Modeling Non-Standard Language Use in Adolescents' CMC: The Impact and Interaction of Age, Gender and Education

Lisa Hilte, Reinhild Vandekerckhove, Walter Daelemans
CLiPS Research Center, University of Antwerp
Postal address: Prinsstraat 13, B-2000 Antwerp, Belgium
E-mail: {lisa.hilte, reinhild.vandekerckhove, walter.daelemans}@uantwerpen.be

Abstract

The present paper deals with Flemish adolescents' informal computer-mediated communication (CMC) in a large corpus (2.9 million tokens) of chat conversations. We analyze deviations from written standard Dutch and possible correlations with the teenagers' gender, age and educational track. The concept of non-standardness is operationalized by means of a wide range of features that serve different purposes, related to the chatspeak maxims of orality, brevity and expressiveness. It will be demonstrated how the different social variables impact on non-standard writing, and, more importantly, how they interact with each other. While the findings for age and education correspond to our expectations (more non-standard markers are used by younger adolescents and students in practice-oriented educational tracks), the results for gender (no significant difference between girls and boys) do not: they call for a more fine-grained analysis of non-standard writing, in which features relating to different chat principles are examined separately.

Keywords: computer-mediated communication, non-standardness, teenage talk, language modeling

1. Introduction

Adolescents' informal CMC tends to deviate from formal standard writing in many ways: alternative spelling, non-standard capitalization, emoticons, … These deviations can be related to the three main principles behind chatspeak, i.e. the principles of orality, economy and expressive compensation (Androutsopoulos, 2011, 149; see section 3.2 for definitions and examples). While many CMC-studies report on just one type of features or present a small selection, the present study examines a wide array of 11 non-standard features and relates their frequency to three independent variables.

In the following sections, we will describe the goal of this study (section 2), as well as the dependent and independent variables (section 3). Next, we present the corpus and methodology (section 4), and finally, we will discuss and evaluate the results (section 5).

2. Goal of the Paper

We try to capture the impact of three aspects of the adolescent authors' profile on their CMC writing practices: their gender, age, and educational track. The latter variable has been largely neglected in CMC research. The same accounts for potential interactions between these variables: as boys and girls age, do their online writing practices evolve in a similar way? And do the same age and gender patterns emerge in different education types? In the end we want to demonstrate that the inclusion of a wide range of both independent and dependent variables is a prerequisite for a correct assessment of variation patterns in adolescents' CMC.

3. Dependent and Independent Variables

3.1. Independent Variables

All participants are high school students living in Flanders, the Dutch-speaking part of Belgium. We examine three social variables: the adolescents' gender, their age and their type of education (i.e. educational track).

Both gender and age are treated as binary variables: boys are compared to girls, and younger teenagers (13-16) to older ones (17-20). For educational track, we distinguish the three main types of secondary education in Belgium: ASO, TSO and BSO. ASO or General Secondary Education is theory-oriented and prepares students for higher education, whereas BSO or Vocational Secondary Education is practice-oriented, preparing students for a manual profession. TSO or Technical Secondary Education constitutes a more hybrid in-between level.

3.2. Dependent Variables

We selected 11 different linguistic features which are all deviations from the formal writing standard.

The largest set of features consists of 7 expressive markers which convey emotional or social involvement (see Hilte, Vandekerckhove & Daelemans, 2016 for a detailed analysis of these expressive markers):

1. non-standard capitalization
   e.g. *IK ZWEER HET* 'I swear'
2. emoticons and emoji
   e.g. *dammn we look so hot* 😍🔥💋❤
3. combinations question and exclamation marks
   e.g. *Echt**?!*** 'Really?!'
4. deliberate repetition ('flooding') of letters
   e.g. ***yeeeesss*** 'yes'
5. deliberate repetition ('flooding') of punctuation marks
   e.g. *Wat**???*** 'What???'

6. onomatopoeic rendering of laughter
   e.g. *hahaha*
7. rendering of kisses and hugs
   e.g. *Dankje **xxx*** 'Thank you xxx'

For orality (i.e. the underlying chatspeak principle to write 'as you speak'), we take one feature into account:

8. non-standard Dutch lexemes (informal Dutch, ranging from colloquial speech, regiolect/dialect words and slang to youth language ...)
   e.g. *vertel het **sebiet*** (std. Dutch: *vertel het **straks**,* 'tell it later')

The economy principle ('make your message as concise as possible') was operationalized with:

9. chatspeak abbreviations and acronyms (i.e. non-standard shortened words or phrases)
   e.g. ***Omg** yes* (full version: '**Oh my god** yes')

The final category contains two features that do not really fit into one of the three chat principle categories but are characteristic of (Dutch) CMC and atypical of formal standard writing:

10. English words[1] (used in a Dutch conversation)
    e.g. *echt **nice*** 'really nice'
11. Discourse markers # (*hashtag*, to indicate a topic or express a feeling about it) and @ (*at*, to directly address one person in a group conversation)
    e.g. ***#**bestfriends*
    e.g. ***@**sarah*

4. Corpus and Methodology

4.1 Corpus

The corpus consists of Flemish teenagers' informal chat conversations and contains 2 885 084 tokens or 488 014 posts. The number of chatters in the corpus is 1384. The distributions for the social variables age, gender and education can be found in Table 1. We note that (dialect) region is a quasi-constant: almost all tokens (over 96%) are collected from participants living in the central Antwerp-Brabant region. 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. Students consented to donate their conversations, and for minors, parents' consent was asked too. All chat material was anonymized before analysis – the participants' names were replaced by serial numbers, which are linked to the features of their social profiles (e.g. gender).

Variable	Subgroups	Tokens
Gender	Boys	985 928 (34%)
	Girls	1 899 156 (66%)
Age	Younger (13-16)	1 584 373 (55%)
	Older (17-20)	1 300 711 (45%)
Education	General (ASO)	920 114 (34%)
	Technical (TSO)	1 213 483 (42%)
	Vocational (BSO)	751 487 (26%)
Total		2 885 084

Table 1: Distributions for gender, age and education.

4.2. Methodology

4.2.1. Feature Extraction

All feature occurrences were extracted automatically using Python scripts. For a random test set, the software's output was compared to human annotation, which rendered a satisfying f-score of 0.90 (average for all 11 features).

4.2.2. Statistical Language Modeling

We statistically analyze the use of non-standard features by constructing a generalized linear mixed model (GLMM) on a token-level, using the R package 'lme4' (Bates et al., 2017). The GLMM tries to model and predict the response variable, which is the probability of a token containing at least one non-standard feature. As a random effect, we add the chatters' ID to account for individual variation between the participants as well as for their unbalanced contributions.

5. Results

5.1 Modeling Non-Standardness

Our model of non-standard language use (conceptualized in a binary way, i.e. the probability that a token contains at least one non-standard feature) lets the three social factors age, gender and education interact with each other, while adding a random effect for individual variation among the chatters. Table 2 shows the raw output of the model, i.e. the estimates and significance scores for the different levels of the factors, always in comparison to the reference category (older teenage boys in the General Education System/ASO). To evaluate a factor's significance as a whole (and not just in comparison to the reference group, but to all other levels as well), we performed extra Anova analyses. These results are shown in Table 3. Furthermore, we added extra effect tests using the 'Effects' package in R (Fox et al., 2016; Fox, 2003).

[1] Although the use of English words in Dutch conversations could also be related to the orality principle, we argue that it should be dealt with separately, since it is indicative of the extent to which youngsters connect with international chat culture.

	estimate	std. error	z	p	sig.
(Intercept)	-1.18023	0.03466	-34.06	< 2e-16	***
ageYoung	0.07991	0.02159	3.70	0.000215	***
genderGirls	-0.15825	0.04613	-3.43	0.000603	***
eduBSO	0.10890	0.05611	1.94	0.052263	.
eduTSO	0.03414	0.05153	0.66	0.507682	
ageYoung: genderGirls	0.24210	0.02486	9.74	< 2e-16	***
ageYoung: eduBSO	0.03532	0.04537	0.78	0.436242	
ageYoung: eduTSO	0.06508	0.02485	2.62	0.008826	**
genderGirls: eduBSO	0.17528	0.07487	2.34	0.019227	*
genderGirls: eduTSO	0.06882	0.07001	0.98	0.325614	
ageYoung: genderGirls: eduBSO	-0.09562	0.04986	-1.92	0.055141	.
ageYoung: genderGirls: eduTSO	-0.06868	0.02905	-2.36	0.018077	*
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Table 2: Output of the GLMM in comparison to the reference level (older boys ASO).

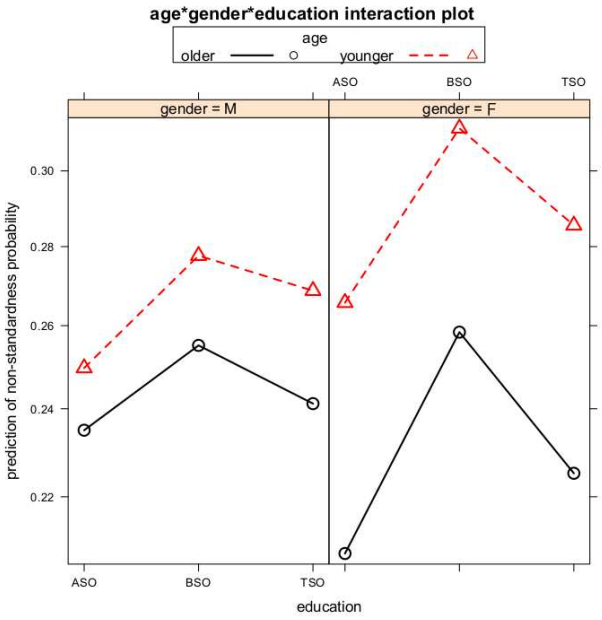

Figure 1: Interaction age*gender*education for non-standardness.

	Chisq	Df	Pr(>Chisq)	Sig.
age	2207.4925	1	< 2.2e-16	***
gender	0.4481	1	0.503223	
education	30.1873	2	2.786e-07	***
age:gender	232.6871	1	< 2.2e-16	***
age:education	10.6975	2	0.004754	**
gender: education	3.0855	2	0.213793	
age:gender:education	6.5998	2	0.036887	*
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 3: Output of the GLMM's Anova.

5.2. Effects and Interpretation

The different effects captured by the model are visualized in Figure 1 (i.e. the plot of the three-way interaction), on which the predicted probabilities for non-standardness are plotted for the different social variables.

The red dotted lines, representing the younger teenagers, are consistently higher than the black solid ones, representing the older teenagers, across all gender and education groups. This is a very consistent main effect for age, which is also significant (see Table 3, also confirmed by additional effect tests in which the other variables are kept constant): the younger teenagers (13-16 years old) use significantly more non-standard features than the older ones (17-20 years old). These results correspond to our expectations: non-standard language use is said to peak during adolescence, around the age of 16 ('the adolescent peak' – which is also the boundary between our two age categories) and thus decreasing as the teenagers age (Holmes, 1992, 184).

The two panels in Figure 1 represent the two genders, with the males on the left and females on the right. Clearly, at a younger age (compare the red dotted lines), girls outperform boys in non-standardness in each education type. However, this is no longer true at an older age, where girls only very slightly outperform boys in the Vocational System but use fewer non-standard features in the General and Technical Education Systems. The Anova (Table 3) and additional effect tests reveal that there is no significant *main* effect for gender, i.e. the model does not predict significantly different probabilities for non-standard features for girls compared to boys. However, the interactions between gender and age and between gender, age and education are significant. Consequently, gender is still an important factor in the model, as it is part of higher-order (interaction) terms which significantly impact on the response variable: in other words, in order to truly capture the gender effect, age and education have to be included in the analyses. As for the interaction between age and gender, Figure 1 shows that the decrease in non-standardness as the adolescents age is much stronger for the girls than for the boys. Again, these results correspond to our expectations, as in previous research, girls were found to converge more towards the adult standard as they grew older than boys (see Eisikovits, 2006, 43-44). Eisikovits ascribes this different age pattern to a difference between (working class) boys' and girls' attitude towards society when they graduate from high school; while accepting the responsibilities of adulthood, girls converge towards mainstream societal norms, whereas boys more strongly insist on their autonomy (2006, 48-49). We note that these preference patterns are confirmed for middle class participants by Vandekerckhove (2000, 302).

Finally, the Anova (Table 3) and additional effect tests reveal a significant main effect for type of education. The separate data points in Figure 1 reveal a consistent pattern across gender and age groups: the lowest probability of non-standardness is predicted for the teenagers in the General (theoretical) System (ASO), followed by the ones in the Technical (hybrid) System (TSO), and then by the ones in the Vocational (practical) System (BSO). Furthermore, additional effect tests showed that all three types are significantly different from each other. A possible explanation for these results concerns the level of proficiency in and familiarity with written standard Dutch in the different education types, which might increase as the school type becomes more theoretical. Apart from linguistic *skills*, *attitudinal* differences might be a factor too, as the prototypical chatspeak features may simply be more popular and considered to be *cooler* among students in the Vocational System. (For a more thorough analysis, see Hilte, Vandekerckhove & Daelemans, *fc*) Finally, the differences between education types are larger for the girls than for the boys. This could indicate a higher sensitivity for girls for this social factor.

Below, we present an alternative way to visualize the effects captured by the model. Figure 2 facilitates grasping the different 'age*gender' interactions in the three school systems. Clearly, in the more theoretical education types (General and Technical Education / ASO resp. TSO), the gender effect is opposite in the two age groups. At a younger age, the girls outperform the boys in non-standardness, but at a later age, they use fewer non-standard markers. In the Vocational System (BSO), however, the girls outperform the boys in non-standardness at a younger age and use more or less the same number of non-standard markers at an older age. Although there is still an interaction (girls' use of non-standard features decreasing more strongly than boys'), it is much less outspoken than the 'classical' pattern in the other two education types, and results in a convergence of the two genders rather than in an (opposite) divergence.

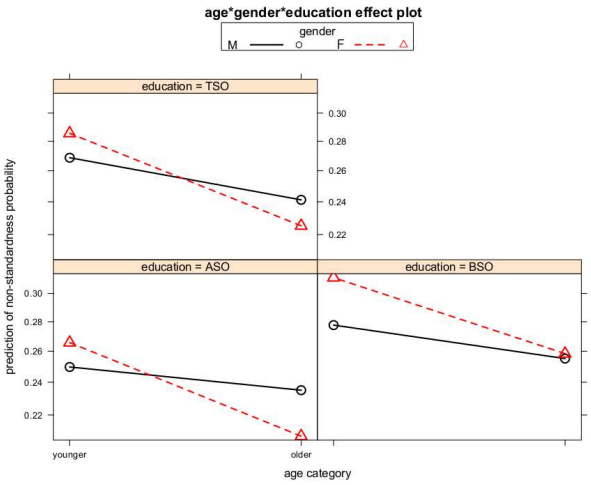

Figure 2: Interaction age*gender*education for non-standardness, alternative visualization.

6. Conclusion

We modeled Flemish adolescents' non-standard language use in their informal computer-mediated communication. We found that age, gender and education interact and influence the occurrence of non-standard features. Whereas the impact of age (lower frequencies in older teenagers' CMC) and education (lower frequencies for students in more theoretical educational tracks) might confirm expectations based on related research, the gender findings are quite surprising. The observation for the main effect of gender (i.e. no significant difference) does not correspond to previous research, as female language use is generally found to be more 'standard-oriented'.

However, this might be related to the operationalization of the notion of non-standardness in our research design: clearly expressive markers, which appear to be highly favored by women (see Hilte, Vandekerckhove & Daelemans, 2016, 31-32), might behave completely different in terms of indexing non-standardness from markers of regional non-standard speech. Consequently, a priority for future research will be the declustering of the set of 'non-standard' features and the consequent construction of different models for each subset, so that potential different preference patterns for these subsets can emerge. Still then, as we have shown in this preliminary study, gender cannot be studied in isolation, since the interactions with age and education are a prerequisite for a correct and nuanced evaluation of its impact.

7. Acknowledgements

We thank Giovanni Cassani, Dominiek Sandra and Koen Plevoets for their help and advise with the statistical modeling.

8. References

Androutsopoulos, J. (2011). Language Change and Digital Media: A Review of Conceptions and Evidence. In T. Kristiansen & N. Coupland (Eds.), *Standard Languages and Language Standards in a Changing Europe*. Oslo: Novus Press, pp. 145--161.

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

Eisikovits, E. (2006). Girl-Talk/Boy-Talk: Sex Differences in Adolescent Speech. In J. Coates (Ed.), *Language and Gender: A Reader*. Oxford: Blackwell, pp. 42--54.

Fox, J. (2003). Effect Displays in R for Generalised Linear Models. *Journal of Statistical Software*, 8(15).

Fox, J., Weisberg, S., Friendly, M., Hong, J., Andersen, R., Firth, D., & Taylor, S. (2016). Package 'effects'. Url: http://www.r-project.org, http://socserv.socsci.mcmaster.ca/jfox/

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.

Hilte, L., Vandekerckhove, R., & Daelemans, W. *(forthcoming).* Adolescents' Social Background and Non-Standard Writing in Online Communication.

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

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), pp. 46--54.

Vandekerckhove, R. (2000). Structurele en sociale aspecten van dialectverandering. De dynamiek van het Deerlijkse dialect. Gent: Koninklijke Academie voor Nederlandse Taal- en Letterkunde.

Wolf, A. (2000). Emotional Expression Online: Gender Differences in Emoticon Use. *Cyberpsychology & Behavior,* 3(5), pp. 827--833.