Methodology for measuring polarization of political discourse: case of comparing oppositional and patriotic discourse in online social networks

Tamara Shcheglova, Galina Gradoselskaya and Ilia Karpov

National Research University Higher School of Economics, 101000 Moscow, Russian Federation, e-mail: tshcheglova@hse.ru

Abstract The paper analyzes speech markers and semantic concepts typical for patriotic and oppositional discourse in social networks. About 100 000 posts from Facebook, VKontakte, and LiveJournal were analyzed, and 35 000 most frequent speech markers were processed, of which 1800 markers were selected for analysis. The alternative method to tf-idf metric for specific text markers identification is proposed. The features of oppositional discourse in comparison with the patriotic discourse were formulated. On the one hand, the analysis of sets of speech markers that characterize political groups allows us to understand social models and attitudes embedded in the discourse and the subsequent behavior of representatives of these groups. On the other hand, it is possible to extend a set of keywords for text search of a certain political orientation, based on the obtained results.

Introduction

The Internet space and social media has a dual nature. On the one hand, it is as a structural formation, where actors (persons, groups, pages, etc.) are connected by information flows and social ties. On the other hand, information flows form a kind of a general discursive space, where speech markers merge into higher-level concepts. Speech markers and concepts have a significant semantic load. They have a social function by implementing models of social influence and manipulation like "us-them" model [Shipilovб 2003], and ideological function by demonstrating values, projecting models of the future, etc.

Therefore, another view on the space of social networks is possible - as a constructed space of meanings, which is a generalized reflection of the discourse of social groups that influences real socio-political processes. A similar view on the role of discourse and communicative space is presented in the works of Yu.M. Lotman [Lotman, 2010], L.B. Makeeva [Makeeva, 2011], T. Tsyvyan [Tsivyan, 2009].

We should also mention the founders of this approach – the classics of the Geneva linguistic school: F. Sossur [Saussure, 1977] and S. Bally [Bally, 2003].

Our study will show key speech markers and semantic concepts that characterize the discursive space of opposing groups in politics on social networks. One group represents a pro-government position, which in the modern political field is defined as "patriotic". Another group represents an oppositional position.

It is necessary to emphasize that the terms "oppositionists" and "patriots" in our study are conventionally accepted, by the principle of self-identification by representatives of these groups, and by their labeling of groups of opponents. That is such identification is also a derivative of "collective intelligence" in a discursive Internet space.

The aim of the study is to determine the speech markers and concepts peculiar for the identified political groups – the bearers of certain political attitudes (in our study – "patriotic" and "oppositional").

The results of the research can have both scientific-methodological and applied significance. To identify speech markers that characterize the discourse of political groups, we developed a special method. The analysis of sets of speech markers that characterize political groups is independent value, it allows us to understand social models and attitudes embedded in the discourse and the subsequent behavior of representatives of these groups. Applied value of the research is that "reverse search technology" is possible – texts search of a certain political orientation or designed social processes (for example, protest actions, strikes, pickets, etc.) according to established specific speech markers.

Solving the problem of comparing the speech behavior of two different political groups required development of special tools. A familiar linguistic tool for most significant words search showed uninterpreted results on our corpus of texts. Perhaps the peculiarity of the object of our research – two large politically opposed corpora of texts – was not taken into account. Therefore, we had to develop our own method of texts polarization and highlight keywords that mark this differentiation – discourse differentiation index.

Literature review of discourse research and allocating key speech markers

Theoretical background for the study of discourse

In critical discourse analysis (CDA) discourse is much more than a sequence of linguistic signs and symbols. Even more, discourse is a multidimensional substance, that includes texts itself, discursive practice, and sociocultural practice. It is a text

as it is the product of language. It is discursive practice because it is associated with established type of discourse connected to particular kind of activity. And it is sociocultural practice so as it explains the relationship between discursive and social processes [Fairclough, 1995].

T. van Dijk pays special attention to the functioning of the language in the mass media. Van Dijk examines the impact of socio-cultural factors on the mechanism of language use. An important component of the general theory of communicative-linguistic interaction, according to van Dijk, is the cognitive theory of language use, which not only give access to the processes and structures that provide cognitive processing of sentences and statements, but also explains how planning, production and understanding of speech is happening [Dijk van, 1985].

Van Dijk adopts the idea of presenting positive-self and negative other by using specific speech markers in discourse. He studies the strategies of foregrounding positive practices of oneself and de-emphasizing any positive aspect of the other [Dijk van, 1988].

Studies of political language features

Politics is a struggle for power in order to achieve certain political, economic and social goals. The analysis of political discourse should treat discourse as an instrument of doing politics. In this context language plays a significant role since every political action is born, prepared, controlled, influenced, and performed by language [Horvath, 2009]. Internet and social media have dramatically changed the study of political communication as researches access massive feeds of data on online social media behavior, networks and language [Gonawela, 2017].

In one research author investigated ideological structures of polarized discourse coded in the reports of two online news websites: egyptindependent and Ikhwanweb. Author found out features of the ideologies of polarized discourse and concluded with a discussion on how both websites establish a dichotomy of “we” vs. “them” [Eissa, 2014].

In the other research Obama’s political discourse is investigated. The authors oppose liberal discourse to conservative and highlight its main features. Concepts of *freedom* and *justice* constitutes liberal discourse in US. *Freedom* is defined as social and political rights of individuals that protect them from interference by others in their lives. Justice is understood in terms of equal rights and the end of oppression in social world [Horvath, 2010].

Linguistic features of measuring distinctive words

It is common practice in computational linguistics to model documents by the words that have been weighted by their term frequency-inverse document term (TF-IDF). It has been the most commonly adopted document representation method for various text-processing tasks. It provides a weight to each word in a document according to the frequency of its occurrence in text and the rareness of its use in the other documents un the corpus of texts. TF-IDF metric works on the basis of bag of words, which involves in the assumption that the document is simply a collection of words and a vector can be computed by estimating the relative distance between words [Kim et al., 2018]. TF-IDF reliably captures what is distinctive about a particular document and it could be interpreted as a feature evaluation technique. According to the logic of this approach most distinctive words are the ones spoken by one party and not spoken once by the other [Monroe et al., 2008].

The problem with that metric is that it allows to pick out distinctive but not widely used words. It is also should be noted that the standard linguistic approach ignores nonstandard words use, considering them marginal and erroneous, while the cognitive approach allows interpreting non-standard uses as specific operations on knowledge. Thus, it becomes possible to detect the hidden intentions of the speaker [Issers, 2008].

Method for differentiating speech markers

Data description

To understand the quantitative trends in the discourse of patriots and oppositionists, the corresponding publications in social networks were investigated. The study was conducted in October – November 2015. At the first stage, 230 patriotic and 240 oppositional resources were expertly selected in three social networks: Facebook, VKontakte and LiveJournal.

The resources were groups and open pages, from which about 100 000 posts were downloaded (over the previous six months). Based on the downloaded frequencies of texts (speech markers) for each political group (patriots and oppositionists) were counted, as well as the total frequency for all groups. 35 000 most frequent speech markers were processed: all indexes, marking the peculiarities of the discourses of patriots and oppositionists were calculated. The final basis for analysis, containing socially significant and differentiating discourses of patriots and oppositionists, was

over 1800 words. Traditionally, the metric TF-IDF is used to identify specific text markers or subsamples [Manning et al., 2008]. However, according to the data obtained in the study, this metric showed not very adequate, rare words (with a frequency of about 5-10 words in the entire array). Perhaps this result would be acceptable for linguists, but for our purposes (further use of markers for targeted material search and classification of texts) such a result will not be relevant. Therefore, to determine specific speech markers we developed an alternative method for differentiating speech markers.

Characteristics of frequency distributions

In the database uploaded from the primary data of social networks, there is a list of speech markers that are used in posts and comments of patriotic and oppositional resources. Speech markers were counted after the normalization of texts (putting the words in the nominative case of the singular).

For each speech marker, the baseline data was calculated as the initial data, which were subsequently used to calculate differentiation indices:

- total frequency of use (in all texts);
- frequencies for each group of texts (patriotic and oppositional);
- percentage of occurrence of a given speech marker in the entire discourse (relative frequency);
- relative frequencies of use for each group of texts (normalization is carried out by dividing the frequency of the speech marker by the total volume of the discourse in this group of texts).

Methodological issues

After weighting the words with TF-IDF metric large absolute frequencies and small relative frequencies of speech markers did not allow to objectively compare the prevalence of speech markers in a particular discourse. Therefore, there was a need for developing an indicator (or a system of indicators) that shows the predominance of the speech marker in a particular discourse, as well as a general indicator that allows to identify key speech markers polarizing the discourse of groups of two different political orientations.

Here we propose the system of indicators that measure the difference between word use in two discourses. That system consists of two basic indices and one final index (as a combination of two basic ones): *PO Index*, *OP Index* and *Total Index*. $WF_{patriotic}$ is the number of occurrences of the speech marker in the patriotic

discourse and $WF_{oppositional}$ is the number of occurrences of the speech marker in the oppositional discourse:

- *Index of prevalence of patriotic discourse over the oppositional (PO Index)* is calculated as the ratio of the relative frequency of the occurrence of the speech marker in the oppositional discourse to the patriotic discourse. This index shows how many times the word prevails in the oppositional discourse in relation to the patriotic:

$$PO\ Index = \frac{WF_{patriotic}}{WF_{oppositional}} \qquad (1)$$

- *Index of prevalence of oppositional discourse over patriotic (OP Index)* is calculated as the ratio of the relative frequency of the occurrence of the speech marker in the patriotic discourse to the oppositional one. It shows how many times this word prevails in patriotic discourse in relation to the opposition:

$$OP\ Index = \frac{WF_{oppositional}}{WF_{patriotic}} \qquad (2)$$

- *Index of differentiation of speech markers between discourses (Total Index)* is calculated as the square root of the difference between the *PO Index* and the *OP Index*. It shows the degree of discrepancy between the usage of the given word in different discourses - patriotic and oppositional. Usually, the size of the index is 1 or more (which corresponds to the predominance of the word in some discourse more than twice):

$$Total\ Index = \sqrt{(PO\ Index - OP\ Index)^2} \qquad (3)$$

Practical results of the study

Key speech markers common to all political groups (speech markers ranked in descending order of the total absolute frequency) are shown in Table 1. The importance of key geopolitical concepts for patriots and oppositionists coincides.

Speech marker	Total frequency	OP Index	PO Index	Total Index
Russia	478 924	1.025	0.976	0.048
them	458 490	1.018	0.982	0.036

us	397 217	1.065	0.939	0.125
country	234 049	1.148	0.871	0.277
world	130 245	0.877	1.140	0.263
state	92 923	1.163	0.860	0.302
people	85 771	1.228	0.815	0.413
RF (Russian Federation)	84 449	1.142	0.875	0.267
history	69 487	0.923	1.083	0.160
politics	63 348	1.093	0.915	0.178
Crimea	60 819	0.811	1.233	0.421
West	57 781	1.187	0.843	0.344
worldwide	57 675	0.933	1.072	0.139
victory	51 499	0.772	1.295	0.523
government	51 119	1.191	0.840	0.351
western	50 814	0.779	1.284	0.504
USSR	49 706	0.804	1.244	0.441
international	49 678	1.076	0.929	0.146
population	46 188	1.062	0.941	0.121
European	42 081	0.912	1.097	0.185
national	40 162	0.976	1.025	0.049

Table 1. Key speech markers common to all political groups.

In all discourses there are the country names - in the present and past tense: “Russia”, “RF” (Russian Federation), “country”, “state”, “USSR”. Equally significant are the references to the people of the country: “people”, “population”, “national”. Also, there are references to major international actors: “West”, “European”, “International”.

The name of the Crimea peninsula after the events of 2014 can be designated as situational speech markers. This event became significant in both discourses – oppositional and patriotic.

It is significant that in the second and third places there are speech markers “us” and “them”, which indicates the prevalence of the model of socio-political differentiation “us-them” [Shipilov, 2003] in the discourses of all political groups.

Key speech markers of patriotic discourse (the speech markers are first selected from the most frequent words, and then ranked in descending order of the PO Index – the predominance of patriotic discourse over the oppositional discourse) are shown in Table 2.

Speech marker	Total frequency	OP Index	PO Index	Total Index
battle	64 366	0.099	10.117	10.018
Donetsk	37 140	0.119	8.409	8.290
tank	42 179	0.149	6.691	6.542

combat	67 959	0.162	6.166	6.003
Poroshenko	42 909	0.169	5.913	5.744
DPR (Donetsk People's Republic)	55 316	0.173	5.771	5.598
Novorossia	41 806	0.189	5.299	5.110
fire	40 994	0.200	5.010	4.810
enemy	42 848	0.202	4.952	4.750
rocket	33 553	0.223	4.481	4.258
troops	74 335	0.246	4.072	3.826
army	93 353	0.253	3.951	3.698
defense	42 233	0.301	3.327	3.027
hero	38 318	0.308	3.250	2.943
Donbass	72 301	0.309	3.236	2.927
front	37 562	0.319	3.131	2.812
military	142 767	0.438	2.283	1.845
Ukraine	295 467	0.482	2.076	1.594
American	79 282	0.498	2.007	1.509
force	111 356	0.586	1.707	1.121

Table 2. Key speech markers of patriots.

In the patriotic discourse, the first place takes situational lexicon which is associated with the events at the time of the research (autumn 2015) taking place in the south-east of Ukraine: “Donetsk”, “DPR”, “Novorossia”, “Donbass”. Also, ideological opponents of different levels are recalled: “Poroshenko”, “American”. Military terms predominate: “battle”, “combat”, “tank”, “military”, “front”, “defense”, “troops”, “army”, etc.

The features of patriotic discourse can be formulated as follows:

- Discourse is directed to the past: the history of the Soviet Union, its achievements, victories are recalled;
- Discourse is militarized: the names of weapons and military terms prevail;
- In the patriotic discourse situational speech markers devoted to current events in Ukraine and Donetsk.

Key speech markers of oppositional discourse (speech markers are first selected from the most frequent words, and then ranked in descending order of the OP Index (the predominance of oppositional discourse over patriotic) are shown in Table 3.

Speech marker	Total frequency	OP Index	PO Index	Total Index
court	34 822	4.120	0.243	3.877
oil	25 145	3.164	0.316	2.848
Kremlin	21 328	3.094	0.323	2.771
ruble	47 952	3.040	0.329	2.711

society	35 268	2.770	0.361	2.409
price	41 361	2.756	0.363	2.393
elections	30 797	2.743	0.365	2.378
network	23 305	2.504	0.399	2.105
action	25 981	2.346	0.426	1.920
freedom	28 703	2.330	0.429	1.901
law	44 680	2.295	0.436	1.859
civil	31 019	2.283	0.438	1.845
bank	26 569	2.255	0.444	1.811
Putin	135 541	2.083	0.480	1.603
company (firm)	40 783	2.082	0.480	1.602
social	26 581	2.043	0.490	1.553
crisis	28 091	1.972	0.507	1.466
sanction	42 667	1.762	0.567	1.195
power	109 463	1.688	0.592	1.096
article	33 171	1.634	0.612	1.022

Table 3. Key speech markers of oppositionists.

In addition to the high-frequency words reflected in Table 3, that characterize the oppositional discourse, there are less frequent, but very popular terms (more than 300 in the subsample) that can also be grouped in meaning. A large semantic group of speech markers that characterize power in Russia (for example, “Kremlin”, “federalism”, “clamp”, “kleptocracy”, etc.) and the head of country (“VVP” (Vladimir Vladimirovich Putin), “putler”, etc.). Separately the supporters of power are characterized by “Kremlebot”, “troll”, “Edinaya Rossia” etc. And the information space of the country (“zombiebox”, “propaganda”, “pro-Kremlin”, “hurray-patriotism”, etc.). Specific names of state corporations, names of state officials, names of regions of the country that are in the zone of attention of the opposition are also listed. Also, there are specific persons of influence, resources of influence. The directions of the opposition's actions are listed, and as usual the protest action (“appeal”, “petition”, “action”, “picketing”, “hunger strike”, “rally”, “procession”, “unauthorized”, etc.) and active action (“terror”, “violence”, “revolutionary”, “lustration”, “anarchist”, “ultra-right”, “bolotnyi” (after protest events in May 2012 on the Bolotnaya square), etc.).

The opposition resources are actively discussing the activities of large state corporations. The names of companies are often mentioned: “Rosbank”, “Gazprom-Media”, “Lukoil”, “Rosneft”, “VTB”, “RZD”. The context of the discussion concerns situations that are possible carriers of corrupt practices: government contracts, government procurement.

Representatives of oppositional discourse actively link to resources – significant and respected for them sources of information: “SvobodaNews”, “Libernews”, “Opir”, “Rabkor”, “Open Russia”, “Forbes”, “TVrain”, “Grani”, “Snob”, “Slon”, “Novaya Gazeta”, “Echo”, “Obozrevatel”, “Vedomosti”, “Rosbalt”, “Transparency

International", "Inosmi", "Kommersant", "Meduza", "Euronews", "Interfax". At the same time, there are no pro-government sources in the list.

Among the representatives of the oppositional discourse, a clear self-identification is built: political prisoner, dissident, dissenter. There are also target groups that are carriers of opposition views: Democrat, political prisoner, dissent, dissent, youth, student, intelligence. The social positioning of the opposition is accompanied by emotional speech marks: self-defense, protest, hybrid, anti-Putin. The actions of the authorities in relation to the opposition are characterized in such a way as to justify their own protest actions. They are characterized by the following negative markers: arbitrariness, prohibit, dispersal, discrimination, redistribution, illegal, police, censorship.

The features of oppositional discourse in comparison with the patriotic discourse can be formulated as follows:

- The discourse is more specific than the discourse of patriots – the names of modern politicians, ministries and state corporations are much more common in use.
- Most modern economic terms predominate - the discourse of oppositionists claims a monopoly of scientific character and objectivity.
- Economic evaluation of the country's future is depressing: "sanctions", "oil", "crisis", etc.
- Legal terms prevail as a guarantee of the legal basis of political activity. The discourse of the oppositionists represents both legal terminology and prison slang ("pakhan" (crime boss), "zek" (convict), "skhod" (descendant of thieves), etc.).
- The terms that became the ideological norm in the 90s are fully used: "society", "public", "civil", "freedom", etc.
- Social technologies of manipulation are mentioned: "action", "picket", etc.

We can draw a general conclusion that oppositional discourse is the result of careful sociolinguistic reflection and social design. Patriotic discourse is formed spontaneously, there is no ideological basis and organizational component of work with patriotically minded groups of the population. In general, patriotic discourse loses to the opposition on ideological and methodological grounds.

Conclusion

According to the results of the study, conclusions can be drawn in two directions: informative and methodological.

Informative conclusions

Polarization of discourse is observed in the Russian political information space of social networks. There are two political groups opposing each other (the names are given according to their self-determination): “patriots” supporting the actions of the authorities, and “oppositionists” challenging the activities of the authorities. There are practically no overlapping and common topics for discussion between them; they live in alternative, parallel social reality.

A comparison of these discourses allows us to say that the “patriotic” discourse is extremely poor in comparison with the “opposition” discourse, it is socially led. “Opposition discourse”, on the contrary, is active, it constantly updates the dictionary in accordance with the current socio-political situation. A priori advantage to the oppositional discourse is given by the presence of the dominant liberal-democratic ideology, the key concepts in Russian society, that considered as a basic value. Oppositional discourse projects social reality, and not only states events.

An extremely interesting result of the analysis of oppositional discourse is the identification of clear socially projection techniques in verbal form. These are the methods of building up the identification of a political group and its mobilization, the image of the enemy, the moral justification of their own protest actions, etc.

Methodological conclusions

Classical linguistic approaches to the identification of key distinctive words are not always suitable for solving sociological problems. The reason is most likely in different objects of study. For linguistics, the focus of research is on texts, and for sociologists – social processes that are labeled or are accompanied by these texts. Therefore, the breadth of distribution of the identified keywords, their representativeness in a sociological sense is crucial. For socio-linguistic research, the main thing is the understanding of language as a participant in the social process both in a theoretical and applied sense. So, the found keywords will help to identify the manipulative, socially projective actions from the side of different political groups.

Final thoughts

The proposed methodology makes it possible to identify speech markers specific to a particular discourse from an array of widely used words. The method of differentiation of speech markers can be used not only for analysis of oppositional and patriotic discourse, but also for any other opposing social groups: for analyzing the discourses of different generations, nationalities, religions, movements, etc.

Identified words that differentiate opposed discourses (discourses of various social and political groups) can be subjected to additional types of statistical analysis and expert coding. It is possible to group differentiating speech markers according to the roles they perform in the overall socio-linguistic projection of the activity of the groups under study. For example, it could be ideological markers that build group's identity, slogans that motivate proactive social actions, etc.

It is possible to include the proposed method in more complex types of socio-linguistic analysis, identify socio-projecting models hidden in the texts. The identification of such socio-linguistic models and manipulative techniques in radical social movements could help counteract the spread of these movements in society.

Acknowledgments The study has been funded by the Russian Academic Excellence Project "5-100".

References

Bally, S.: Language and Life. URSS, Moscow (2003).
Dijk van, T. Discourse and communication. Walter de Gryuter, Berlin (1985).
Dijk van, T. News as discourse. Hillsdale, NJ, Erlbaum (1988).
Eissa, M. POLARIZED DISCOURSE IN THE NEWS. In: Procedia - Social and Behavioral Sciences 134 (2014), 70-91 (2014).
Fairclough, N.: Critical discourse analysis. The critical study of language. Longman, London (1995).
Gonawela, J: Studying political communication on Twitter: the case of small data. Current Opinion in Behavioral Sciences (18), 97-102 (2017).
Horvath, J.: Critical Analysis of Obama's Political Discourse. In: INTERNATIONAL CONFERENCE OF LANGUAGE, LITERATURE AND CULTURE IN A CHANGING TRANSATLANTIC 2009 (2009).
Issers, O. Communicative strategies and tactics of the Russian language. URSS, Moscow (2008).
Kim, D., Seo, D., Cho, S., Kang, P.: Multi-co-training document classification using various document representations: TF-IDF, LDA, and Doc2Vec. Information Sciences (2018).
Lotman, Y.: The Semiosphere, Art-SPB, Saint-Petersburg (2010).
Makeeva, L.: Language, ontology and realism. Publishing house of the Higher School of Economics, Moscow (2011).
Manning, C., Raghavan, P., Schutze, H.: Scoring, term weighting, and the vector space model. Introduction to Information Retrieval, Cambridge University Press New York, NY, USA (2008).
Monroe, B., Colaresi, M., Quinn, K.: Fightin' words: lexical feature selection and evaluation for identifying the content of political conflict. Political Analysis 16(4), 372-403 (2008).
Saussure, F. de.: Works on linguistics. Progress, USSR (1977).
Shipilov, A.: Opposition "us – them" in sociocultural development. Philosophical and legal thought: almanac (5), 280-304 (2003).
Tsivyan, T.: Model of the world and its linguistic basis. URSS, Moscow (2009).