




ARTICLE



<https://doi.org/10.1057/s41599-021-00968-7>

OPEN

How analysis of mobile app reviews problematises linguistic approaches to internet troll detection

Sergei Monakhov  ¹✉

State-sponsored internet trolls repeat themselves in a unique way. They have a small number of messages to convey but they have to do it multiple times. Understandably, they are afraid of being repetitive because that will inevitably lead to their identification as trolls. Hence, their only possible strategy is to keep diluting their target message with ever-changing filler words. That is exactly what makes them so susceptible to automatic detection. One serious challenge to this promising approach is posed by the fact that the same troll-like effect may arise as a result of collaborative repatterning that is not indicative of any malevolent practices in online communication. The current study addresses this issue by analysing more than 180,000 app reviews written in English and Russian and verifying the obtained results in the experimental setting where participants were asked to describe the same picture in two experimental conditions. The main finding of the study is that both observational and experimental samples became less troll-like as the time distance between their elements increased. Their 'troll coefficient' calculated as the ratio of the proportion of repeated content words among all content words to the proportion of repeated content word pairs among all content word pairs was found to be a function of time distance between separate individual contributions. These findings definitely render the task of developing efficient linguistic algorithms for internet troll detection more complicated. However, the problem can be alleviated by our ability to predict what the value of the troll coefficient of a certain group of texts would be if it depended solely on these texts' creation time.

¹Friedrich Schiller University, Jena, Germany. ✉email: sergei.monakhov@uni-jena.de

Introduction

Troll writing has enjoyed significant scholarly attention for a long period of time, starting from 1980s, when trolling was investigated within the frameworks of computer-mediated communication (Sia et al., 2002; Douglas and McGarty, 2001; Siegel et al., 1986) and hate speech (Carney, 2014; Chakraborti, 2010; Herring et al., 2002; Fraser, 1998), and up to the present day, when troll messages are scrutinised as a powerful weapon of disseminating propaganda in modern hybrid warfare (Lundberg and Laitinen, 2020; Zannettou et al., 2019; Elyashar et al., 2018; Egele et al., 2017; Volkova and Bell, 2016). Nowadays, an important task for the academic community is to provide a tool for identifying internet troll accounts as quickly and accurately as possible. Though it seems that such a task can be effectively fulfilled on purely linguistic grounds, until very recently, little work has been done that could help to explain the discourse-specific features of this type of writing.

In 2020, Monakhov showed that a number of features inherent in trolls' tweets are grounded in the sociolinguistic limitations of this type of discourse, which, in essence, is an imitation, make-believe game (Monakhov, 2020a, 2020b). Internet trolls want to achieve their goals without being identified as trying to achieve them. In other words, their language attitudes must be predefined and moulded by a combination of two factors: first, speaking with a purpose; second, trying to mask the purpose of speaking. Monakhov then contended that the orthogonal nature of these factors must necessarily result in the skewed distribution of different language parameters of trolls' messages and showed some very pronounced anomalies in the distribution of topics and associated vocabulary in Russian trolls' tweets.

This view seems to be intuitively clear if we agree that troll writing is characterised by the omnipresence of a target message (or a small cluster of such messages) underlying each and every concrete topic, however great the range. Suppose that a troll has to write a great number of messages using the word *vaccine*. It is not possible to simply continuously repeat the same tweet because that will lead to the exposure of the troll. Hence, it is necessary to use the target word in a variety of different contexts, including those where it may seem incongruous to most speakers. This, in turn, has consequences for the target word's lexical compatibility: its distribution markedly increases, its neighbours become more numerous, and the co-occurrence links between it and other words become artificially strengthened.

It means that, though troll messages are usually thought of as being highly repetitive, their most essential feature is anomalous distribution of repeated words and word pairs. This anomaly is inevitable because a task of delivering a target message multiple times without being suspected of such can only be performed by using a limited number of signal words in a wide variety of different contexts.

Building upon this theory, Monakhov proposed a simple and effective algorithm for the identification of troll writing, which was based on calculating the ratio q of the proportion of repeated content words among all content words to the proportion of repeated content word pairs among all content word pairs. He found that, regardless of the distribution of topics in tweets and the number of content words within a message, tweets written by trolls were characterised by greater values of q than tweets written by congresspeople and Donald Trump, which were used for comparison. The reason for this was that the denominator always had a higher value in the latter case than in the former, since repeated content word pairs were more frequent in non-troll writing (Monakhov, 2020a, 2020b).

Theoretically speaking, the algorithm that proved to identify troll messages with an accuracy of more than 98 % should work

just as well with other types of paid texts distributing false information. However, when it was tested on app reviews, one interesting phenomenon emerged. For analysis, we chose the users' reviews of Russian official Corona app 'Social monitoring' (<https://play.google.com/store/apps/details?id=ru.mos.socmon>) that was severely criticised in Russia for an outrageously bad design and poor functionality. Its current rating in the Google Play Store is 1.3, averaged across more than 7000 reviews. However, there is also some small amount of highly positive five-star reviews, the authors of which rebut all criticism and praise the effectiveness of the app. Since we had good reasons to believe that these positive reviews were suspicious, we checked them with Monakhov's algorithm and, in line with our expectations, they were identified as troll-like. Absolutely contrary to our expectations was the fact that strictly negative one-star reviews, which we analysed in the next turn, were also identified as troll-like.

It is highly unlikely that one-star reviews are actually not genuine but written for whatever reason by a group of paid authors. Puzzled by the findings and wanting to know whether it is this particular app or probably Russian language specific anomaly, we put to the same test all the one-star English reviews of the National Health Service official app (<https://play.google.com/store/apps/details?id=com.nhs.online.nhsonline>; rating of 3.1, averaged across more than 5200 reviews). The results were the same: they were identified as troll-like.

There appears to be only two logically consistent ways to account for this phenomenon: either the algorithm for some reason works only with tweets but not with app reviews, or the very communicative situation of posting an app review somehow leads to the emergence of collective 'troll effect'. The first explanation seems implausible: there are no apparent features that set apart, with regard to the numbers of repeated content words and word pairs, topically related tweets, on the one hand, and app reviews, on the other, as two varieties of short internet messages. The second explanation, however, is psychologically credible. One might assume that it is very common for internet commenters to first read what other people have written and only after that share their own opinion. Thus, the hypothesis that we want to test is that acquaintance with previous messages shapes each consequent message in what concerns the choice of words.

If it is observed that mobile apps reviews' troll coefficient can be approximated by some function of their creation time, we may conclude that there actually takes place some kind of reviews' repatterning that is not indicative of any malevolent practices in online communication. Thus, we expect to find a significant negative correlation between the value of the troll coefficient and the amount of time that separates different groups of reviews. It must be the case since people are more likely to take into account the reviews that have been written recently and not care too much about those written a longer time ago.

The rest of the paper is structured as follows. In Study 1, we present and discuss the results of the analysis of more than 180,000 app reviews written in English and Russian. In Study 2, we show how these results were verified in the experiment where participants were asked to describe the same picture in two experimental conditions. The General Discussion session is dedicated to the question of whether there exists a connection between the domains of troll writing and the creative aspect of language use.

Study 1: data and methods

In order to put our hypothesis to test, we downloaded from Google Play store all the one-star reviews of the 'Social

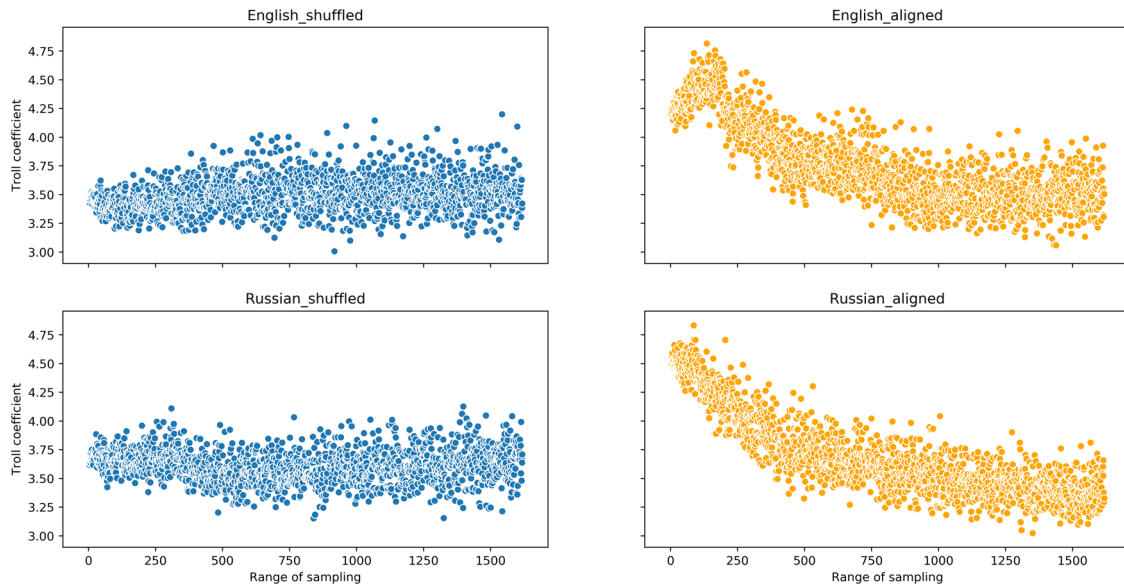


Fig. 1 Distribution of troll coefficients. Patterns observed in randomly shuffled and chronologically aligned English and Russian reviews.

monitoring’ app (4072 messages in Russian) and the NHS app (1987 messages in English). For each review we obtained its text as a sequence of stemmed content words and a Python datetime object specifying the exact date and time of its publishing. The reviews for each app were chronologically aligned, from the newest to the latest $\{r_1, \dots, r_n\}$, where r_1 is the first, most recent review and n is the total number of reviews. After that, we divided the data into a number of samples by means of the following procedure: (1) a sampling window of 250 reviews was chosen at the initial stage and the first 250 reviews were subsequently taken as sample $S_1 = \{r_1, \dots, r_{250}\}$; (2) at each subsequent step, the size of the sampling window was incremented by one element resulting in the family of sets $F = \{S_2, \dots, S_{n-250}\}$, where $S_2 = \{r_1, \dots, r_{250+1}\}$, $S_3 = \{r_1, \dots, r_{250+2}\}$ and so on; (3) from each set of the family F , 250 reviews were randomly sampled to make their troll coefficients comparable.

Next, for each sample in $\{S_1, \dots, S_{n-250}\}$, a measure of average time distance between different reviews within it was evaluated. The process of calculation was as follows: for each review in a sample, a number of days separating it from each other review in the same sample was obtained; then the absolute values of those numbers were summed up and divided by the number of pairwise comparisons.

The troll coefficients were obtained for each sample in accordance with the Monakhov’s formula:

$$q = \frac{w/W}{p/P},$$

where w is the number of repeated content words in a sample, W is the total number of content words, p is the number of repeated content word pairs, and P is the total number of content word pairs.

In Fig. 1, distributions of troll coefficients are plotted for both apps against the ever-widening ranges of sampling in chronologically aligned reviews (right-hand plots) and, for comparison, in randomly shuffled reviews (left-hand column). For better comparability, we trimmed the number of Russian samples to be equal to the number of English ones.

The Fig. 1 may be interpreted as follows. When reviews are chronologically aligned, increasing the range of sampling leads to the increase of average time distance between different reviews. That is why we find a very strong negative correlation between the average numbers of days separating reviews in a sample and the values of troll coefficients obtained for respective

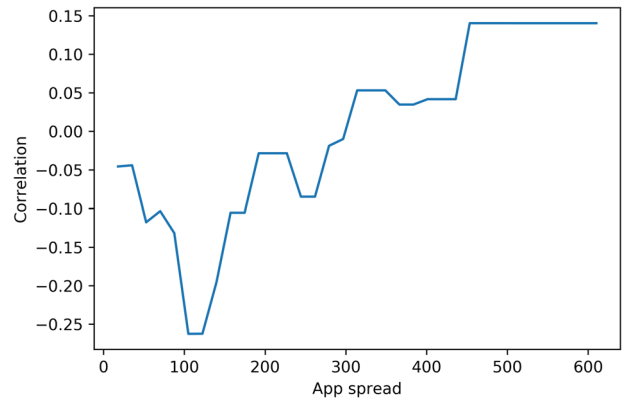


Fig. 2 Correlation between troll coefficients and average number of days for apps with different time spread.

samples ($r = -0.87$, $p < 0.0001$ for Russian data; $r = -0.84$, $p < 0.0001$ for English data). With randomly shuffled reviews, however, time distances become approximately equal for each sample, which results in uniform distribution of the values of troll coefficients.

The remarkable parallelism of English and Russian patterns seems to lend credence to our initial hypothesis. However, up till now, we only looked at the one-star reviews of the two apps from the ‘Health’ category of Google Play store. To be able to generalise our findings, we randomly selected 59 apps across different categories and downloaded 3000 reviews written in English for each app disregarding the number of stars associated with them. The data were preprocessed along the same lines described above.

For our analysis, we, first, checked for each app the correlation between the average numbers of days separating reviews in a sample and the values of troll coefficients obtained for respective samples. Surprisingly, some of the apps revealed positive correlation meaning that their troll coefficients became higher as the average time distances became longer. As can be inferred from Fig. 2, this tendency is characteristic for the apps where reviews are very time sparse, that is, separated, on average, by more than 300 days.

Our task was to build a model capable of predicting values of troll coefficients for any app having as its input only average time

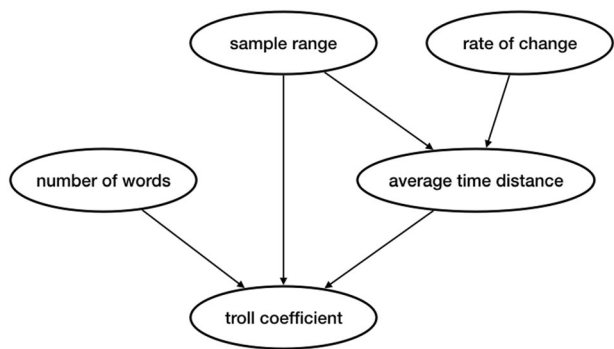


Fig. 3 Causal diagram of the app reviews model.

Table 1 Troll coefficients and average time distances correlation for training data.

App	Correlation coefficient
com.nhs.online.nhsonline	-0.74
com.teacapps.barcodeScanner	-0.43
us.zoom.videomeetings	-0.16
com.canva.editor	0.01
org.telegram.messenger	0.24
com.google.android.spotlightstories	0.37
com.borauonline.idates	0.76

distances between reviews in the samples. However, a number of potential confounders have to be controlled for. To better understand the structure of the model, we created a causal diagram (Pearl, 1995) presented in Fig. 3.

The logic behind this diagram is as follows. First, we need to control for varying average number of words in the reviews of different samples: some apps may be characterised by longer reviews than others and the greater number of words, the lower troll coefficient will be. Second, we need to take into account the range of sampling, which acts as a confounder, influencing both values of troll coefficients (here negative and positive correlation are equally likely) and average time distances (here only positive correlation is expected: the greater sampling range, the greater average time distance). Finally, average time distances are themselves the results of some data-generating process that we designated on the diagram as rate of change. Some apps are more popular than others and, therefore, are characterised by reviews posted at a higher rate. Assuming that this rate is constant for any particular app, there would be no need to control for it but for the fact that average time distance in this diagram is a collider, so our back-door adjustment set should include both sample range and rate of change (Pearl et al., 2016; Pearl, 2009).

To train the model, we chose seven out of 59 apps collected: one with a correlation coefficient near zero, three with negative correlation coefficients, and three with positive correlation coefficients, so that incremental step was approximately equal to 0.2 (Table 1).

We preferred Bayesian multiple linear regression to the classical one. Though, under a standard noninformative prior distribution, the Bayesian estimates coincide with the classical regression results, posterior simulations are useful for predictive inference and model checking (Gelman et al., 2003). The Bayesian regression model was specified as follows:

$$Q_i | \beta_0, \dots, \beta_6, \sigma \stackrel{\text{ind}}{\sim} N \left(\begin{matrix} \beta_0 + \beta_1 x_{i,\text{range}} + \beta_2 x_{i,\text{rate}} + \beta_3 x_{i,\text{days}} + \beta_4 x_{i,\text{words}} \\ + \beta_5 x_{i,\text{range}} * x_{i,\text{rate}} + \beta_6 x_{i,\text{range}} * x_{i,\text{days}}, \sigma \end{matrix} \right),$$

where $x_i = (x_{i,\text{range}}, x_{i,\text{rate}}, x_{i,\text{days}}, x_{i,\text{words}})$ is a vector of predictors

Table 2 Bayesian regression coefficients and probability intervals.

Coefficients	Probability intervals		
	Lower PI	Mean	Upper PI
Intercept	6.129e+00	6.173e+00	6.217e+00
Sample range	1.580e-04	1.873e-04	2.167e-04
Rate of change	-1.336e-01	-9.450e-02	-5.534e-02
Average time distance	-4.065e-04	-2.855e-04	-1.644e-04
Number of words	-2.047e-01	-2.010e-01	-1.972e-01
Sample range * Rate of change	-4.255e-04	-3.794e-04	-3.334e-04
Sample range * Average time distance	-1.236e-06	-1.104e-06	-9.724e-07

and σ is the standard deviation in the Normal model shared among all responses Q_i 's, that is, troll coefficients. Assuming independence, the prior density for the set of parameters $(\beta_0, \dots, \beta_6, \sigma)$ can be written as a product of the component densities:

$$\pi(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \sigma) = \pi(\beta_0) \pi(\beta_1) \pi(\beta_2) \pi(\beta_3) \pi(\beta_4) \pi(\beta_5) \pi(\beta_6) \pi(\sigma),$$

where $\beta_0 \stackrel{\text{ind}}{\sim} N(m_0, s_0), \dots, \beta_6 \stackrel{\text{ind}}{\sim} N(m_6, s_6)$, and the precision parameter $\phi = 1/\sigma^2$, the inverse of the variance, is $\Gamma(a, b)$. Since our prior information about the parameters' values is very limited, we assigned to them noninformative priors that would have little impact on the posterior. For regression parameters $(\beta_0, \dots, \beta_6)$, we chose prior mean to be equal to 0 and prior precision to be equal to $1/1e6$. For the precision ϕ , the prior values for shape and scale parameters were specified as $a = b = 0.001$ (Albert and Hu, 2019).

As for the data values, Q and X_{days} were obtained in accordance with the procedure described above for the health apps; X_{range} is the index of chronologically aligned samples; X_{words} is the mean of the number of words in all reviews of a particular app; X_{rate} was averaged for each app across 5000 draws of λ parameter from the posterior exponential distribution of time distances, with the conjugate noninformative prior $\lambda \sim \Gamma(0.001, 0.001)$. Values of X_{rate} are easy to interpret, for example, $X_{\text{rate}} = 0.02$ means that average time distance between reviews in different samples equals $1/0.02 = 50$ days.

Having specified the parameters, we used the JAGS software (Plummer, 2003) to draw MCMC samples from this multiple linear regression model. We ran three MCMC chains with an adaptation period of 2000 iterations, a burn-in period of 5000 iterations, and an additional set of 50,000 iterations to be run and collected for inference. Given the results of all standard diagnostic tests, we may be confident that our Markov chain has converged and we can treat it as a Monte Carlo sample from the posterior distribution.

Study 1: results and discussion

The obtained coefficients and the 95 % probability intervals are given in Table 2. Since none of them includes zero, all our explanatory variables are helpful in predicting troll coefficients of the app reviews.

The well-known technique for checking the fit of a model is to draw simulated values from the joint posterior predictive distribution of replicated data and compare these samples to the observed data. We computed the logs of the posterior density given range of different possible values of mean and variance and found those that maximise our function for both replicated and

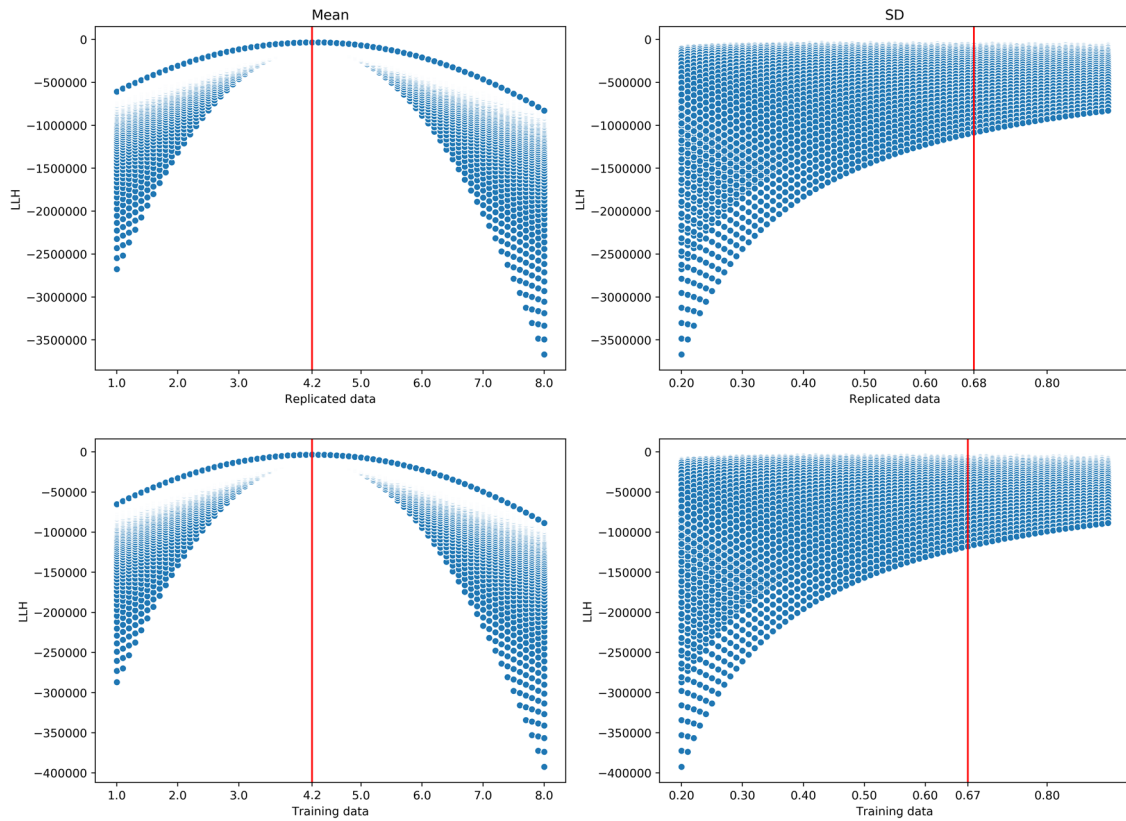


Fig. 4 Logs of the posterior density. Mean and variance for replicated and observed troll coefficients.

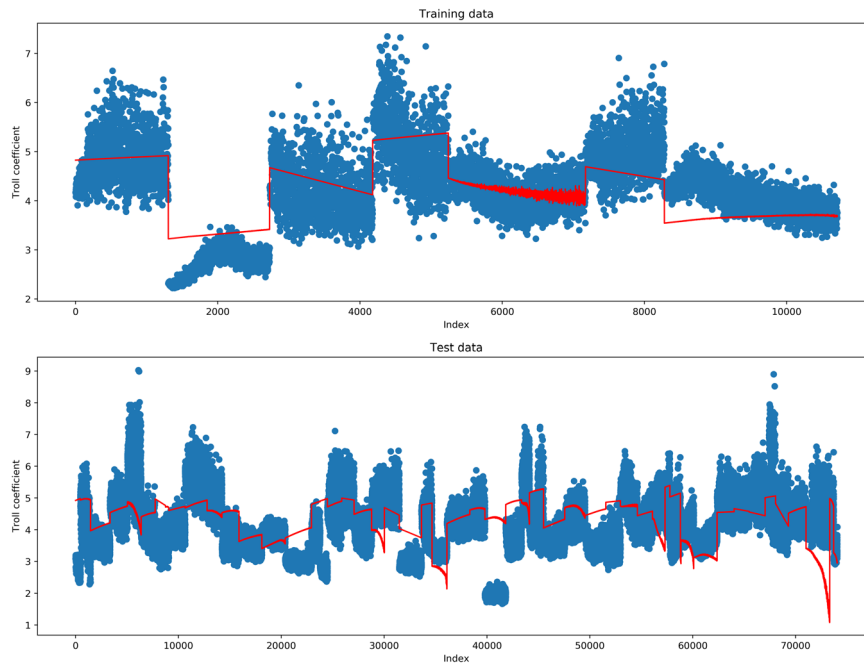


Fig. 5 Apps' troll coefficients. In-sample and out-of-sample predicting.

observed troll coefficients. Judging by the results in Fig. 4, the model seems to describe the data distribution fairly close.

In order to find out how well the model generalises, we fitted it to the remaining 52 apps reviews in our collection that were excluded from the training process. The predicted troll coefficients are plotted against the observed ones in the lower subplot

of Fig. 5 (in the upper subplot, for comparison, the same is done with data from the seven apps on which the model was trained). It looks like the model, though being an oversimplification and making a number of errors, is able to capture some relevant pattern in the distribution of troll coefficients for the majority of test apps.

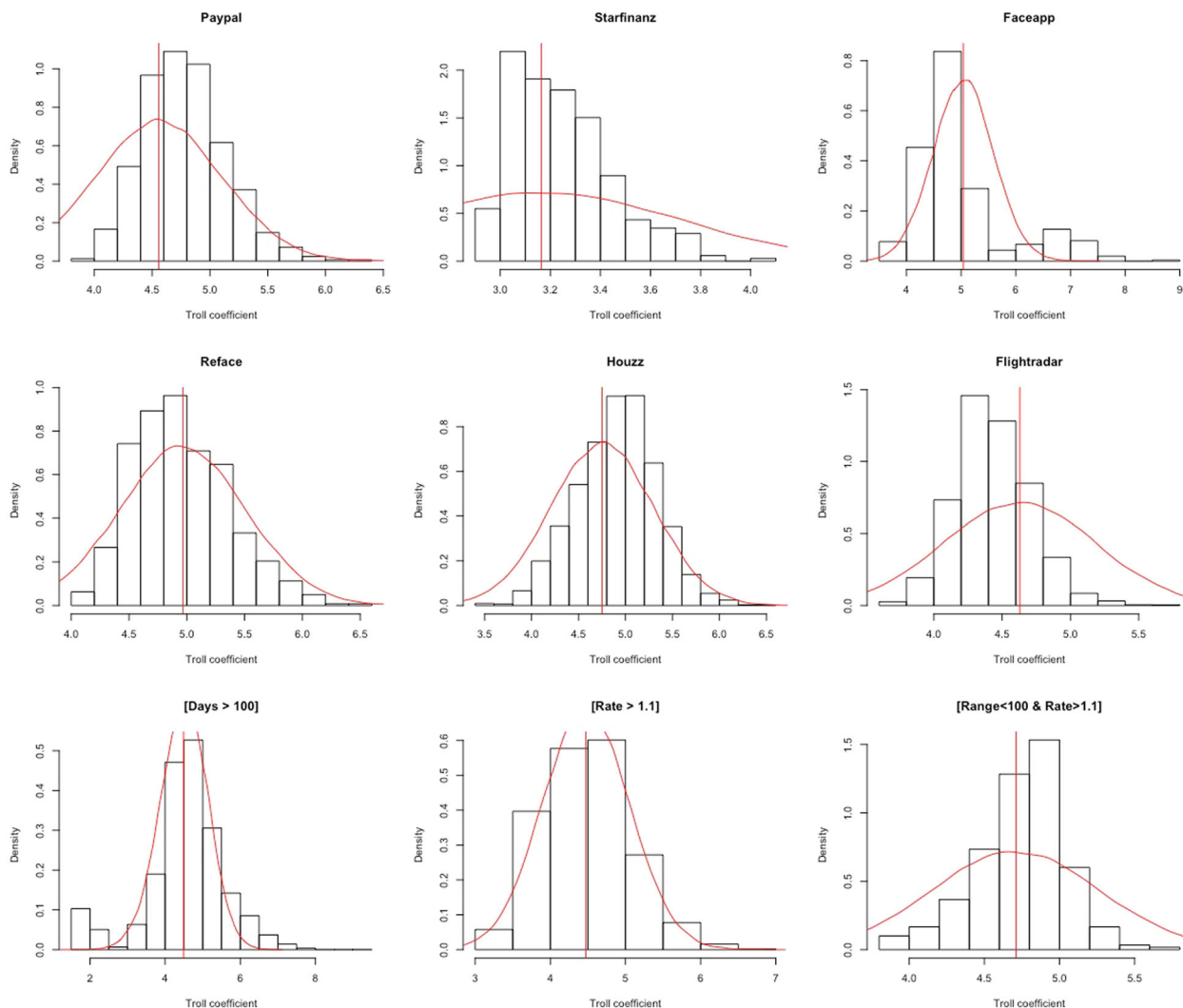


Fig. 6 Actual and replicated distributions of troll coefficients. Different apps and parameter values obtained by sampling from the posterior distribution.

More importantly, thanks to the MCMC sampling, we can create full probability distribution of troll coefficients for different apps as well as for different parameters of interest. Thus, the first two rows of subplots in Fig. 6 give us histograms of actual values of troll coefficients for six apps not included in the training set; overlaid upon them are the densities of troll coefficients for respective apps drawn from the posterior distribution. In the last row of Fig. 6, the same procedure is repeated for different values of the parameters of interest (average time distance > 100, rate of change > 1.1, sample range < 100 and rate of change > 1.1), across all apps in the dataset. Again, one can see that the posterior distribution in these cases describes the actual data reasonably well.

With regard to the regression coefficients, we observe that all of them are negative with one notable exception. When average time distance and rate of change are at their reference values, 0.5 and 0.002, respectively, the effect of sample range on troll coefficient is positive. It means that given a very slow rate of new reviews' arrival or a very small time distance between the adjacent reviews, commentators tend to gather more information about what other people have already said to adjust their own messages accordingly.

Study 2: data, methods, and results

All this time we have been talking about app reviews. To find out whether the observed tendency is truly a factor in human communication in general, it is needed to move from observational data to an experimental setting. We designed an experiment in which Russian speaking participants were asked to write a short (one or two sentences) comment describing one picture. It was a photo of a young man in a tuxedo, standing with a portable sewing machine in hands in front of a truck that has slid into a ditch. We chose it for two reasons: first, it has some kind of mystery to it and is interpretation-inducing, which allowed us to mask the true purpose of study under the pretence that we are interested in elucidation; second, it has several knots and possible hermeneutical lines (Schmitt, 2014), which allowed us to mimic the actual multiplicity of stories that is somewhat akin to the communicative situation of many customers describing their personal experiences with one and the same app.

The participants were randomly distributed between two experimental conditions: (1) in the first condition, they had to communicate their message without being able to see what anyone else had written; (2) in the second condition, they had the possibility (but no necessity) of reading what others had written

before them. The order in which participants in the second condition performed the task was random as well. Overall, there were 192 Russian native speakers in the first experimental group and 193 Russian native speakers in the second group. Each person could leave only one comment. All submissions were accepted without any censoring provided that they included no less than two content words.

To conduct the experiment, we used Yandex.Toloka (<https://toloka.yandex.com>), a Russian crowdsourcing service analogous to Amazon Mechanical Turk that facilitates collecting large volumes of data in a short time. On the platform, two special task templates were programmed to match two experimental conditions. The first template contained a link to the picture that participants were asked to describe and an input field where they were supposed to type in their comments. The second template contained a link to the specially created webpage, and two input fields. In this condition, participants were supposed to (1) go to the webpage with the picture on top of it and two-column and 250-row table below, (2) look at the picture, scroll down to the first empty row of the table, type in the first column of this row next consecutive number, type in the second column of this row their comment, 3) copy their number and comment, return back to the task page on Yandex.Toloka, paste their number and comment from the table into respective input fields.

After creating task templates, we assembled two pools of users registered on the platform who met the only criterion of being a native speaker of Russian. People were distributed among the pools randomly, so that they knew the general task but were not aware of which experimental condition they will be assigned to. The instructions for the participants of the experiment were, apart from describing the formal ways to proceed, identical and written so as not to reveal the true purpose of study. For each task, a time limit of 10 min was imposed and each following task was distributed only after the previous was closed. No user could see any tasks other than those assigned to their pool and was dismissed from the project immediately after submitting the first assignment, so that no one had the possibility to leave more than one comment. After completing the tasks, each participant was rewarded in the amount of \$1.0 USD for their submission.

We took all the necessary precautions to verify the consistency of results in the second experimental condition by comparing numbers and answers in the tables exported from the project webpage and from the Yandex.Toloka platform row by row. Everything matched, no discrepancies were detected, which suggests that the operating procedure was well-planned.

The null hypothesis H_0 of the experiment was that there would be no significant difference in the distributions of troll coefficients between two groups of tasks. The alternative hypothesis H_1 , given our prior state of knowledge, can be formulated as follows:

- $H_{1,1}$) submissions of the second group will be characterised by significantly greater values of troll coefficients than submissions of the first group;
- $H_{1,2}$) submissions of the second group, when sorted from earliest to latest, will reveal positive correlation between troll coefficients and sample indices (the more previous answers will be available for commentators, the more they will try to be linguistically creative); as for the submissions of the first group, no significant correlation between troll coefficients and sample indices will be observed;
- $H_{1,3}$) submissions of the second group, when sorted from latest to earliest, will reveal an inverted U-shape pattern of association between troll coefficients and sample indices (latest commentators will be taking into account only the answers that have been left at some reasonable distance from their ones and will disregard the earliest messages); as

for the submissions of the first group, again, no such pattern will be observed.

The obtained data were preprocessed along the same lines described above, with the only exception of the sampling window having been made 50 comments instead of 250 due to the smaller total numbers. The results of the analysis are plotted in Fig. 7.

To test our hypothesis, we fitted two first-order linear regression models to the data arranged from earliest to latest and two second-order polynomial models to the data arranged from latest to earliest. In all models, troll coefficient was the response; the only predictor was sample index (for the first-order models) or its quadratic term (for the second-order models). The coefficients and 95% confidence intervals as well as models' p -values and adjusted R^2 values are given in Table 3.

The data show that all three parts of our null hypothesis may be rejected. Troll coefficients of the comments in the first and second experimental condition, when adjusted for varying numbers of content words, differ significantly, the former are characterised by significantly lower values than the latter ($t(257.59) = -37.2$, $p < 0.0001$ for the samples aligned from earliest to latest; $t(262.88) = -46.7$, $p < 0.0001$ for the samples aligned from latest to earliest).

Comments produced in the first experimental condition, when sorted from earliest to latest, reveal no association between troll coefficients and range of sampling, while in the comments produced in the second experimental condition, given the same ordering, each increase in index predictor leads to an increase in troll coefficient response.

For the comments produced in the first experimental condition, reversing of order does not result in any change of a pattern of (no) association; however, the comments produced in the second experimental condition, when sorted from latest to earliest, reveal the anticipated inverted U-shape pattern of association between troll coefficients and sample indices.

General discussion

The results presented in Studies 1 and 2 suggest that people, when being able to make themselves acquainted with what other people have written before them on the same topic, are willing to take into account not only the communicated information but also the choice of words. Surprisingly, in the light of what we know about cognitive priming effects, this prior knowledge forces interlocutors to refrain from verbatim repetitions and explore language space in the search of new lexemes and constructions that they will be the first to introduce. However, people only do so in case they perceive the flow of conversation as uninterrupted and discussion as ongoing.

This observation is even more interesting if we contemplate the fact that it applies to a very special communicative situation in which mobile app reviews, especially negative ones, are produced. People who write such reviews are not forced to take previous messages into account (plagiarise or paraphrase them); in fact, they do not even have to read them. On the other hand, there is no such thing as public appraisal of their eloquence and wit. That is why the repatterning that actually takes place can only be accounted for by the commenters' desire to make the language of their own contributions in some respect different from that of the contributions of others.

The question now is what cognitive mechanism underlies this process of repatterning? Since this paper is not a study in psycholinguistics, we are not in a position to draw reliable inferences about the interlocutors' real motives. Nevertheless, without falling into the sin of psychologising textual data, we can base our hypothesis on some intuitively clear premises, by a simple process

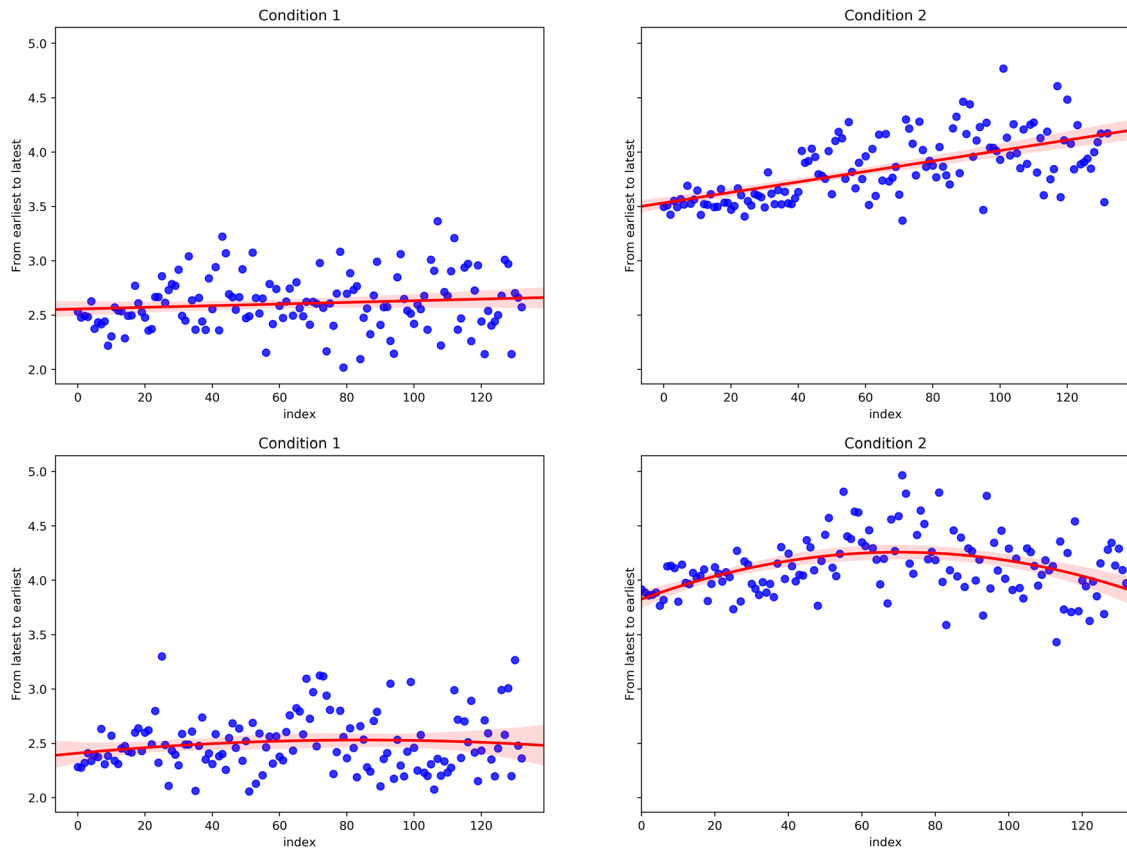


Fig. 7 Correlation of troll coefficients with sample indices. Two experimental conditions with comments sorted from earliest to latest and form latest to earliest.

Table 3 Coefficients of the first-order and second-order models and their confidence intervals.

Models	Confidence intervals			p-values	R ²
	Lower CI	Estimate	Upper CI		
M1.1	-0.0003	0.0007	0.001	0.17	0.006
M2.1	0.003	0.004	0.005	<0.0001	0.39
M1.2	-0.72	-0.15	0.42	0.26	0.005
M2.2	-1.82	-1.33	-0.84	<0.0001	0.18

Note. M1.1—first-order model for the first experimental condition, M2.1—first-order model for the second experimental condition, M1.2—second-order model for the first experimental condition, M2.2—second-order model for the second experimental condition.

of elimination. First, since the repatterning is time-dependent, it must be a discourse-level phenomenon rather than anything else. Second, since there are no discernible pragmatic reasons, practical considerations for the interlocutors in this particular communicative situation to avoid being repetitive (which is especially clear in the experimental setting where nothing prevented the participants from copy-pasting any earlier-made comment), it must be the case that some recency effect opposed to priming has a significant impact on people’s choice of linguistic means.

One cannot but notice that the workings of this recency effect, as observed in the current paper, are reminiscent of the notions of ‘schema refreshing’, ‘heteroglossia’ (Blackledge and Creese, 2014; Tagg, 2013; Bailey, 2007; Cook, 1994), and ‘people’s strategic deployment of language resources from across their repertoire’ (Maybin, 2016: p. 36), all of which have come from the field of language creativity studies. While everyday language creativity has been an area of extensive research for decades (Cameron,

2011; Semino, 2011; Mendoza-Denton, 2008; Holmes, 2007; Maybin and Swann, 2007; Tannen, 2007[1989]; Carter, 2004; Cook, 2000; Norrick, 2000; Crystal, 1998; Chomsky, 1982), there is definitely a lack of agreement about the precise definition and scope of creativity itself. Still three almost universally present ideas can be identified: (1) an idea of a reflexive manipulation of language structure as the language creativity’s form of existence, (2) an idea of dialogicality as a necessary precondition for the arising of language creativity, and (3) an idea of participants’ enjoyment of language play as a tentative explanation of the phenomenon of language creativity.

It is easy to see in which respects our findings do not quite fit this well-established matrix. First, the notion of a reflexive manipulation of language form should be made more specific and more contextual. We can represent it, for any message Y_{i+1} , as a function $\text{argmax}_{Y_{i+1}} (Y_{i+1} - Z - \{Y_1 \dots Y_i\})$ over the number of words in the message Y_{i+1} that are neither in some hypothetical set of topical words Z , necessary to maintain the cohesiveness of discussion, nor in the set of words introduced by the messages $\{Y_1 \dots Y_i\}$ where $\{1 \dots i\}$ delineates timeframe that is relevant for the message Y_{i+1} .

Second, the notion of dialogicality also should be widened to include the communicative situations like online commenting. This type of discourse was shown to reveal multiple cohesive analogies to written monologues while also exhibiting some features of prototypical spoken dialogues (Hoffmann, 2010). Thus, it definitely occupies some intermediate position on the cline between monologic and dialogic interactions. If we adapt for our purposes a terminological distinction made by Cowan and Arsenault (2008), this type of discourse may be called *collaboration*. Cowan and Arsenault used this term to refer to initiatives that feature an effort by citizens of different countries to complete a common project or achieve a common goal. In the

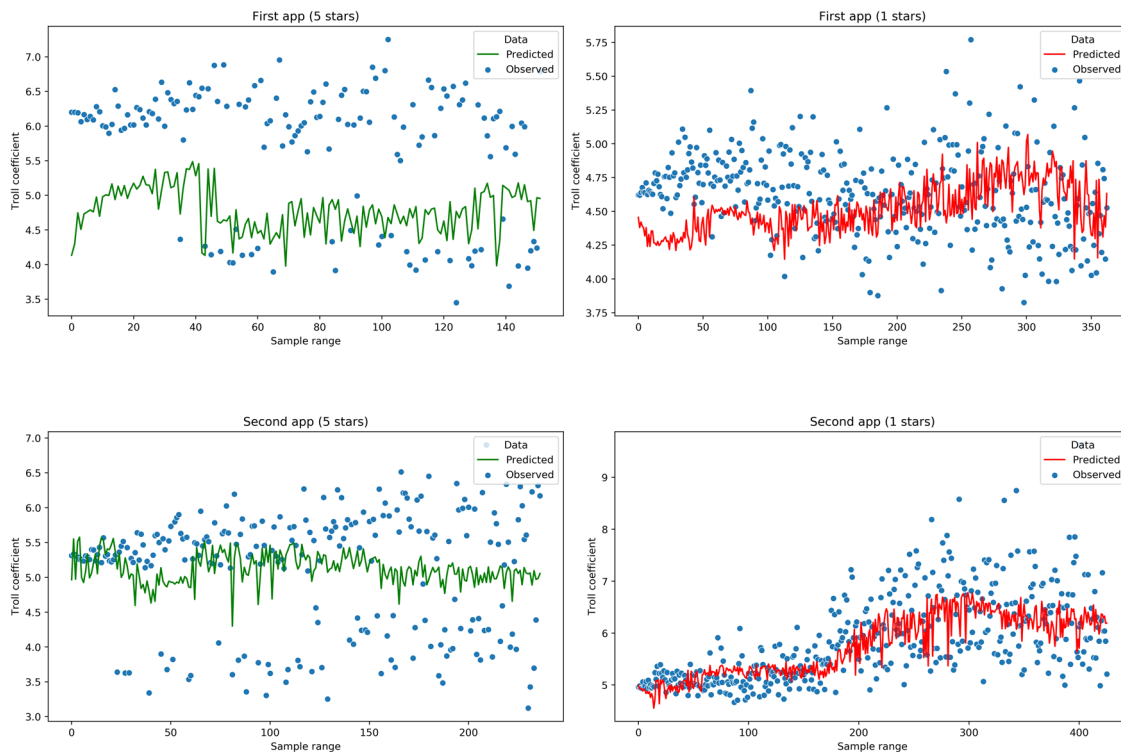


Fig. 8 Observed and predicted values of troll coefficients. Five- and one-star reviews of two suspicious apps.

domain of language, it can be understood as an effort to exhaust all possible linguistic means that can be sensibly used while discussing a topic, in essence, a ‘third-type phenomenon’ (Keller, 1994) resulting from conditionally dependent efforts of many speech-act participants.

Pursuing this line of thought, we might state that the notions of schema refreshing and manipulation of language form are, on the face of it, easily coupled with the assumed troll strategy of writing. If so, two completely different discourse scenarios—troll writing and collaborative repatterning—happen to produce the same result. With troll writing, a person uses different contexts as simple proxies to get through a limited number of signals. With app reviews and other forms of online commenting, a person adjusts their message with regard to what other people have already said within some timeframe that he or she considers relevant. The reason for this adjustment, as one may contend, is the unwillingness to be repetitive and the unsolicited desire to use words that have not yet been used while contributing to the same topic.

Conclusion

This paper reports the results of two studies and comprises both observational and experimental data. For observational part (Study 1), we analysed more than 180,000 app reviews written in English and Russian. For experimental part (Study 2), participants were asked to describe the same picture in two experimental conditions. In the first one, they had to communicate their message without being able to see what anyone else had written. In the second one, they had the possibility (but no necessity) of reading what others had written before them.

The observational and experimental results match well. In essence, it was found that troll coefficients obtained for different groups of app reviews and online comments could be modelled as a complex function of time distance between separate individual contributions. Though approximating this function by linear regression brings tolerable results, its true nature is probably

more sophisticated, with several variables involved and interacting dynamically. The data that we have suggest that people are more likely to engage in this game of deployment of language resources when the communicative space is conceptualised as continuous, which presupposes a very small time distance between the adjacent contributions or a very slow rate of new contributions’ arrival.

On the one hand, our theoretical and methodological framework shows that the same algorithm that has proven highly efficient in detecting internet trolls can give us a reasonable estimate of how much previous information online commentators are willing to take into account. It also allows to predict under which time and rate of arrival conditions people will shift towards either pole of the monologic-dialogic continuum. Insights provided by this framework can help better understand the nature of online communication and foretell the possible communicative scenarios of its participants.

On the other hand, these findings definitely render the task of developing efficient linguistic algorithms for internet troll detection more complicated. In the process of devising such algorithms, it should now be taken into account that a certain troll-like effect may arise as a result of collaborative repatterning that is not indicative of any malevolent practices in online communication. However, the problem can be alleviated by our ability to predict what the value of the troll coefficient of a certain group of texts would be if it depended solely on these texts’ creation time.

For example, consider the plots in Fig. 8 where the observed and predicted values of troll coefficients for five- (left column) and one-star (right column) reviews of two anonymised apps from our data are visualised. Remarkably, while troll coefficients for one-star reviews form uninterrupted sequences that are reasonably well approximated by our model, with five-star reviews, troll coefficients break up into two groups, of which only one is close to the predicted developmental trend.

This observation may give rise to reasonable suspicions. One may speculate that some of the 5-star reviews of these apps were

written by specially hired people in order to increase the apps' average ratings and attract more potential users. If so, we can hypothesise that two different scenarios of producing fake reviews were selected: in one case, creative rewriting that resulted in higher than expected values of troll coefficients (the upper left plot in Fig. 8); in the other case, simple reposting that resulted in lower than expected values of troll coefficients (the lower left plot in Fig. 8). Needless to say, these considerations are only preliminary and require further testing and elaboration.

Data availability

The datasets analysed during the current study are available in the Zenodo repository: <https://zenodo.org/record/4295546#.X8LSWS3Mzq0>.

Received: 7 December 2020; Accepted: 2 November 2021;

Published online: 18 November 2021

References

- Albert J, Hu J (2019) Probability and Bayesian modeling. CRC Press
- Bailey B (2007) Heteroglossia and boundaries. In: Heller M (ed.) Bilingualism: a social approach. Palgrave, pp. 257–276
- Blackledge A, Creese A (2014) Heteroglossia as practice and pedagogy. In: Blackledge A, Creese A (eds) Heteroglossia as Practice and Pedagogy. Springer, pp. 1–20
- Cameron LJ (2011) Metaphor and reconciliation: the discourse dynamics of empathy in post-conflict conversations. Routledge
- Carney T (2014) Being (im)polite: a forensic linguistic approach to interpreting a hate speech case. *Lang Matters* 45(3):325–341
- Carter R (2004) Language and creativity: the art of common talk. Routledge
- Chakraborti N (2010) Hate crime: concepts, policy, future directions. Willan
- Chomsky N (1982) A note on the creative aspect of language use. *Philos Rev* 91(3):423–434
- Cook G (2000) Language play, language learning. Oxford University Press
- Cook G (1994) Discourse and literature. Oxford University Press
- Cowan G, Arsenault A (2008) Moving from monologue to dialogue to collaboration: the three layers of public diplomacy. *Ann Am Acad Polit Soc Sci* 616:10–30
- Crystal D (1998) Language play. Penguin
- Douglas KM, McGarty C (2001) Identifiability and self-presentation: computer-mediated communication and intergroup interaction. *Br J Soc Psychol* 40(3):399–416
- Egele M, Stringhini G, Kruegel C et al. (2017) Towards detecting compromised accounts on social networks. *IEEE Trans Depend Secure Comput* 14(4):447–460
- Elyashar A, Bendahan J, Puzis R (2018) Is the online discussion manipulated? Quantifying the online discussion authenticity within online social media. Preprint at <https://arxiv.org/abs/1708.02763>
- Fraser B (1998) Threatening revisited. *Forensic Linguist* 5(2):159–73
- Gelman A, Carlin J, Stern H et al (2003) Bayesian data analysis. Chapman and Hall
- Herring SC, Job-Sluder K, Scheckler R et al. (2002) Searching for safety online: managing 'trolling' in a feminist forum. *Inf Soc* 18:371–384
- Hoffmann CHR (2010) From monologue to dialogue? Cohesive interaction in personal weblogs. Dissertation, University of Augsburg
- Holmes J (2007) Making humour work: creativity on the job. *Appl Linguist* 28(4):518–537
- Keller R (1994) On language change: the invisible hand in language. Taylor & Francis
- Lundberg J, Laitinen M (2020) Twitter trolls: a linguistic profile of anti-democratic discourse. *Language Sciences* 79. <https://doi.org/10.1016/j.langsci.2019.101268>
- Maybin J (2016) Everyday language creativity. In: Jones RH (ed.) The Routledge handbook of language and creativity. Routledge, pp. 25–39
- Maybin J, Swann J (2007) Everyday creativity in language: Textuality, contextuality and critique. *Appl Linguist* 28(4):497–517
- Mendoza-Denton N (2008) Homegirls: language and cultural practice among latina youth gangs. Wiley-Blackwell
- Monakhov S (2020) (2020a) Understanding troll writing as a linguistic phenomenon. In: Arai K, Kapoor S, Bhatia R (eds) Intelligent systems and applications. *IntelliSys 2020. Advances in intelligent systems and computing*, vol 1251. Springer, Cham
- Monakhov S (2020b) Early detection of internet trolls: Introducing an algorithm based on word pairs/single words multiple repetition ratio *PLoS ONE* 15(8). <https://doi.org/10.1371/journal.pone.0236832>
- Norrick NR (2000) Conversational narrative John Benjamins, Amsterdam
- Pearl J, Glymour M, Jewell NP (2016) Causal inference in statistics: a primer. Wiley
- Pearl J (2009) Causality: models, reasoning, and inference. Cambridge University Press
- Pearl J (1995) Causal diagrams for empirical research. *Biometrika* 82(4):669–710
- Plummer M (2003) JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. <http://citeseer.ist.psu.edu/plummer03jags.html>
- Schmitt A (2014) Knots, story lines, and hermeneutical lines: a case study. *Story-worlds: J Narrative Stud* 6(2):75–91
- Semino E (2011) Metaphor, creativity and the experience of pain across genres. In: Swann J, Pope R, Carter R (ed) Creativity, Language, Literature: The State of the Art. Palgrave Macmillan, pp. 83–102
- Sia CL, Tan BCY, Wei KK (2002) Group polarization and computer-mediated communication: effects of communication cues, social presence, and anonymity. *Inform Syst Res* 13(1):70–90
- Siegel J, Dubrovsky VJ, Kiesler S et al. (1986) Group processes in computer-mediated communication. *Organiz Behav Human Decis Process* 37(2):157–187
- Tagg C (2013) Scraping the barrel with a shower of social misfits: Everyday creativity in text messaging. *Appl Linguist* 34(4):480–500
- Tannen D (2007[1989]) Talking voices: repetition, dialogue and imagery in conversational discourse. Cambridge University Press
- Volkova S, Bell E (2016) Account deletion prediction on RuNet: A case study of suspicious Twitter accounts active during the Russian-Ukrainian crisis. In: Proceedings of NAACL-HLT. Association for Computational Linguistics, San Diego, pp. 1–6.
- Zannettou S, Caulfield T, De Cristofaro E et al. (2019) Disinformation warfare: understanding state-sponsored trolls on Twitter and their influence on the web. Preprint at <https://arxiv.org/abs/1801.09288>

Acknowledgements

None

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The author declares no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Correspondence and requests for materials should be addressed to Sergei Monakhov.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021