Expressiveness in Flemish Online Teenage Talk: A Corpus-Based Analysis of Social and Medium-Related Linguistic Variation

Lisa Hilte, Reinhild Vandekerckhove, Walter Daelemans
CLiPS, University of Antwerp
E-mail: lisa.hilte@uantwerpen.be, reinhild.vandekerckhove@uantwerpen.be, walter.daelemans@uantwerpen.be

Abstract

We analyze linguistic expressiveness in an extensive corpus (2 million tokens) of Flemish online teenage talk, focusing on the use of typographic chatspeak features, an onomatopoeic and a lexical variable and its correlation with the chatters' profile and the online medium. General quantitative findings are that girls outperform boys in the expression of emotional involvement, and younger adolescents outperform the older group. However, medium has the largest impact: much more expressive markers are used in asynchronous social media posts than in synchronous instant messaging. On a qualitative level, utterances written by girls, by younger teenagers and on the asynchronous platform contain more expressive markers related to love or friendship.
Apart from the medium's (a)synchronicity and its public or private character, the nature of the interaction appears to be a determining factor too. The asynchronous social media posts involve a lot of flirting or pleasing, which drastically increases linguistic expressiveness.

Keywords: computer-mediated communication, adolescents, computational sociolinguistics

1. Introduction

Since the rise of informal computer-mediated communication (CMC), both laymen and linguists have been fascinated by the prototypical features that they identified in several forms of digital writing (see Crystal, 2001). Androutsopoulos relates these features to three dimensions or themes: "orality, compensation, and economy" (2011: 149). While orality refers to the use of spoken language features in written discourse 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; Androutsopoulos, 2011: 149). The latter dimension is at issue in the present paper, which examines the use of expressive markers in Flemish online teenage talk.

2. Goal of the Paper

We examine social and medium-related linguistic variation concerning expressiveness in a corpus of Flemish online teenage talk. The linguistic variables include several typographic features that are generally associated with chat discourse (e.g. emoticons), an onomatopoeic variable (rendition of laughter) and a lexical variable (intensifiers[1]). All features will be discussed more elaborately in section 3. We investigate the potential (quantitative and qualitative) correlations between the use of the selected expressive markers and the profile of the chatters (in terms of age and gender) as well as the impact of the synchronicity and (largely) public versus private character of the medium on which the utterances were written.

3. Expressive Markers

First of all, the present study includes six typographic expressive markers:

- flooding (i.e. deliberate, expressive repetition) of letters
  e.g. *s**uuu**per*
- flooding of punctuation marks
  e.g. *nice**!!!***
- combinations of exclamation and question marks
  e.g. *wtf**?!?***
- capitalization of words or entire utterances
  e.g. ***FAIL***
- emoticons
  e.g. *dude* ***:P***
- typographic rendering of kisses or hugs and kisses
  e.g. ***Xxxx***

The onomatopoeic marker studied in this research is the rendering of laughter in CMC, which includes all variants of *haha* and *hihi*.

e.g. ***hahahaha***

Finally, we added a lexical variable, i.e. the use of intensifiers: "items that amplify and emphasize the meaning of an adjective or adverb" (Stenström, Andersen & Hasund, 2002: 139). In Dutch, these items can either be adverbs or intensifying prefixes.

e.g. ***Super**mooie t-shirt* '**super** nice T-shirt'

4. Corpus and Methodology

4.1. Corpus

Our corpus consists of 400 808 online messages or 2 066 521 tokens[2]. The messages were produced between 2007 and 2013 by adolescents from Dutch-speaking northern Belgium (Flanders), all aged between 13 and 20 years old. The utterances were written on both a synchronous electronic medium (private instant messaging) and an asynchronous electronic medium (private and public messages on a social media site). Table 1 shows the distribution of the tokens over the age and gender groups

[1] We sincerely thank Jens Vercammen for the data processing for this variable.

[2] These tokens are the result of splitting the text on whitespace. A token can be a word, but also an emoticon or isolated punctuation marks.

and the two media. We note that, although there is an imbalance 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.

	GIRLS		BOYS		
	YOUNGER	OLDER	YOUNGER	OLDER	**total**
SYNC.	118 694	176 233	29 146	973 061	1 297 134
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 1: Distribution of variables in the corpus.

4.2. Methodology

The typographic and onomatopoeic expressive markers were automatically detected and counted using Python scripts. The coverage of the software was evaluated and judged accurate on a test set of 1000 randomly chosen posts from the corpus by comparing a human annotator's feature extraction to the software's output. The intensifiers were automatically extracted using a predefined list[3] (which covered most of the intensifiers used in the corpus) and a frequency cutoff to not take into account very infrequent variants. The software's output was manually screened and filtered. To evaluate the human judgment, finally, a test set of 700 utterances was screened by two annotators, who obtained a low error rate (1.57%).

5. Results and Discussion

To verify the statistical significance of our quantitative findings, we combined chi square tests with a bootstrapping approach (with Monte Carlo resampling), to obtain more solid results than when performing one single chi square test on the entire data set[4]. The statistical values we report in the next paragraphs (p-values, Cramer's V scores and odds ratios) are the mean of the values for all samples.

5.1. Quantitative Findings

We quantified the degree of expressiveness by counting all markers in the subcorpora and dividing these counts by the number of tokens in the subcorpora. 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% (in terms of tokens – in terms of types: 21 427 markers, or a ratio of 11.88%). An overview of the ratios per independent variable is shown in Table 2. The asynchronous posts contain the highest relative number 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%
Asynchronous posts	Synchronous posts
28.35%	5.94%

Table 2: Overview of expressiveness ratios per subcorpus.

General tendencies for the social variables are that the girls use significantly more expressive markers than the boys ($p < .001$), that younger teenagers use significantly more expressive features than older ones ($p < .001$) and that significantly more expressive writing is used on asynchronous media ($p < .001$). These general tendencies also hold for each of the analyzed expressive markers: the female (resp. younger, resp. async.) texts contain *each* expressive marker significantly more often than the male (resp. older, resp. sync.) texts.

As for the strength of the correlation between the linguistic and independent variables, the strongest correlation can be found for medium (Cramer's V = 0.31), followed by age (Cramer's V = 0.24) and gender (Cramer's V = 0.17). The same order can also be found for effect size: medium has the largest effect size (odds ratio = 6.27), followed by age (odds ratio = 4.02) and gender (odds ratio = 2.71). These scores should be interpreted as follows: the odds that a token contains an expressive marker are 6.27 times higher if the token is produced on the asynchronous platform than when produced on the synchronous platform[5]. Medium seems to be the most interesting independent variable when it comes to expressiveness, as the correlation with the linguistic variables is very high and the actual effect size is large as well.

Some expressive features both heavily correlate with the social variables and are used very differently (quantitatively) by the subgroups of the same social variable. This is the case for letter flooding (i.e. deliberate, expressive letter repetition) and the rendition of kisses (e.g. 'xxx'), especially with regards to medium. The odds ratios are respectively 51.85 (kisses – medium) and 16.33 (letter flooding – medium): for each occurrence of kisses (flooding letters, resp.) in the synchronous chat messages, 51.85 occurrences (16.33, resp.) can be expected in the asynchronous posts.

5.2. Qualitative Findings

On a qualitative level, some constants could be found among all different subgroups. The most popular expressive markers in all groups are emoticons and punctuation flooding (deliberate repetition of question and exclamation marks). These features' popularity could be

[3] 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.*

[4] We thank Giovanni Cassani and Dominiek Sandra for their help and advice in the statistical aspect of the research.

[5] Note that these numbers differ from the ratios reported in Table 2. 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 ratio and not to calculate the straightforward percentages.

due to their 'explicit' expressive nature: many emoticons represent facial expressions and question and exclamation marks are the most expressive punctuation marks. Apparently, because of the explicit nature of these features, they are very obvious and favored markers.

As for letter flooding, we note that in all subgroups, mainly vowels are repeated, and hardly ever plosives. This supports the hypothesis that flooding is the orthographic representation of an oral phenomenon (Darics, 2013: 144), i.e. the lengthening of sounds, which is easiest for vowels and impossible for plosives.

A third general tendency is the top position of the Dutch first person singular pronoun 'ik' (I) among the lexemes written in capital letters. As pronouns are function words, they are automatically used more frequently (Newman et al., 2008: 216; Pennebaker, 2011: 27). However, the top position of 'ik' could also be symptomatic of the fact that when the teenagers write in a very expressive way, they often talk about something personal. This finding also suggests that quite often entire utterances are written in capitals, as merely capitalizing function words would make less sense (although the chatters could, of course, only emphasize the word 'I' in their utterance to stress its importance).

Finally, the qualitative in-depth analyses for each of the expressive markers also lay bare correlations between the independent variables. Strikingly, similar tendencies could be noted for texts written by female participants, by younger teenagers, and on the asynchronous medium. These texts contain a lot more expressive markers related to love and friendship. The most popular emoticons were related to love (e.g. heart-emoticons: <3) and many of the top lexemes that were written in allcaps concerned love or friendship (e.g. 'LOVEYOU', 'BFF': *best friend forever*). These results are incongruent with male texts, the texts written by older adolescents or the synchronous posts. E.g.: While heart-emoticons were much favored by girls, they were at the bottom of the list of the emoticons produced by boys.

However, some caution might be needed when interpreting these correlations, as there is an imbalance in our dataset which could (partially) influence our results: many of the female participants are also younger adolescents, often writing on the asynchronous medium, whereas many of the male participants are also older teenagers, often writing on the synchronous chat platform. Still, linguistic correlations between gender and age have been reported on before (Argamon et al., 2007; Pennebaker, 2011; Schwartz et al., 2013). Stylistic correlations concern the use of function words: men and older people use more articles and prepositions, whereas younger people and women use more pronouns, conjunctions and auxiliary verbs (Pennebaker, 2011: 66; Argamon et al., 2007: n.pag.; Schwartz et al., 2013: 8-9). On a content-related note, 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.pag.). These findings correspond to the younger and female teenagers' preference 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 flirting and/or pleasing (e.g. in positive reactions to other users' pictures). In this respect, our asynchronous medium differs from other social media, like Twitter, where the writing is less personal and more targeted at informing a wider audience, rather than at bonding or pleasing[6]. The latter focus prevails in our asynchronous data, which could explain the higher rate of love-related expressive markers in this subcorpus.

6. Conclusion

This paper discussed linguistic expressiveness in (Belgian) Dutch informal 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 authors' profile (gender, age) versus the CMC medium. Girls appeared to outperform boys in the use of expressive markers, 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 than in synchronous and mainly private posts). 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 asynchronous than in the 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) and combines 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.

7. References

Androutsopoulos, J. (2011). Language Change and Digital Media: A Review of Conceptions and Evidence. In: T.

[6] We thank Lieke Verheijen for pointing out this difference.

Kristiansen & N. Coupland (Eds.), *Standard Languages and Language Standards in a Changing Europe.* Oslo: Novus, pp. 145--161.

Argamon, S., Koppel, M., Pennebaker, J.W., & Schler, J. (2007). Mining the Blogosphere: Age, Gender and the Varieties of Self-Expression. *First Monday,* 12(9), n.pag.

Baron, N.S. (1984). Computer Mediated Communication as a Force in Language Change. *Visible Language*, 18(2), pp. 118--141.

Crystal, D. (2001). *Language and the Internet.* Cambridge: Cambridge University Press.

Darics, E. (2013). Non-verbal Signalling in Digital Discourse: The Case of Letter Repetition. *Discourse, Context and Media,* 2, pp. 141--148.

Newman, M.L., Groom, C.J., Handelman, L.D. & Pennebaker, J.W. (2008). Gender Differences in Language Use: An Analysis of 14,000 Text Samples. *Discourse Processes*, 45(3), pp. 211--236.

Pennebaker, J.W. (2011). *The Secret Life of Pronouns. What Our Words Say About Us*. New York: Bloomsbury Press.

Schwartz, A.H., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M. et al. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*, 8(9), e73791.

Stenström, A.B., Andersen, G., & Hasund, I.K. (2002). Non-Standard Grammar and the Trendy Use of Intensifiers. In: A.B. Stenström, G. Andersen & I.K. Hasund, *Trends in Teenage Talk. Corpus Compilation, Analysis and Findings*. Amsterdam: John Benjamins, pp. 131--163.