion on the different deviations from standard Dutch by indicating their (dis)agreement with the following statement: 'This "mistake" would bother me in a chat message on Facebook or WhatsApp'. A predominantly tolerant tendency could be noted: most participants claimed not to be bothered at all by most of the features in a CMC-context. The only clear exception was the deviant spelling of *schattig* ('cute') as *sgattig* – a typical form of Dutch chatspeak spelling where the consonant cluster 'sch' (/sX/) is replaced by the phonologically equivalent (but non-standard Dutch) spelling 'sg':

(2) original: *Jij bent **sgattig***
correction: *Jij bent **schattig***
translation: 'You are cute'

Surprisingly, 49% of the participants claimed this mistake would bother them in a social media context. This is a strikingly high percentage, as none of the other deviations bothered more than 11% of the participants. This specific spelling deviation appears to be typical of young

[3] Some participants, for instance, replaced the word *klank* ('sound') with a synonym, such as *geluid*.

teenagers' chatspeak: while occurrences of *schattig* (i.e. correct formal spelling) in the reference corpus are quite evenly distributed among the age groups (54% of all 529 occurrences are produced by younger and 46% by older adolescents), *sgattig* is used much more frequently by the younger group (89% of all 47 occurrences by younger and 11% by older teenagers). The participants' negative attitude towards example (2) may thus be linked to their negative evaluation of younger teenagers' CMC (see Section 4.1.2).

Finally, additional GLMM-analyses revealed a significant influence of the participants' age and educational background on their performance in this correction task: higher probabilities for correct answers were associated with older teenagers and teenagers in General Education. These findings thus indicate a stronger proficiency in formal standard writing for students in the most theory-oriented educational track compared to students in more practice-oriented tracks, which might reflect the extent to which formal writing is focused on in different educational systems. In addition, regardless of educational background, all students' proficiency in standard Dutch seems to increase as they age. We can compare these results to the findings by Verheijen and Spooren, who provided Dutch youths with a similar correction task: their participants were instructed to detect and correct linguistic 'errors', which could either be CMC-related deviations or more classical spelling errors in Dutch (2017, 7). No information was provided in the paper on the types of errors that were harder to identify or correct. The youths' performance, however, was positively predicted by their educational level, and surprisingly, also by their gender: both higher educated and female participants obtained higher scores in the task. Unlike in the present survey however, Verheijen and Spooren (2017, 9) found no significant age differences.

4.3. Block 6: The relevance of standard Dutch and self-reported proficiency

In view of the concerns with respect to the impact of CMC on the formal literacy of youths, we included some questions that relate to the perception of standard language proficiency and the reflection of standard language ideologies. The answers show a broad consensus with highly similar attitudes amongst the different teenage groups.

Almost all participants (92%) subscribe to the importance of standard Dutch in written school assignments. With regards to electronic communication, the students showed proof of register sensitivity: 95% indicated to use another register when writing an email to a teacher than when doing this to a friend. Concerning their teachers, 79% of the participants expected the *Dutch* teacher to speak in a standard register, whereas only 58% did so for teachers of other courses. The responses for this last question, however, were significantly influenced by the participants' age: older adolescents attached more importance to the use of standard Dutch by teachers regardless of their subject. Furthermore, almost all participants (92%) believed a good proficiency in standard Dutch would increase their chances of finding a job. However, less than two third of all students (62%) claimed to actually *be* proficient in the standard register. The potential use of standard Dutch in social media was generally met with

indifference: while only a small minority of the participants (9%) explicitly appreciated the use of the standard register in online chat conversations, an equally small minority (9%) claimed to be bothered by it.

Finally, none of the reported tendencies – except for the question on teachers – were significantly influenced by the participants' profile. Consequently, Flemish adolescents with different backgrounds appear to have very similar attitudes on standard language use. Strikingly, the different focus on formal Dutch proficiency in the distinct school systems does not seem to influence the students' opinions on the importance of the register in particular contexts.

4.4. Block 7: The social indexicality of (CMC-)features

The seventh question block in the survey concerned potential negative or positive connotations of certain linguistic features in chat messages. For different chat utterances, the participants had to evaluate how friendly or kind they thought the author was. Several of these utterances were very similar except for one specific element.

Three groups of variations on the same sentence were presented to the participants. In the different variations, the original sentence either ended with emoticons or emoji, with a full stop, or with no emoji or punctuation whatsoever (see example (3) below). These related messages were not presented together to the participants, as the order of all utterances in this block was randomized. The following tendencies were observed: when the sentence ended with no punctuation marks or emoji, as in (3a), most participants had a neutral opinion on the author's friendliness. When the message ended with a full stop, as in (3b), most participants considered the author to be unfriendly, whereas when it ended with emoji (either hearts or smiley faces), as in (3c), most of them considered the author to be friendly. These findings support the idea that full stops (and to a lesser extent the absence of punctuation marks whatsoever) may be perceived as unfriendly, and even rude or somewhat passive aggressive. Emoji, on the other hand, appear to mitigate the message expressed in a chat utterance.

(3a) *Ook goe* '[That's] fine too'
(3b) *Ook goe.* '[That's] fine too.'
(3c) *Ook goe* 💕 '[That's] fine too 💕'

Next, we examined the connotation of the thumb-emoji used as a reply to another chatter's message. Again, the same tendencies could be observed for the two examples included in the survey, with 'thumb-replying' authors being perceived as unfriendly by most participants. In example (4), especially, author B appeared to come across as highly unfriendly, with 78% 'unfriendly' votes. This very outspoken non-appreciation could be linked to the fact that the thumb-emoji is used as a response to a fairly personal message, which may be a context in which such a short non-verbal reply is considered 'not done'.

(4) Author A: *Sorry*

Author B: 

Interestingly, the participants' profile interfered with their responses. Additional GLMM-analyses indicated that girls and students in more theory-oriented educational tracks are significantly more sensitive to the indexicality of particular non-verbal features in social media utterances. For all teenagers, however, this sensitivity appears to increase as they grow older, which suggests that teenagers gradually acquire CMC-norms.

4.5. Block 8: Ranking chat messages

In the final block of the survey, the participants had to rank 13 authentic chat messages (written by their peers) by preference or appeal. Since their own practices were the reference point, the results potentially point to the role of particular features in their personal online identity construction: which features carry enough positive connotations for them to be included in their own self-reported writing practices and which do not? Below, we focus on the extremes of the scales: i.e. the features that got an overall high or low ranking.

Most messages that were clearly popular among many participants contained English words or abbreviations:

(5) *Hellooooo xx* ('Hello xx')
(6) *Thanks* ('Thanks')
(7) *Wtf haha* ('What the fuck haha')

Consequently, the incorporation of English in Dutch chat conversations seems to hold a strong contemporary prestige ('coolness'/'dynamism') in the eyes of many Flemish adolescents, regardless of their socio-demographic profile, and has much potential for identity construction.

Utterances containing an abundance of either new (example (8)) or old vernacular features (example (9)) were evaluated negatively by most participants. While e.g. the use of expressive markers such as emoji certainly tends to be appreciated by the adolescents, they generally dissociate themselves from the excessive use of them. In other words, proportions matter.

(8) 😂😂❤❤💖💖😊 *hahaha als ge rustig fietst komt alles in orde* ❤😂🙈🙈😝
('hahaha if you bike slowly, everything will be fine')
(9) *Vorwa da na wer??!!* ('Now what is that for??!!')

However, there is a clear gender and education divide concerning utterances that are very typographically expressive, such as (8). These messages were evaluated negatively by most boys, whereas the girls' reactions were more varied. In addition, while such highly expressive messages were evaluated negatively by almost all General Education students, responses were more varied among Vocational and Technical students. We recall that a quantitative

difference in emoticon use could be attested in the reference corpus and for gender in related research too. Girls use significantly more emoticons than boys ($p < .0001$, chisq. = 7101.96, odds ratio = 1.71) (see also e.g. Baron 2004, 415; Herring & Martinson 2004, 436; Kucukyilmaz et al. 2006, 282; Parkins 2012, 52). Students in General Education use significantly fewer emoticons than their peers in other tracks, although the odds ratio is very small ($p < .0001$, chisq. = 28119.82, odds ratio = 2.46 for emoji; $p < .0001$, chisq. = 127.50, odds ratio = 1.07).

These findings on adolescents' attitudes with respect to online writing can complement previous results on youths' production of CMC, as they show that teenagers with distinct socio-demographic profiles do not only use certain chatspeak features to different extents, but that they appear to do so out of a difference in appreciation of these linguistic markers.

5. Conclusion

This attitudinal study analyzed adolescents' perception of their peers' writing practices on social media, reporting on a survey conducted among 168 Flemish high school students. The questions and tasks were designed to examine the participants' linguistic attitudes, their awareness of sociolinguistic patterns in online language use, and to a minor extent their (formal) writing skills.

With respect to the awareness of social patterns in CMC, very different results emerged for the estimated effect of education compared to age and gender: While the participants performed fairly well for age and gender detection, they hardly believed in educational differences in online writing and performed much worse in the education detection tasks. In addition, the linguistic cues they used in their decision-making were less accurate for this specific social variable. These results are quite striking, since clearly distinct writing patterns for teenagers with different educational backgrounds can be attested in our social media corpus, which consists of online conversations produced by their peers. However, this discrepancy between teenagers' perception and production of CMC in terms of educational patterns might – at least partially – be related to the more sensitive nature of this topic: respondents appeared to be quite reluctant when confronted with questions on the impact of education.

The tasks that focused on the detection of deviations from formal standard writing, both in the form of chatspeak markers and common spelling errors, displayed a striking combination of a fairly high register sensitivity with poor spelling skills: typical chatspeak markers were detected with high accuracy, whereas performance for classical spelling errors was much worse. This suggests that adolescents use typical chatspeak features intentionally, and that they are aware of the genre-specific (in)appropriateness of these linguistic markers. These results can therefore contribute to the debate on the potential negative effects of CMC on literacy, offering a more positive perspective by showing how teenagers are mostly unaware of classical language errors, whereas they do show awareness and register sensitivity when it

comes to CMC-specific deviations from formal writing. Furthermore, the results clearly showed that most of the deviations from standard Dutch did not bother the participants when used in social media contexts. Additional questions on the importance of standard Dutch indeed revealed that the participants only considered this register to be vital in formal (e.g. school-related) contexts. With the latter attitude, they pay lip service to classical standard language ideologies. Yet, we cannot but conclude that the adolescents display some degree of indifference with respect to the standard language as well. Moreover, less than two third of all participants considered themselves to actually be proficient in standard Dutch, which might also point to a certain dissociation from the standard register.

Finally, although many linguistic variants and varieties – ranging from a very non-standard to a very standard register – seem to be 'accepted' on social media, they are not all appreciated to the same extent. The use of certain non-verbal elements in chat messages appeared to evoke negative connotations: for instance, authors who ended their chat messages with a full stop were often perceived as unfriendly. This points to the existence of alternative norms for online writing: it may be wise to avoid using full stops at the end of an utterance if you want to create goodwill. A moderate use of emoji for closing messages in many cases seems a preferable strategy, since both the responses to the survey questions and the analyses of online writing practices in our reference corpus reveal that adolescents appreciate the use of typical CMC expressive markers (e.g. emoji). However, it is crucial to use them in the right doses. Moreover, the right dosage tends to be different for different social groups: the tolerance level for e.g. the use of emoji is much lower amongst high educated adolescents and boys than amongst girls and lower educated adolescents. Interestingly, these differences in appreciation perfectly correspond to actual frequency patterns in adolescents' CMC as attested in our corpus. Finally, with respect to the appreciation of particular features, another strategy for increasing media coolness appears to be the integration of English slang (i.e. English that is no part of Standard Dutch).

Strikingly, unlike the results for the appreciation of particular chatspeak features discussed in the previous paragraph, most of the survey responses were not significantly influenced by the participants' socio-demographic profile, which shows that when it comes to linguistic attitudes, most Flemish teenagers share a common ground, regardless of their specific age, gender or educational track. However, some subtle but interesting differences could be noted. For instance, girls showed a significantly weaker awareness of or 'belief' in educational linguistic differences – although, as has been shown in previous research, more pronounced educational differences can be found in girls' social media writing than in boys' (Hilte et al. forthcoming). Consequently, the discrepancy between CMC production and perception in terms of educational patterns seems to be larger for teenage girls than boys. With regards to linguistic skills, we found that while all teenagers performed rather well in the correction task, older teenagers and teenagers in the theory-oriented General Education were more likely to detect and adequately correct the linguistic deviations from formal writing. These findings suggest that although highly educated teenagers are more proficient in the standard register

– which is likely to be an effect of their more theory-oriented education – *all* adolescents, no matter their educational track, become more proficient as they grow older. A final attitudinal difference consisted in girls and higher educated teenagers showing a higher sensitivity to particular negative connotations evoked by certain non-verbal features in chat messages. This sensitivity also appeared to increase as teenagers grow older.

We can conclude that this attitudinal study on teenagers' CMC can contribute to the debate on the effects of online writing on youths' literacy, and can be combined with variationist sociolinguistic studies to provide more insight in adolescents' production of CMC, answering not only the question of *how* teenagers write on social media, but also *why*. Whereas Flemish adolescents clearly appear to share a common ground concerning their attitudes on online language use, the subtle differences and nuances that emerged from the analyses show that, just like for adolescents' *production* of computer-mediated communication, their *perception* of CMC is more complex and fascinating than one might initially think. Finally, with respect to the question formulated in the subtitle of this paper, it seems there is no straightforward answer: Has non-standard become the new standard? We might say to some extent it has, at least in social media contexts, since there appears to be quite a lot of indifference with respect to the use of standard language in online writing. However, following traditional standard language ideologies, the importance of standard language in formal contexts is clearly acknowledged. Moreover, both the appreciation and disapproval of particular CMC features and the way they are used or the proportions in which they are used, points to the existence of alternative norms for informal writing. In other words, though they cannot always clearly be delineated, there are standards for online writing too and though there seems to be a broad consensus with respect to the appreciation of particular features, these standards are not completely identical for different social groups. This brings us back to the starting point of this paper: informal writing has indeed contributed to a pluralization of (written) language norms (see Androutsopoulos 2011).

Acknowledgments

We are grateful towards Brahim Zarouali for his advice in setting up the survey. We also wish to thank all teachers and students for their participation.

References

Androutsopoulos, Jannis. (2011). Language change and digital media: A review of conceptions and evidence. In Tore Kristiansen, & Nikolas Coupland (Eds), *Standard languages and language standards in a changing Europe* (pp. 145-161)*,* Oslo: Novus.

Argamon, Shlomo, Moshe Koppel, James W. Pennebaker, & Jonathan Schler. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM. Inspiring Women in Computing* 52(2), 119-123.

Auer, Peter. (2011). Dialect vs. standard: A typology of scenarios in Europe. In Bernd Kortmann, & Johan van der

Auwera (Eds), *The languages and linguistics of Europe: A comprehensive guide (The world of linguistics 1)* (pp. 489-504), Berlin / Boston: De Gruyter Mouton.

Baron, Naomi S. (2004). See you online: Gender issues in college student use of instant messaging. *Journal of Language and Social Psychology* 23, 397-423.

Baron, Naomi S. (2008). *Always on: Language in an online and mobile world.* Oxford: Oxford University Press.

De Decker, Benny, & Reinhild Vandekerckhove. (2017). Global features of online communication in local Flemish: Social and medium-related determinants. *Folia Linguistica* 51(1), 253-281.

[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* (pp. 8-18)*,* Brussels: Department of Education and Training*.*

Grondelaers, Stefan, Paul van Gent, & Roeland van Hout. (2016). Destandardization is not destandardization. Revising standardness criteria in order to revisit standard language typologies in the Low Countries. *Taal en Tongval* 68(2), 119-149.

Herring, Susan C., & Anna Martinson. (2004). Assessing gender authenticity in computer-mediated language use: Evidence from an identity game. *Journal of Language and Social Psychology* 23, 424-446.

Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans. (2018a). Adolescents' social background and non-standard writing in online communication. *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. *International Journal of Society, Culture & Language* 6(2), 73-89*.*

Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans. (2018c). Expressive markers in online teenage talk: A correlational analysis. *Nederlandse Taalkunde* 23(3), 293-323*.*

Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans. (forthcoming). Modeling adolescents' online writing practices: The sociolectometry of non-standard writing on social media. *Zeitschrift für Dialektologie und Linguistik*.

Holmes, Janet. (1992)*. An introduction to sociolinguistics*. London / New York: Longman.

Kristiansen, Tore, Peter Garrett, & Nikolas Coupland. (2005). Introducing subjectivities in language variation and change. *Acta Linguistica Hafniensia* 37, 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* (pp. 274-283), Berlin: Springer.

Labov, William. (1972). *Sociolinguistic patterns.* Philadelphia: University of Pennsylvania Press.

Labov, William. (2001). *Principles of linguistic change. Volume 2: Social factors.* Malden: Blackwell.

Lybaert, Chloé. (2014). *Het gesproken Nederlands in Vlaanderen. Percepties en attitudes van een spraakmakende generatie.* Ghent: Ghent University (doctoral thesis).

Mehl, Matthias R., & James W. Pennebaker. (2003). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality & Social Psychology* 84, 857-870.

Newman, Matthew L., Carla J. Groom, Lori D. Handelman, & James W. Pennebaker. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes* 45(3), 211–236.

Parkins, Róisín. (2012). Gender and emotional expressiveness: An analysis of prosodic features in emotional expression. *Griffith Working Papers in Pragmatics and Intercultural Communication* 5, 46-54.

Sandra, Dominiek, Steven Frisson, & Frans Daems. (2004). Still errors after all those years...: Limited attentional resources and homophone frequency account for spelling errors on silent verb suffixes in Dutch. *Written Language & Literacy* 7(1), 61-77.

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. *PLoS ONE* 8(9).

Stenström, Anna-Brita, Gisle Andersen, & Ingrid Kristine Hasund. (2002). Non-standard grammar and the trendy use of intensifiers. In Anna-Brita Stenström, Gisle Andersen, & Ingrid Kristine Hasund (Eds), *Trends in teenage*

talk: Corpus compilation, analysis and findings (pp. 131-163), Amsterdam: John Benjamins.

Vandekerckhove, Reinhild, & Dominiek Sandra. (2016). De potentiële impact van informele online communicatie op de spellingpraktijk van Vlaamse tieners in schoolcontext. *Tijdschrift voor Taalbeheersing* 38(3), 201-234.

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. *Reading and Writing* 23, 719-733.

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

Verheijen, Lieke. (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).

Verheijen, Lieke. (2016). De macht van nieuwe media: Hoe Nederlandse jongeren communiceren in sms'jes, chats en tweets. In Dorien Van De Mieroop, Lieven Buysse, Roel Coesemans, & Paul Gillaerts (Eds), *De macht van de taal: Taalbeheersingsonderzoek in Nederland en Vlaanderen* (pp. 275-293), Leuven / The Hague: Acco.

Verheijen, Lieke, & Wilbert Spooren. (2017). The impact of WhatsApp on Dutch youths' school writing. In Egon W. Stemle, & Ciara R. Wigham (Eds), *Proceedings of the 5th conference on CMC and social media corpora for the humanities (cmccorpora17), 3-4 October 2017, Eurac Research, Italy* (pp. 6-10).

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

CHAPTER 9

Conclusion

This dissertation aimed to lay bare correlations between teenagers' socio-demographic profile and their informal online writing practices through a diversified operationalization of the variables and their interactions. We collected a large and representative dataset for this purpose, containing over 400,000 instant messages produced on Facebook Messenger and WhatsApp by more than 1000 Flemish high school students with varied socio-demographic profiles. Through the multidimensional conceptualization of both the linguistic and social variables and through the inclusion of interactions between the latter, we aimed to obtain a more nuanced insight in the impact of social factors on youths' online writing practices, and reveal more subtle and complex patterns of sociolinguistic variation than have previously been attested. Furthermore, we included several social variables in the research design that have hardly been examined before with respect to (youths') online writing practices, i.e. social class indicators such as the adolescents' educational track and the profession of their parents.

In Section 1, the main findings of the dissertation are summarized, and in Section 2, their importance and relevance is discussed. Section 3, finally, contains suggestions for further research.

1. Main outcomes of the dissertation

Below, we discuss the main findings of the seven research papers presented in Chapters 2 to 8. The results are summarized per (cluster of) research question(s) that the dissertation aimed to answer.

RESEARCH QUESTION 1: Which patterns of sociolinguistic variation can be attested in adolescents' informal online writing with respect to age, gender and social class indicators such as educational track and parental profession? Can significant interactions between these socio-demographic variables be observed?

Throughout the dissertation, both divergent tendencies and striking similarities were attested between different groups of Flemish adolescents with respect to their social media writing. As for the *content* of their instant messages, a strong common ground could be observed: regardless of their specific age, gender or educational background, all teenagers appear to share largely similar interests and preoccupations and discuss these in online conversations (e.g. friends, family, school and social media). With respect to the *stylistic* properties of online writing, however, clear patterns of sociolinguistic variation could be attested. While certain linguistic features can be considered 'prototypical' markers of the genre of informal online

communication as they are used by almost all youths, the extent to which they are favored appears to be socially determined. In addition, differences were observed among distinct social groups of teenagers with respect to more general linguistic properties of their texts, related to 'traditional literacy' rather than to digital communication specifically (e.g. average sentence length). Below, we summarize the main results per social variable. In addition to these main effects, significant interactions have been attested between the social variables, which indicates that their impact on youths' online writing practices is not always independent and that they should thus not be studied in isolation. These interactions are discussed in the final paragraph of this section.

EDUCATIONAL TRACK

Educational track has largely remained out of scope in sociolinguistic studies focusing on adolescents' online writing practices, even though it is a major factor of their social profile. We distinguished between students in the three main types of Belgian secondary education, i.e. General, Technical and Vocational Secondary Education (ranging from a very strong theoretical to a very strong practical orientation). Our findings indicate that educational track is a strong determiner of teenagers' online writing practices. A clear distinction was observed between teenagers on the two ends of the educational continuum from theory to practice: vocational students systematically insert more non-standard markers in their online writing than their peers in General Education. Furthermore, they appear to favor features of both new and old vernacular (i.e. 'digital' and 'traditional' non-standardness) to a greater extent. Strikingly, while technical students occupy an intermediate position on the educational scale, they do not appear to occupy an intermediate *linguistic* position. Rather, their online writing style has more variable and unpredictable linguistic properties. A possible explanation is that technical students are truly a diverse social group due to the hybrid orientation of their specific educational tracks, which might cause a greater interpersonal linguistic variation among these students (with e.g. more varied and unpredictable frequencies for various linguistic features).

With respect to more general linguistic properties of the teenagers' instant messages, educational differences (with respect to the focus on formal language skills) seem to be reflected in social media writing to some extent, as a stronger verbal orientation could generally be observed for more theory-oriented students. Especially when it comes to conveying sentiment or emotion in informal online texts, youths with distinct educational backgrounds seem to prefer different repertoires: while more theory-oriented students show proof of a stronger verbal expression, their vocational peers are more inclined to opt for typographic or pictorial means in this respect.

In conclusion, while they may generally be less verbally- or standard-oriented, vocational students do not score worse for *digital literacy*, as they truly exploit the communicative possibilities and tools typical of digital media.

SOCIAL CLASS

The adolescents' social class was the most challenging variable to operationalize, and its implementation was adapted and improved in different stages of the research project. The final implementation combines educational track and parental profession, and leads to a distinction between prototypical upper class, middle class and working class youths. A clear linguistic distinction could be observed between the teenagers on the extreme ends of the social continuum, with working class youths systematically inserting more non-standard markers into their instant messages than their upper class peers. While middle class youths occupy an intermediate linguistic position when all non-standard markers are clustered, a more varied pattern emerges for the individual markers, which suggests that these youths' online writing practices have their own distinct properties. Another main finding is that working class teenagers are not only attracted to *old* vernacular (which is in line with older sociolinguistic findings, see e.g. Labov 1972) but to *new* vernacular as well, which indicates that working class youths strongly connect to the digital writing culture too.

Finally, we examined teenagers with 'hybrid' social profiles. For youths with a major discrepancy between their educational track and the profession of their parents, deviant linguistic practices were observed that may be indicative of sociolinguistic hypercorrection. These findings suggest that social mobility might make people's language use more dynamic and that social aspiration might favor hypercorrective linguistic behavior (see also Aitchison 2013; Labov 1966; Labov 2006).

AGE

As for age, we make a distinction between young teenagers (13 to 16 years old) and older teenagers or young adults (17 to 20 years old). Concerning youths' deviations from standard writing norms in instant messages, our results support the idea of an *adolescent peak* (see e.g. Holmes 1992, 184; Coates 1993, 94), as significantly fewer non-standard markers occur in texts produced by teenagers over the age of sixteen. However, while this age effect could be observed for distinct types of non-standard markers (e.g. expressive markers, oral features, non-standard abbreviations), it is not always as outspoken for boys as it is for girls (see below). Furthermore, instant messages produced by older teenagers do not only contain fewer non-standard markers, but they also show proof of a stronger 'traditional literacy'.

Finally, teenagers seem to prefer distinct repertoires when it comes to the online expression of sentiment or emotion, depending on their age: older adolescents favor verbal expression, whereas younger teenagers seem to prefer typographic or pictorial means. In other words, while younger teenagers' traditional literacy and their verbal and standard language orientation may generally be less well developed than older adolescents', their digital expression appears to be stronger.

GENDER

With respect to gender, our results show that classical sociolinguistic patterns are actually reproduced in informal online writing. One of the most consistent findings in western sociolinguistics (though there are counterexamples) is that traditional substandard language (dialect, 'working class speech') more strongly appeals to men than to women. Trudgill (1983, 161) already pointed to the robustness of this pattern. In addition, studies in interactional sociolinguistics have indicated that women focus more on establishing emotional and social connections (e.g. Tannen 1990; Holmes 1995). This translates to our findings as follows: boys use more traditional non-standard features (regional language and slang) in their instant messages, whereas girls opt more for the expressive-typographic markers that are specific to the online genre (e.g. emoji). Furthermore, we found that girls do not only use this typographic repertoire to a greater extent than boys to express emotional or social involvement, but that their verbal expression of this type of involvement is stronger too. In other words, classical gender patterns are not blurred in new media, on the contrary, they might even be reinforced by the availability of certain pragmatic tools (e.g. emoji). In terms of traditional literacy, finally, no clear gender differences could be observed.

INTERACTIONS BETWEEN THE SOCIAL VARIABLES

In addition to the general effects for each of the social variables, important interactions *between* these variables have been observed, which lead to a more accurate and nuanced insight in the impact of social factors on youths' informal online communication. While in the first chapters of the dissertation potential interactions between the social factors are not yet systematically operationalized, distinct age and gender dynamics with respect to the use of non-standard markers are observed depending on the teenagers' social class. In later chapters, interactions between the social predictors are systematically included, and it is demonstrated that the linguistic impact of teenagers' age, gender and educational track is not always independent. For instance, important interactions could be observed between age and gender. While all teenagers tend to use fewer expressive and oral (non-standard) chatspeak markers as they grow older, this decrease is much stronger – and only significant – for girls than for boys. Consequently, our findings do not only reveal strong general gender effects (i.e. girls using more expressive markers than boys at whichever age, and boys using more oral markers than girls at every age, see above), but they also suggest that the prestige that girls and boys derive from both old and new vernacular seems to evolve differently as they reach adulthood. While the appeal of (especially 'old') vernacular stays more or less the same for boys regardless of their age, a clear drop in popularity can be attested for features of both old and new vernacular among girls. These results suggest an attitudinal gender difference with regards to the acceptance of external social norms (see also Eisikovits 2006), with girls appearing to aim more for a standard, adult linguistic 'appearance' on social media as they grow older, and boys barely seeming to adapt their online language practices, as far as the use of non-standard markers is concerned.

Additional interactions between age and gender were attested with respect to certain aspects of traditional literacy: e.g. a significant increase in verbal orientation by age could be attested for girls only. Furthermore, the adolescents' gender and educational track appeared to interact too. Between girls in different educational tracks, larger differences could be observed with respect to how frequently they use oral non-standard markers in instant messages compared to boys (i.e. between boys in distinct educational tracks, much smaller differences could be attested in this respect). This greater sociolinguistic variation in girls' texts may be related to a stronger female awareness (and signaling) of social status differences (see below). A final interaction between gender and education concerns the teenagers' use of chatspeak abbreviations: significant gender differences in this respect could only be observed among students in General Education.

In conclusion, these findings reveal distinct linguistic age and gender dynamics in different educational tracks and different social classes, and show a greater linguistic impact of (and a greater linguistic variation related to) both age and educational track for girls than for boys.

RESEARCH QUESTION 2: Are sociolinguistic variation patterns in youths' informal online writing sufficiently robust to be used in quantitative (descriptive and predictive) modeling?

The dissertation contains studies on two types of quantitative models that aim to answer opposite yet complementary research questions: we wanted to model adolescents' online writing practices given relevant aspects of the authors' socio-demographic profile, and we aimed to predict teenagers' educational track based on a sample of their instant messages. Below, we summarize our findings.

DESCRIPTIVE MODELS: LINGUISTIC FEATURES AS RESPONSE VARIABLE

Using generalized linear mixed models (GLMMs), we modeled how frequently the adolescents use different (sets of) non-standard features of online writing. In addition, we used linear mixed models to model more general linguistic properties of their instant messages, related to traditional literacy rather than to the specific characteristics of online communication. Important advantages of these (G)LMM-analyses compared to other descriptive statistic methods are the simultaneous inspection of multiple social predictors, the inclusion of interactions between these social predictors, and the inclusion of a random effect for subjects (in order to take the impact of the individual participants into account). The modeling of teenagers' online writing practices based on their socio-demographic profile revealed interesting and nuanced sociolinguistic patterns (see above). Below, we compare these results to the predictive models' findings and show in which ways they complement each other.

PREDICTIVE MODELS: EDUCATIONAL TRACK AS RESPONSE VARIABLE

The opposite task concerns the prediction of teenagers' educational track based on their online writing. Again, we distinguish between three classes: students in General, Technical and Vocational Secondary Education. The results from the pilot study are promising and indicate that the task is doable. In addition, they suggest that the most informative features in this respect are specific occurrences of stylistic chatspeak phenomena.

While the distinction between general and vocational students appears to be relatively easy to make, the detection of students in the intermediate technical track is much harder. These findings are in line with the descriptive models' results, which showed that linguistic distinctions between general and vocational students' instant messages are highly consistent, while the messages produced by technical students appear to have more varied linguistic properties. The different models' complementary findings suggest that technical students are truly a distinct class, with more varied and less predictable linguistic practices. Finally, the classification task appeared to be easier for girls than for boys, which may reflect the greater linguistic variation that was observed between girls in distinct educational tracks compared to boys.

RESEARCH QUESTION 3: Do teenagers' attitudes on their peers' online writing practices reflect the attested sociolinguistic patterns? Or do discrepancies emerge between adolescents' production and perception of the linguistic genre?

Finally, we complemented our analyses on Flemish adolescents' *production* of instant messages with a study on their *perception* of the genre. For this purpose, we conducted an anonymous survey among Flemish high school students with varied socio-demographic profiles. In general, the results of this study indicate very similar perceptions and attitudes among different groups of youths. Below, we summarize our findings.

SOCIOLINGUISTIC AWARENESS

The survey shows to what extent teenagers are aware of attested patterns of sociolinguistic variation in their peers' online writing. With respect to age and gender patterns, the strong awareness reflects the clear patterns that have been found in the corpus. Regarding educational track, however, a striking discrepancy was observed between production and perception: while we attested systematic linguistic differences between instant messages produced by teenagers in distinct educational tracks, especially for girls, the survey participants are hardly aware of these. In addition, the topic even seems to be somewhat sensitive. This discrepancy appears to be especially large for teenage girls, as their belief in the existence of such educational differences is significantly *lower* than that of their male peers, whereas the female utterances in the corpus show significantly *more* variation in this respect compared to the boys' texts (see above).

The (self-reported) linguistic cues that the teenagers used in intuitive profiling tasks are highly informative too. First of all, these cues are much less accurate (i.e. with the attested differences in the corpus as a point of reference) for educational track than for the age and gender tasks, especially with respect to technical students' writing practices. This seems to reflect the actual more varied and unpredictable writing practice of this group. In addition, we recall that the computational classification of teenagers' educational track based on their instant messages was also particularly challenging for these students.

REGISTER SENSITIVITY AND STANDARD LANGUAGE IDEOLOGY

A spelling test reveals poor spelling skills with respect to classical orthographic errors but a strong register sensitivity with respect to prototypical (non-standard) markers of informal online writing: regardless of their socio-demographic profile, the teenagers appear to be strongly aware that these markers do not belong to a formal writing context. With respect to their attitudes on standard Dutch, the teenagers subscribe to classical standard language ideologies: all teenagers, irrespective of their social profile, acknowledge the importance of the standard register in formal contexts. In the context of social media, however, indifference towards classical linguistic norms predominates.

APPRECIATION OF CHATSPEAK FEATURES

Finally, the survey results offer more insight in the appreciation of or tolerance towards certain prototypical features of online writing. The results correspond to attested patterns in the corpus, and suggest that teenagers use certain linguistic markers to different extents because they value these markers differently. In addition, very similar attitudes could be attested among the teenagers with respect to the (potential) negative connotations of certain non-verbal chatspeak features. Girls and students in more theory-oriented educational tracks, however, seem slightly more sensitive to the indexicality of these features, and for all teenagers, this sensitivity appears to increase with age, which suggests that teenagers gradually acquire CMC-norms.

2. Relevance of the findings

The findings of the dissertation, which shed light on multiple underresearched aspects of sociolinguistic variation in youths' online writing, are relevant in several respects (e.g. on a methodological and sociological level). Below, we discuss their relevance.

GENERAL METHODOLOGY: COMPLEX SOCIAL VARIABLES AND INTERACTIONS

The multidimensional conceptualization of both the linguistic and social variables enables a more accurate understanding of socially determined variation patterns in youths' online communication. Furthermore, the systematic inclusion of interactions between the social variables has led to a more nuanced insight in teenagers' online writing practices, by revealing

that the included social factors' impact is often not independent in this respect. Consequently, when these interactions are taken into account, traditional sociolinguistic patterns – which are clearly present in the data – are shown to be more subtle and complex than has been assumed before. A highly relevant finding in this respect concerns the teenagers' gender: not only can classical gender patterns be observed in the corpus (which, in addition, shows that 'classical' patterns are actually reproduced in the setting of new media), but a consistently greater linguistic variation was attested in the online language use of girls, e.g. with respect to age and education. These findings suggest that adolescent girls experience a stronger impact of (multiple aspects of) their socio-demographic profile than adolescent boys, and that social distinctions may thus be more relevant among teenage girls than boys. From the early days of sociolinguistics onwards, it has been suggested that women tend to engage more strongly in signaling their social status linguistically (see e.g. Trudgill 1983, 167). Our findings suggest that while over the past decades many things have changed in terms of gender equality, these older tendencies still hold to a certain extent, even among the youngest generations in dynamic new media contexts: nowadays, girls still seem more sensitive to the social indexicality of linguistic markers.

Finally, the combination of different foci of research yields informative complementary results. For instance, through the combined study of teenagers' production and perception of informal online writing, we aimed to answer not only the research question of *how* adolescents write in their online messages but also *why* they appear to do so, and found e.g. that the different extent to which distinct groups of teenagers use and favor certain linguistic markers corresponds to attested differences in appreciation of these markers. This suggests that within social groups, some sort of consensus exists about desired or expected linguistic behavior. Furthermore, interesting patterns could be observed with respect to youths' traditional versus digital 'literacy', which suggest that non-standard writing in informal online communication does not necessarily point towards weaker traditional or formal language skills, but may be a deliberate choice of repertoire (recall e.g. the divergent preferences for verbal or typographic expressions of emotional involvement that were observed for distinct groups of teenagers).

INCLUSION OF SOCIAL CLASS INDICATORS

Another methodological contribution of the present dissertation concerns the inclusion of several social variables that had hardly been examined in previous work on online writing, i.e. social class indicators such as adolescents' educational track and the profession of their parents. While these variables remained underresearched up to now in this context, our findings show that they are strong determiners of teenagers' online writing practices, and that they interact with other major aspects of adolescents' socio-demographic profile (e.g. gender). Furthermore, in related studies, participants generally have a middle or upper class profile – consequently, the linguistic practice of working class youths in a CMC-context remained largely unexplored. Our dataset is more representative in this respect through the inclusion of a large group of working class youths. The inclusion of participants with varied

social class backgrounds revealed that some previously attested findings (e.g. with respect to age- and gender-related linguistic variation) do not hold for youths in all social classes, but that distinct linguistic patterns can be observed for adolescents with an upper, middle or working class background. The observation of divergent linguistic practices and language dynamics for youths with different social and/or educational backgrounds points to the relevance of the concept of 'social class' in today's society, even amongst the younger generations (see below).
Finally, we note that the operationalization of teenagers' social class was challenging. While the current implementation is an important step forward with respect to the systematic conceptualization and inclusion of this variable, improvements can still be made. Limitations of our approach and suggestions for further work are discussed in Section 3.

As social class factors are seldom included in computational linguistic research too, our study on education profiling offers a pioneering case study that yields promising preliminary results. We recall that this classification task is not only relevant in a Belgian context, as the three educational tracks that served as class labels correspond to secondary education programs in quite a lot of countries. On a technical level, the inclusion of stylistic chatspeak features increases the generalizability of the models, since certain chatspeak phenomena are language-independent characteristics of informal online communication. Consequently, we argue that these models – when further improved, see Section 3 – may be used for different languages and different societally relevant applications. For instance, the addition of an educational compound might increase the performance of existing profiling tools, which are important in different tasks (e.g. the detection of fake accounts on social media).

SOCIAL (IM)MOBILITY

A sociological contribution of the dissertation concerns the strong correlations that have been observed between different parameters of teenagers' social class, i.e. educational track, home language and parental profession. For instance, the teenagers' (choice of) educational track appears to be strongly determined by their social background. For upper and working class families, the relation between the parents' profession and the teenagers' educational track reveals a strong tendency towards *social stagnation*, i.e. new generations staying in the same social layer as the previous one. Strikingly, however, the impact of parental profession on educational track appears to be much less outspoken in what we have delineated as middle class families. We recall that social stagnation or 'immobility' was observed for half of the participants, and social 'mobility' for the other half (with a quarter of the teenagers moving 'up' and a quarter 'down' the social ladder compared to their parents' position). These tendencies of social (im)mobility attested among our teenage participants are of societal relevance, as they reveal large-scale intergenerational dynamics with respect to social class. In addition, the strong correlations between different aspects of people's social profile emphasize the relevance of the concept of 'social class' in today's society.

REGISTER SENSITIVITY AND TRADITIONAL VERSUS DIGITAL LITERACY

Finally, the present dissertation contributes to the ongoing debate on whether informal online writing practices negatively influence youths' formal literacy skills. While the genre has received much negative media attention in this respect in the past decade (see Vandekerckhove & Sandra 2016) and many people, especially parents, caretakers and teachers, express negative opinions on the matter and seem worried (see the overview study by Verheijen 2018, 36-44), our results show proof of register sensitivity among adolescents with respect to prototypical non-standard markers of online writing. Regardless of their specific age, gender or educational track, Flemish adolescents appear to be clearly aware that these prototypical chatspeak features belong to the informal setting of social media. The dissertation thus complements and supports previous conclusions on register sensitivity amongst the young digital natives (e.g. Vandekerckhove & Sandra 2016; Verheijen 2018).

In addition, our findings may lead to a better appreciation of vocational youths' and young teenagers' linguistic practices. While these particular groups of adolescents are often assumed to have 'bad' language practices, our results reveal that they too show strong register sensitivity with respect to prototypical chatspeak markers, but that they seem to prefer digital over classical repertoires for online self-expression. Furthermore, we showed that vocational students connect to the interactive online writing culture to (at least) the same extent as students in more theory-oriented educational tracks.

3. Suggestions for further research

In this final section, we discuss limitations of the dissertation and suggest some paths for further research.

OPERATIONALIZATION OF SOCIAL CLASS

Potential issues concerning a too strict or 'narrow' approach of the complex variable of social class were dealt with in the dissertation: based on the outcomes of a pilot study, we decided to exclude the teenagers' home language as a parameter of their social class in subsequent analyses, and to restrict the operationalization to the parameters of educational track and parental profession. While this final implementation takes multiple aspects of teenagers' social profile into account and renders very consistent patterns of sociolinguistic variation, it could be further refined.

A first possible improvement concerns the classification of the parents' professions based on the widely accepted sociological EGP-scheme (Erikson, Goldthorpe & Portocarero 1979). Some social positions that may be relevant with respect to social class (e.g. unemployed people, students, housewives/-men) fall outside the scheme's scope and could thus not be classified in the present research project. A suggestion for future research consists in

'updating' the EGP-scheme in this respect so that it includes an even more varied range of social positions.

A second possible improvement concerns participants for whom *both* parents' profession is known. In the dissertation, we only use the one that ranked highest for the variable of parental profession, since we assumed that the highest ranked profession might have a major impact on general living conditions in the families, in several respects. However, a more accurate conceptualization of teenagers' social profile could be obtained by taking both parents' profession into account, as the (social) family background of a child with two upper class parents might, in some respects, significantly differ from the family situation of a child with e.g. an upper class and a working class parent.

Furthermore, in a more detailed implementation of social class, weights could be assigned to the different social parameters, since we can assume that certain aspects of teenagers' social background might be more determining than others. We recall that we excluded home language as a parameter after the pilot study since we risked oversimplifying social reality by including it. However, when using a weighted combination of the subvariables, home language could still be included (but with a smaller weight compared to e.g. educational track or parental profession).

A final suggestion to improve the current implementation – which was already addressed in the dissertation to some extent – concerns the focus on 'prototypical' social layers (i.e. prototypical upper, middle and working class youths), while many teenagers appeared to have a more hybrid social profile. In a small-scale analysis, we already showed that the language use of 'socially hybrid' teenagers was deviant and deserved further examination. Consequently, further research on the linguistic practices of teenagers with non-prototypical social profiles may lead to a more nuanced sociolinguistic understanding of the entire social continuum rather than of three (discrete) social layers.

IMPROVEMENT OF THE EDUCATION CLASSIFIER

The pilot study on education profiling is relevant since social class factors such as educational track are seldom included in profiling studies. Although the preliminary results are promising and indicate that the task is doable, methodological improvements can still be made, and additional research on the matter is necessary. For instance, our findings indicate that the abstract representations of stylistic chatspeak phenomena may still be improved, as they are currently of lesser importance (for the classifier) compared to specific occurrences of these phenomena (e.g. specific lexemes containing a certain typographic manipulation). We note that the improvement of abstract representations of chatspeak features is highly relevant, for two reasons. First of all, these representations are likely to generalize better to unseen data or to other datasets than the specific occurrences of features, and second, several (e.g. typographic or pictorial) chatspeak phenomena are not language-specific and could thus be used in language-independent profiling tools for social media texts (e.g. while a specific Dutch lexeme written in capital letters is obviously bound to Dutch or Flemish data, the use of

'allcaps' in general is a language-independent stylistic feature). Furthermore, the classification of technical students' texts in particular could be improved. In future research, additional experiments could be conducted in which more data are used (or in which the current data are e.g. processed on a post- rather than on an author-level).

ANALYSIS OF PATTERNS OF ACCOMMODATION

A final suggestion for follow-up research concerns the inclusion of conversational factors that fell outside the scope of the present dissertation. For instance, two variables that were annotated in the dataset but were ultimately not analyzed in the research project, concern the number of interlocutors in a conversation and the interlocutors' gender: we distinguished between 'dyadic' (one-on-one) chats and group chats (with more than two interlocutors), and between same-gender conversations (i.e. girls or boys only) and mixed-gender chats (i.e. including at least one boy and one girl). In further research, it could be analyzed whether and how teenagers adapt their informal online writing practices to their conversation partner(s), and which social and contextual factors (co-)determine this process of *accommodation*.

References

Aitchison, Jean. (2013). *Language change: Progress or decay?* Cambridge: Cambridge University Press.

Coates, Jennifer. (1993). *Women, men and language: A sociolinguistic account of gender differences in language.* London / New York: Longman.

Deumert, Ana, & Kristin Vold Lexander. (2013). Texting Africa: Writing as performance. *Journal of Sociolinguistics* 17(4), 522- 546.

Eisikovits, Edina. (2006). Girl-talk/boy-talk: Sex differences in adolescent speech. In Jennifer Coates (Ed.), *Language and gender. A reader* (pp. 42-54), Oxford: Blackwell.

Erikson, Robert, John H. Goldthorpe, & Lucienne Portocarero. (1979). Intergenerational class mobility in three Western European societies: England, France and Sweden. *The British Journal of Sociology* 30(4), 415-441.

Holmes, Janet. (1992). *An introduction to sociolinguistics*. London / New York: Longman.

Holmes, Janet. (1995). *Women, men and politeness*. London: Longman.

Labov, William. (1966). Hypercorrection by the lower middle class as a factor in linguistic change. In William Bright (Ed.), *Sociolinguistics* (pp. 84-113), The Hague: Mouton.

Labov, William. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Labov, William. (2006). *The social stratification of English in New York City*. New York: Cambridge University Press.

Tannen, Deborah. (1990). *You just don't understand. Women and men in conversation.* New York: Ballantine Books.

Trudgill, Peter. (1983). Social identity and linguistic sex differentiation. Explanations and pseudo-explanations for differences between women's and men's speech. In Peter Trudgill, *On dialect. Social and geographical perspectives* (pp. 161-168), Oxford: Blackwell.

Vandekerckhove, Reinhild, & Dominiek Sandra. (2016). De potentiële impact van informele online communicatie op de spellingpraktijk van Vlaamse tieners in schoolcontext. *Tijdschrift voor Taalbeheersing* 38(3), 201-234.

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

BIBLIOGRAPHY

Bibliography

1. Publications included in the dissertation

Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans. (2018a). Adolescents' social background and non-standard writing in online communication. *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. *International Journal of Society, Culture & Language* 6(2), 73-89.

Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans. (2018c). Expressive markers in online teenage talk: A correlational analysis. *Nederlandse Taalkunde* 23(3), 293-323.

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* (pp. 328-334), Association for Computational Linguistics.

Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans. (forthcoming). Modeling adolescents' online writing practices: The sociolectometry of non-standard writing on social media. *Zeitschrift für Dialektologie und Linguistik*.
(accepted with minor revisions – the revised version is included in the dissertation)

Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans. (submitted). Adolescents' perceptions of social media writing: Has non-standard become the new standard?

Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans. (submitted). Lexical patterns in adolescents' online writing: The impact of age, gender and education.

2. Other publications

Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans. (2016). Expressiveness in Flemish online teenage talk: A corpus-based analysis of social and medium-related linguistic variation. In Darja Fišer, & Michael Beißwenger (Eds), *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities, Ljubljana, Slovenia, 27-28 September 2016* (pp. 30-33).

Hilte, Lisa, Reinhild Vandekerckhove, & Walter Daelemans. (2017). Modeling non-standard language use in adolescents' CMC: The impact and interaction of age, gender and education. In Egon W. Stemle, & Ciara R. Wigham (Eds), *Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities (cmccorpora17), 3-4 October 2017, Eurac Research, Italy* (pp. 11-15).

Tulkens, Stéphan, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, & Walter Daelemans. (2016). A dictionary-based approach to racism detection in Dutch social media. In *Proceedings of the LREC 2016 Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS)* (pp. 11-17), European Language Resources Association (ELRA).

Tulkens, Stéphan, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, & Walter Daelemans. (2016). The automated detection of racist discourse in Dutch social media. *Computational Linguistics in the Netherlands Journal* 6, 3-20.

Vandekerckhove, Reinhild, Darja Fišer, & Lisa Hilte (Eds). (2018). *Proceedings of the 6th Conference on Computer-Mediated Communication (CMC) and Social Media Corpora*. Antwerp: University of Antwerp. ISBN: 9789057285868.
Url: https://repository.uantwerpen.be/docman/irua/de0576/153416.pdf

DUTCH SUMMARY

Dutch summary / Nederlandse samenvatting

Dit proefschrift wil op genuanceerde wijze correlaties blootleggen tussen het sociodemografische profiel van tieners en hun informele online taalgebruik. Verschillende onderbelichte aspecten van dit onderwerp komen aan bod. Zo worden interacties tussen de sociale predictoren onderzocht en wordt de talige invloed bestudeerd van variabelen die zelden eerder werden geanalyseerd in deze context, i.e. sociale-klasse-indicatoren zoals de studierichting van de tieners en het beroep van hun ouders. Het tekstcorpus dat centraal staat, werd speciaal voor dit onderzoeksproject verzameld, en bevat meer dan 400 000 chatberichten die door Vlaamse scholieren zijn geproduceerd op Facebook Messenger en WhatsApp. De scholieren zijn allemaal leerlingen in een van de drie voornaamste studierichtingen in het Belgisch middelbaar onderwijs: Algemeen Secundair Onderwijs of ASO (sterk theoretisch georiënteerd), Beroepssecundair Onderwijs of BSO (sterk praktijkgericht) en Technisch Secundair Onderwijs of TSO (hybride, met zowel een theoretische als een praktische oriëntatie).

Het proefschrift bevat zeven artikelen die samen een antwoord willen bieden op drie (clusters van) onderzoeksvragen. Hieronder vatten we de belangrijkste bevindingen samen.

ONDERZOEKSVRAAG 1: Welke sociolinguïstische variatiepatronen kunnen worden waargenomen in het informele online taalgebruik van tieners met betrekking tot leeftijd, gender en sociale-klasse-indicatoren als studierichting en het beroep van de ouders? Kunnen er significante interacties tussen deze sociodemografische variabelen worden waargenomen?

Inhoudelijk vertonen de chatberichten van de verschillende sociale groepen sterke gelijkenissen, maar op het vlak van de schrijfstijl nemen we verschillende tendensen waar naargelang het profiel van de jongeren. Het gebruik en de frequentie van verschillende soorten 'afwijkingen' van de formele Nederlandse schrijftaal (bv. expressieve typografische features zoals emoji, of traditionele niet-standaardtaligheid in de vorm van regionale taalkenmerken) worden onderzocht, maar ook algemenere teksteigenschappen die meer te maken hebben met 'traditionele geletterdheid' dan met de typische kenmerken van digitale communicatie (bv. gemiddelde zinslengte) komen aan bod.

De **studierichting** van de tieners oefent een sterke invloed uit op hun online taalgebruik. Zo gebruiken praktijkgerichte BSO-leerlingen systematisch meer niet-standaardtalige kenmerken in hun chatberichten dan theoretisch georiënteerde ASO-leerlingen, en nemen TSO-leerlingen – die zich op het midden van het educatieve continuum van theorie naar praktijk bevinden – geen talige tussenpositie in. Hoewel uit de chatberichten van BSO-

leerlingen over het algemeen een minder sterke verbale oriëntatie blijkt, doen deze tieners niet onder op vlak van *digitale geletterdheid*, en maken zij volop gebruik van de nieuwe communicatieve mogelijkheden en middelen die digitale media bieden.

De **sociale klasse** van de jongeren wordt geoperationaliseerd als een combinatie van hun studierichting en het beroep van hun ouders. Die informatie hebben de participanten zelf verschaft tijdens de dataverzameling. We maken een onderscheid tussen tieners met een sociaal profiel dat aansluit bij de prototypische boven-, midden- of arbeidersklassen. Jongeren uit de arbeidersklasse gebruiken systematisch meer niet-standaardtalige kenmerken dan hun leeftijdsgenoten uit de bovenklasse, en jongeren uit de middenklasse nemen geen talige middenpositie in (i.e. hun teksten vertonen een variabel linguïstisch patroon voor de verschillende onderzochte taalkenmerken). Opmerkelijk is dat tieners uit de arbeidersklasse niet alleen sterk aangetrokken zijn tot klassieke niet-standaardtaligheid, wat eerdere sociolinguïstische bevindingen ondersteunt, maar ook tot de digitale schrijfcultuur. Tieners met een hybride sociaal profiel en meer bepaald een sterke discrepantie tussen het beroep van hun ouders en hun eigen studierichting, tot slot, vertonen afwijkend taalgedrag met een neiging tot hypercorrectie.

Wat **leeftijd** betreft, vergelijken we jongere tieners (13-16 jaar) met oudere tieners (17-20 jaar). De oudere groep gebruikt systematisch minder niet-standaardtalige kenmerken en lijkt een sterkere 'traditionele geletterdheid' te vertonen in chatberichten. Opvallend is dat beide groepen andere repertoires lijken te verkiezen om zich (emotioneel/expressief) uit te drukken op sociale media: terwijl een sterkere verbale expressie wordt waargenomen voor oudere tieners, zet de jongere groep veeleer in op een typografisch repertoire.

Met betrekking tot de variabele **gender**, tonen onze resultaten dat klassieke sociolinguïstische patronen een specifieke vertaling krijgen in online interacties. Uit vroeger kwantitatief-correlationeel onderzoek weten we dat traditionele substandaardtaal (dialect, 'working class speech') vooral een aantrekkingskracht uitoefent op mannen. De meer interactionele sociolinguïstiek heeft aangetoond dat vrouwen meer inzetten op het tot stand brengen van emotionele en sociale connecties. In onze bevindingen vertaalt dit zich als volgt: jongens gebruiken meer traditionele niet-standaardtalige kenmerken (regionaal taalgebruik/slang) terwijl meisjes meer de expressieve-typografische markers hanteren die eigen zijn aan het online genre. Wat formele geletterdheid betreft, nemen we geen eenduidige genderverschillen waar. Opvallend is wel dat meisjes in chatberichten niet alleen het typografische repertoire meer benutten dan jongens om betrokkenheid uit te drukken, maar dat in deze specifieke context ook hun verbale expressie sterker is.

Ten slotte leggen we belangrijke **interacties** bloot tussen de sociale variabelen. Deze interacties bieden een meer genuanceerd inzicht in de hierboven beschreven patronen. We observeren bijvoorbeeld andere linguïstische leeftijds- en genderdynamieken voor jongeren uit verschillende sociale klassen, en stellen vast dat de talige impact van leeftijd, gender en studierichting niet steeds onafhankelijk is. Zo schrijven alle tieners standaardtaliger op sociale media naarmate ze ouder worden, maar is deze tendens veel sterker voor meisjes dan voor

jongens. Ook interageren bijvoorbeeld gender en studierichting, en vertonen de teksten van meisjes uit verschillende studierichtingen een grotere talige variatie dan die van jongens.

ONDERZOEKSVRAAG 2: Zijn sociolinguïstische variatiepatronen in het informele online taalgebruik van jongeren voldoende robuust om te worden gebruikt in kwantitatieve (descriptieve en predictieve) modellen?

Het proefschrift bevat studies rond twee types kwantitatieve modellen, die omgekeerde probleemstellingen behandelen: enerzijds willen we het online taalgebruik van tieners modelleren op basis van hun sociodemografische profiel, en anderzijds willen we de studierichting van tieners voorspellen op basis van hun chatberichten.

De **modellering van het taalgebruik van tieners op basis van hun sociale profiel** legt interessante patronen bloot (zie hierboven). Aan de hand van 'generalized linear mixed models' (GLMMs) modelleren we hoe frequent tieners bepaalde niet-standaardtalige kenmerken van online schrijftaal gebruiken. We creëren aparte modellen voor verschillende soorten kenmerken (bv. expressieve features zoals emoji versus spreektaalkenmerken zoals markers van regionaal taalgebruik), en we modelleren ook algemenere talige eigenschappen van de chatberichten (bv. gemiddelde zinslengte). Belangrijke voordelen die (generalized) linear mixed models bieden tegenover andere analyses zijn de mogelijkheid om meerdere sociale variabelen tegelijk te onderzoeken, om interacties tussen de predictoren te analyseren, en om een 'random effect' toe te voegen voor de auteurs (en zo rekening te houden met de impact van de individuele chatters).

De omgekeerde probleemstelling betreft het **voorspellen van de studierichting van tieners op basis van hun chatberichten**. De resultaten van de pilootstudie tonen aan dat dit haalbaar is, en dat de meest informatieve taalkenmerken specifieke voorkomens zijn van stilistische chattaalfenomenen. Het onderscheid tussen ASO- en BSO-leerlingen blijkt relatief eenvoudig te maken, maar het herkennen van TSO-studenten is moeilijker. In dit opzicht versterken de resultaten van de descriptieve en predictieve modellen elkaar, en tonen ze aan dat TSO-studenten echt een hybride klasse vormen, met meer variabele taalpatronen tot gevolg. Ook blijkt de classificatietaak makkelijker voor meisjes dan voor jongens, wat strookt met de grotere talige variatie naargelang studierichting die werd waargenomen voor meisjes dan jongens.

ONDERZOEKSVRAAG 3: Reflecteren de attitudes van tieners m.b.t. het online taalgebruik van hun leeftijdsgenoten de waargenomen sociolinguïstische patronen? Of bestaat er een discrepantie tussen de productie en perceptie van het talige genre door adolescenten?

Naast hoofdstukken over de *productie* van informele online communicatie door adolescenten, bevat het proefschrift ook een studie over hun *perceptie* van het genre. Hiervoor werd een anonieme enquête afgenomen bij Vlaamse middelbare scholieren. De

verschillende onderdelen van de enquête varieerden van stellingen waarmee de deelnemers al dan niet akkoord konden gaan tot korte doe-opdrachten. De resultaten van deze studie tonen bij alle sociale groepen sterk gelijklopende percepties en attitudes.

De enquête laat zien **in welke mate tieners zich bewust zijn van vastgestelde variatiepatronen in online taalgebruik**. Voor leeftijd en gender reflecteert het sterke bewustzijn van de deelnemers de duidelijke patronen die in het corpus (en in eerdere studies) zijn waargenomen. Voor studierichting stellen we echter een opmerkelijke kloof vast tussen productie en perceptie: hoewel systematische verschillen kunnen worden waargenomen tussen de chatberichten van tieners uit verschillende studierichtingen, zijn de deelnemers van de enquête zich hier helemaal niet van bewust en blijkt het onderwerp bovendien gevoelig te liggen.

De resultaten van een spellingtaak, vervolgens, duiden op gebrekkige spelvaardigheden wat klassieke spelfouten betreft, maar op een sterke **registergevoeligheid** met betrekking tot prototypische kenmerken van online schrijftaal. Wat hun attitudes met betrekking tot Standaardnederlands betreft, echoën de jongeren de klassieke **standaardtaalideologie**: ongeacht hun sociale profiel stellen ze dat ze veel belang hechten aan het gebruik van Standaardnederlands in formele contexten. In de context van sociale media overheerst echter onverschilligheid ten aanzien van klassieke taalnormen.

Tot slot biedt de enquête inzicht in de **appreciatie van of tolerantie tegenover bepaalde prototypische kenmerken van online taalgebruik**. De resultaten stroken met de geobserveerde patronen in het corpus, en suggereren dat tieners bepaalde talige kenmerken in verschillende mate gebruiken vanuit een verschil in appreciatie voor die features. Verder blijken de jongeren sterk gelijklopende meningen te hebben wat de (potentiële) negatieve connotaties van bepaalde niet-verbale chatkenmerken betreft, al blijken meisjes en theoretisch geschoolde tieners wel iets gevoeliger te zijn voor de sociale indexicaliteit van deze kenmerken, en lijkt die gevoeligheid ook toe te nemen met leeftijd.

BELANG VAN DE BEVINDINGEN EN PISTES VOOR VERDER ONDERZOEK

De bevindingen van het proefschrift zijn in verschillende opzichten van belang. Zo draagt de algemene methodologie, en in het bijzonder de toevoeging van interacties tussen sociale variabelen, bij tot een genuanceerder inzicht in de impact van sociale factoren op het online taalgebruik van jongeren. Traditionele sociolinguïstische patronen, die sterk aanwezig zijn in de data, blijken immers subtieler en complexer te zijn wanneer deze interacties mee in overweging worden genomen. Zo zijn niet alleen typische genderpatronen waar te nemen in het corpus, maar vertoont het online taalgebruik van meisjes een grotere variatiebreedte (gerelateerd aan leeftijd en studierichting) dan dat van jongens. Ook de toevoeging van sociale-klasse-indicatoren als studierichting en het beroep van de ouders als sociale variabelen is vernieuwend. Sociologisch relevant is de bevinding dat de studiekeuze van jongeren uit de boven- en arbeidersklasse zeer sterk sociaal bepaald is, wat de relevantie van

het concept ‘sociale klasse’ in onze huidige samenleving onderstreept. Verder biedt het proefschrift een positieve bijdrage tot het debat over de mogelijke negatieve invloed van informeel online taalgebruik op de traditionele geletterdheid van jongeren door te suggereren dat tieners – ongeacht hun profiel – over een sterke registergevoeligheid beschikken. Over het taalgebruik van jonge tieners en BSO-leerlingen in het bijzonder wordt vaak negatief bericht, maar onze resultaten tonen aan dat ook zij registergevoelig zijn, maar in hun chatberichten typografische expressie lijken te verkiezen boven traditionele verbale expressie. Verder blijkt uit de resultaten dat BSO-jongeren minstens evenzeer aansluiting vinden bij de interactieve online schrijfcultuur als ASO-jongeren.

Enkele mogelijke pistes voor verder onderzoek zijn het ontwikkelen van een nog preciezere implementering van sociale klasse, met meer aandacht voor mensen met een ‘hybride’ sociaal profiel. Ook een verbetering van het classificatiemodel voor studierichting vormt relevant vervolgonderzoek, evenals de studie van conversationele variabelen die buiten de scope van het huidige onderzoeksproject vielen. Een prioriteit voor vervolgonderzoek is namelijk de analyse van accommodatiepatronen in online taalgebruik.