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

Lisa Hilte, Reinhild Vandekerckhove and Walter Daelemans
University of Antwerp

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, p. 184; Peersman, Daelemans, R. Vandekerckhove, B. Vandekerckhove & Van Vaerenbergh, 2016) – due to peer "group pressure to not conform to established societal conventions" (Nguyen, Doğruöz, Rosé, & De Jong, 2016, p. 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 vari-

https://doi.org/10.1075/dujal.17018.hil
Dutch Journal of Applied Linguistics 7:1 (2018), pp. 2–25. issn 2211-7245 | e-issn 2211-7253
© John Benjamins Publishing Company

eties which more closely approach the standard language] to be taken seriously" (Nguyen et al., 2016 – our insertion, p. 17). 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, pp. 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, pp. 31–32; Kucukyilmaz, Cambazogly, Aykanat, & Can, 2006, p. 282; Parkins, 2012, pp. 48, 50, 53; Wolf, 2000, p. 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) and Trudgill (1983b) 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, p. 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, p. 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, p. 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

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

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

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, p. 157)

As there is not *one* correct way to define class (Braham, 2013, p. 30; Crompton, 2010, p. 155), we treat it as a multidimensional variable with several subfactors, in order to represent its 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, p. 253). As today's society has evolved towards a knowledge-based *meritocracy* – i.e. "social stratification based on personal merit" (Macionis, 2011, p. 206) – education and obtained degrees have become an increasingly important aspect of social status and position (De Jager et al., 2009, pp. 243, 247). In the present case study, we include the three main levels of Belgian secondary education[4] (FMET, 2017, p. 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.

[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, p. 155–157; Goldthorpe & Breen 2007, p. 25), globalization, an increase in female employment (Crompton, 2010, p. 159), as well as a growing importance of education and knowledge (Goldthorpe & Breen, 2007, p. 45; De Jager, Mok, & Sipkema, 2009, p. 243), which will be discussed more elaborately in the next paragraph.

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

- *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, pp. 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, pp. 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:

- 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 & 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, p. 420; Vranken, Van Hootegem, Henderickx, & Vanmarcke, 2017, p. 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, p. 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.

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, pp. 319–325), which is confirmed by our dataset (see Table 1).[6] Although in theory, Flemish children can choose any educational track regard-

[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).

[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.

less 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, p. 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, p. 254; Vranken et al., 2017, pp. 314–315, 319). In general, parents want to avoid downward mobility for their children (Goldthorpe & Breen, 2007, p. 53).

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

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

In our dataset, we find a significant and strong correlation between the participants' level of education and their parents' profession category ($X^2 = 99.638$,

they would fill in 'restaurant', or 'harbor', or the name of a company, without specifying their parent's job).

$p < 0.0001$, 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'.

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

		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

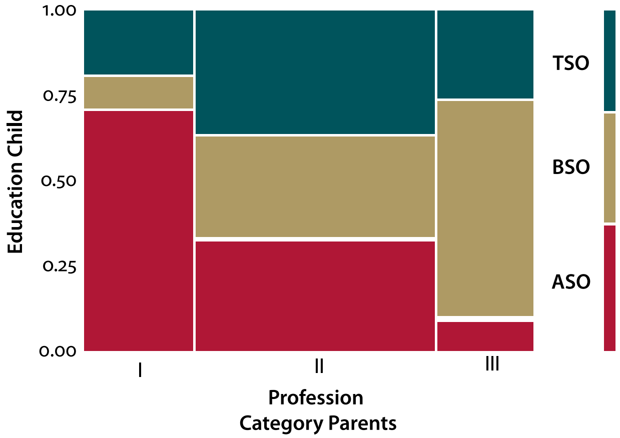

Figure 1. Mosaic plot of the correlation between the adolescents' level of education and their parents' profession

Figure 1 visualizes the distribution for the different education and profession types. The shares of 'extreme' social mobility (cat. I and Vocational Education/

[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.

BSO, cat. III and General Education/ASO) are smallest. For all education types, stagnation is very frequent, with the 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) ($X^2 = 23.249$, $p < 0.0001$, 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 ($X^2 = 16.138$, $p = 0.0028$, 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, p. 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** nie**t** 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).

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

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**

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).

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

[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.

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.

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

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

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.

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

	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)		**$p<0.0001$**	**$p<0.0001$**	**$p<0.0001$**
X^2		338.353	26,518.16	5,993.251
Strength of correlation (Cramer's V)		1.08%	9.59%	4.56%

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.

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

	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)		$p < 0.0001$	$p < 0.0001$	$p < 0.0001$
X^2		633.358	7,914.388	816.783
Strength of correlation (Cramer's V)		1.48%	5.25%	1.69%

5.3 Profession of the parents

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.

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

	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)		$p < 0.0001$	$p < 0.0001$	$p < 0.0001$
X^2		249.098	1,282.129	1,817.297
Strength of correlation (Cramer's V)		1.27%	2.87%	3.42%

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.

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

	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)		$p<0.0001$	$p<0.0001$	$p<0.0001$
X^2		324.240	1,879.366	1,916.853
Strength of correlation (Cramer's V)		3.61%	8.70%	8.79%
Effect size (odds ratio)		2.37	2.42	1.93

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

[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.

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.

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" (p. 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, p. 152; Vandekerckhove, Cuvelier, & De Decker, 2015, p. 355), like in *nrml* (standard Dutch: *normaal,* 'normal'). The third category consists of standard Dutch or English words containing genuine typing or spelling mistakes, like *vrined* instead of *vriend* ('friend' – typing error) or *abbonement* instead of *abonnement* ('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

[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*.

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, p. 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.

Acknowledgements

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

References

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

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

Braham, P. (2013). *Key concepts in sociology*. Los Angeles, CA: Sage.

Coates, J. (1993). Quantitative studies. In J. Coates, *Women, men and language. A sociolinguistic account of gender differences in language* (pp. 61–86). London: Longman.

Crompton, R. (2010). The rise, fall and rise of social class. In A. Giddens & P. W. Sutton, *Sociology: Introductory readings* (3rd ed., pp. 154–160). Cambridge: Polity Press.

De Decker, B., & Vandekerckhove, R. (2012). De mythe van dialectrevival. In S. Kindt, P. Dendale, & A. Vanderheyden (Eds.), *La langue mise en contexte: Essais en l'honneur d'Alex Vanneste* (pp. 27–46). Maastricht: Shaker.

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

De Jager, H., Mok, A. L., & Sipkema, G. (2009). *Grondbeginselen der sociologie*. Groningen: Noordhoff.

Eckert, P. (2000). *Linguistic variation as social practice*. Malden, MA: Blackwell.

Erikson, R., Goldthorpe, J. H., & Portocarero, L. (1979). Intergenerational class mobility in three Western European societies: England, France and Sweden. *The British Journal of Sociology*, 30(4), 415–441. https://doi.org/10.2307/589632

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

[FMET] Flemish Ministry of Education and Training/Vlaams ministerie van onderwijs en vorming. (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).

Goldthorpe, J. H., & Breen, R. (2007). Explaining educational differentials. Towards a formal rational action theory. In J. H. Goldthorpe, *On Sociology, Vol. Two: Illustration and retrospect* (2nd ed., pp. 45–72). Stanford, CA: Stanford University Press.

Hilte, L., Vandekerckhove, R., & Daelemans, W. (2016). Expressiveness in Flemish online teenage talk: A corpus-based analysis of social and medium-related linguistic variation. In D. Fišer & M. 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, J. (1992). *An introduction to sociolinguistics*. London: Longman.

Kucukyilmaz, T., Cambazogly, B. B., Aykanat, C., & Can, F. (2006). Chat mining for gender prediction. In T. Yakhno, & E. J. Neuhold (Eds.), *Advances in information systems. ADVIS 2006. Lecture Notes in Computer Science* (pp. 274–283). Berlin: Springer.

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

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

Nguyen, D., Doğruöz, A. S., Rosé, C. P., & De Jong, F. (2016). Computational sociolinguistics: A survey. *Computational Linguistics*, 42(3), 537–593.

Parkins, R. (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, C., Daelemans, W., Vandekerckhove, R., Vandekerckhove, B., & Van Vaerenbergh, L. (2016). The effects of age, gender and region on non-standard linguistic variation in online social networks. *Arxiv*, 26 January 2016. Retrieved from <http://arxiv.org/abs/1601.02431> (20 September 2016).

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

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

Vandekerckhove, R., Cuvelier, P., & De Decker, B. (2015). The integration of English in Flemish versus African online peer group language: A comparative approach. *Language Matters*, 46(3), 344–363. https://doi.org/10.1080/10228195.2015.1089925

Vranken, J., Van Hootegem, G., Henderickx, E., & Vanmarcke, L. (2017). *Het speelveld, de spelregels en de spelers? Handboek sociologie*. Den Haag: Acco.

Wolf, A. (2000). Emotional expression online: Gender differences in emoticon use. *Cyberpsychology & Behavior*, 3(5), 827–833. https://doi.org/10.1089/10949310050191809

Address for correspondence

Lisa Hilte
University of Antwerp
Prinsstraat 13
B-2000 Antwerpen
België
lisa.hilte@uantwerpen.be

Co-author information

Reinhild Vandekerckhove
University of Antwerp
reinhild.vandekerckhove@uantwerpen.be

Walter Daelemans
University of Antwerp
walter.daelemans@uantwerpen.be