Lexical and Grammatical Features of Russian-Language Tweets in Comparison with Everyday Spoken Russian

Asia Karysheva
National Research University Higher School of Economics, ITMO University
Saint Petersburg, Russia
asyakarysheva@list.ru

Abstract—**This study examines the features of computer-mediated discourse, often perceived as neither purely written nor spoken. Twitter discourse serves as a case in point, reflecting attributes of both spoken and written language. The aim of the study is to discern how closely Russian-language Twitter discourse mirrors everyday spoken Russian. A dataset of 152,223 Russian-language tweets (over 2 million tokens) was examined and juxtaposed against transcripts from the ORD speech corpus, which captures 508 macro episodes of daily conversation, totaling just under 900,000 tokens. Both lexical and grammatical aspects of the tweets and the spoken episodes are analyzed. A detailed comparison of unigrams, discourse words, and pragmatic markers is undertaken, supplemented by a multidimensional analysis spanning 22 grammatical features. The findings indicate that while the lexical attributes of Russian Twitter discourse closely align with spoken Russian, its grammatical features differ. Notably, both the tweets and the speech episodes share a significant overlap in lemmas, discourse words and pragmatic markers. However, when viewed grammatically, the Twitter discourse diverges from spontaneous spoken language. These insights hold potential for refining computer-mediated discourse generation systems for the Russian language.**

I. Introduction

This paper investigates how Russian-language tweets are similar to everyday spoken Russian. Computer-mediated discourse cannot be described as either written or spoken language since it resembles writing in terms of the means of its production but also exhibits the features of orality. The mode of Twitter messages may not be defined unambiguously as well. Although researchers have analyzed Twitter discourse in terms of its mode, Russian-language tweets have not been studied from this perspective. Since Twitter discourse is claimed to be rather close to spoken language [1] [2] [3] [4], it seems reasonable to compare Russian-language discourse on Twitter to Russian spontaneous speech.

This research focuses on lexical and grammatical features of Russian-language tweets and everyday spoken Russian. As for the vocabulary, frequency dictionaries of unigrams are compiled for both samples. The frequencies of discourse words and pragmatic markers in tweets and macro episodes of daily conversation are also compared. As far as the grammar is concerned, multidimensional analysis is conducted based on 22 grammatical features presented in the samples. Besides, proportions of parts of speech in the tweets and the speech episodes are calculated.

The paper is organized as follows: Sections II and III discuss the mode of computer-mediated discourse and of Twitter discourse in particular. In section IV, the data studied and their preprocessing are described. Section V investigates the lexical features of Russian-language discourse on Twitter and everyday spoken Russian, namely, unigrams including stop words, unigrams excluding them, discourse markers and pragmatic markers. Section VI is dedicated to the analysis of the grammatical features found in the samples under consideration. In this section, the methodology and the results of the multi-dimensional analysis are discussed. The final section provides the conclusion drawn in this study.

II. Mode of Computer-Mediated Discourse

This paper explores the characteristics of computer-mediated discourse, which is frequently seen as not strictly oral nor written. The researchers have suggested a number of terms describing the communication taking place on the Internet. Among them are “Netlish”, “Internet language”, “cyberspeak”, “electronic discourse”, “computer-mediated communication” and “Netspeak” [5]. Here, the term “computer-mediated discourse” is used since it focuses on the medium itself [ibid.] and thus seems to be the broadest of the proposed notions.

In this study, computer-mediated discourse is understood according to the definition provided by Herring and Androutsopoulos: as “the communication produced when human beings interact with one another by transmitting messages via networked or mobile computers” [6].

Herring and Androutsopoulos argue that computer-mediated discourse reveals features of both written and spoken language, depending on synchronous or asynchronous nature of the discourse in question. Computer-mediated discourse is similar to writing in terms of the means of its production. However, it also “exhibits features of orality as well as characteristics unique to itself” [6]. The researchers claim that “asynchronous modes such as email” are closer to the written language, while synchronous computer-mediated discourse such as chat has rather “oral” features [ibid.].

The similar idea is put forward by Crystal. Using the term “Netspeak” instead of “computer-mediated discourse”, he maintains that the former “relies on characteristics belonging to both sides of the speech/writing divide” [5]. Crystal distinguishes four Internet-using language situations — electronic mail, chat groups, virtual worlds, and World Wide Web, — claiming that different situations are close to different modes. In the Web, language displays properties of writing in terms of the lack of direct communication between Web page-writers and their readers. In contrast, e-mail, chatgroups and virtual worlds tend to reveal speechlike characteristics as they are "time-governed, expecting or demanding an immediate response" [ibid.]. Crystal also affirms that Netspeak "is more than just a hybrid of speech and writing" since it demonstrates electronically mediated properties unique to itself [ibid.].

Coming to particular linguistic features of computer-mediated discourse, the researchers tend to indicate its speechlike properties, written characteristics and features that are found only in the discourse in question but not in any of the modes.

As it has been brought out in the literature, the similarity of computer-mediated discourse to spoken language is manifested both at the level of vocabulary and at the level of grammar. As for the lexical features, when studying comments on YouTube videos written in Arabic, Abdul-Latif highlights that profanity bears evidence to oralization of computer-mediated discourse [7]. Concerning grammatical properties of the discourse in question, Ferrara and colleagues discover that interactive written discourse exhibits the presence of the first- and second-person pronouns as well as WH questions [8]. According to Biber, these grammatical features argue for interactivity and involvement of the discourse which, in turn, is associated with oral language and face-to-face communication [9].

With regards to written-like properties of computer-mediated discourse, they also exhibit themselves at the two levels discussed. Yates compares type/token ratio and lexical density of computer conferencing messages with the same measures calculated for written texts and spoken language [10]. He comes to the conclusion that computer-mediated discourse resembles writing in terms of vocabulary use. As to grammatical features, according to Ferrara and colleagues, interactive written discourse demonstrates high frequency of relative clauses, adverbial clauses and subordination. Cataphora, or forward sentence, also emerges in computer-mediated discourse [8]. These features are common to written language which is elaborated and expanded [ibid.].

Turning to the characteristics unique to computer-mediated discourse, they are mostly of a lexical and iconographic nature. As Crystal claims, among them are innovations related to word-formation. These include, for example, blends that are illustrated by *netiquette*, *netizen*, *infonet* and the replacement of a word-element by a similar sounding item, e.g., *ecruiting* ["electronic recruiting"] [5]. Abbreviations are a distinctive feature of computer-mediated discourse as well, e.g., *afaik* "as far as I know", *cu* "see you" [ibid.]. According to Herring, emoticons ("smiley faces composed of ascii characters") also belong to the properties unique to the discourse in question [6].

Different modes or language situations of computer-mediated discourse result in its similarity to either speech or writing. The similarity in question is found both at the level of vocabulary and at the level of grammar. Along with that, computer-mediated discourse has the characteristics common to neither spoken nor written language but unique to itself. These properties are mainly lexical and iconographic.

III. Mode of Twitter Discourse

Tweets are part of computer-mediated discourse as well. Twitter is a microblogging service, i.e., "an online platform for posting small messages to the Internet in chronological sequence" [2]. Tweets are sent and received via web, email, SMS (Short Message Service) and third-party clients which are often run on mobile devices [ibid.]. What also indicates that Twitter messages belong to computer-mediated discourse is the interactivity of the former: "Tweets also contain metadata for managing interaction with other, for instance, @ indicating address (or reference) and # labelling topic" [ibid.].

As one may expect from the part of computer-mediated discourse, the mode of tweets is considered ambiguous. Although tweets are close to writing in terms of the means of their production, in the literature, their spokenness is mainly highlighted.

Honeycutt and Herring detect the similarity of tweets to speech and their conversational nature. The researchers argue that Twitter is similar to instant messaging, yet it is more dynamic [11]. Instant messaging is a type of synchronous computer-mediated discourse that reportedly demonstrates the features that are characteristic of spoken language. Besides, the authors claim that extended conversations are found in Twitter [ibid.] which brings tweets closer to speech as well.

Zappavigna also notes that Twitter discourse is similar to spoken language, however putting emphasis on a specific nature of this likeness. She maintains that interactions via Twitter, along with interactions taking place in other social networking services, resemble conversations [2]. These conversations are defined as "searchable talk" [ibid.]. The adjective used by the researcher refers to the possibility to search for the tweets containing a specific word or a hashtag.

The spoken nature of Twitter discourse is underlined by Bounegru as well [1]. She believes that Twitter, like other microblogs, exhibits the features of "secondary orality". Secondary orality is "a mixture of literate, oral and electronic cultures in contemporary discourse". According to Bounegru, posting on Twitter is more similar to a conversation than to a written exchange. Moreover, she puts forward the idea that communication occurring on this microblogging service resembles oral storytelling.

Since the researchers argue for the closeness of Twitter discourse to spoken language, it is reasonable to address the papers defining to what extent and in which linguistic features tweets are similar to speech.

Wikström investigates the spokenness of tweets using computer-mediated discourse analysis (CMDA) [3]. According to Herring, CMDA "adapts methods from the study of spoken and written discourse to computer-mediated communication data" [12]. The research excludes a comparative perspective, i.e., Twitter discourse is not compared with either spoken language or written texts. Instead, Wikström takes a close look at how written and spoken discourse are distinguished and how Twitter data reflect the aspects of the differences in question.

The author presents four case studies that explore aspects of what talk-likeness and orality mean in digital writing [3].

Among these case studies is, for example, the research dedicated to reported speech, which is claimed to be associated with spoken discourse. As Wikström asserts, reported speech is animated in the tweets, i.e., it features non-lexical items, non-verbal and typographical elements, e.g., *& she was like 'O_O'* [ibid.].

Referring to particular linguistic characteristics arguing for spokenness of Twitter discourse, these involve fragmentariness of the tweets in terms of their form and content and the strategies of animation found in them. According to Wikström, Twitter messages are frequently "fragmented" since they contain unspecified propositional meaning and thus depart from the written norm [ibid.]. The researcher also describes the style of the tweets as spokenlike and colloquial. The Twitter posts are interactional as well which is accomplished through the animation strategies [ibid.].

Bohmann addresses the similarity of Twitter discourse to speech conducting multi-dimensional analysis and comparing tweets with four spoken and eight written registers. The spoken data categories range from private dialogue to scripted monologue, the written ones — from student writing to novels and stories [4]. The study is based on 236 linguistic features that, in turn, establish ten-dimensional space.

In his paper, Bohmann reports five dimensions in detail showing that tweets are close to spoken registers in terms of only three of them [ibid.]. These include the dimension marking a colloquial nature of the discourse, the dimension relating to a differentiation between involved and informational production and the one expressing narrative focus. The Twitter messages differ from spoken language in terms of dimensions "Collaborative communication orientation" and "Assertion of factual validity".

As Bohmann claims, the tweets appear to be even more colloquial than the private dialogues [ibid.]. That is, first person and indefinite pronouns, predominantly spoken modal expressions (*wanna*, *gotta*, *need to*), intensifier *so* etc. emerge more frequently in the Twitter posts than in spoken language. As to the second dimension common to both the tweets and the spoken registers, Twitter discourse is involved to practically the same extent the spoken registers are. This manifests itself in the low frequency of prepositions, passive constructions, a number of prefixes (*re-*, *de-*), and suffixes (*-ion*, *-ation*, *-al*, *-ment*) in both tweets and speech. Canonical narrative focus is not typical of neither Twitter posts nor spoken registers. That means third person pronouns, *could*, prefix *be-*, *the*, past perfect forms are not particularly common to neither tweets nor dialogues or monologues [ibid.].

Many researchers have put forward the idea of the similarity between Twitter discourse and spoken language. As for the linguistic features ensuring this similarity, they are both lexical and grammatical.

In the following sections, lexical and grammatical features of speech and tweets are compared. Lexical features studied involve unigrams including stop words, unigrams excluding stop words, discourse words and pragmatic markers. As for grammatical features, 22 features are extracted from tweets and transcribed speech episodes. Five functional dimensions are established as the result of multi-dimensional analysis of the features in question. Besides, proportions of parts of speech in tweets and spoken data are calculated.

Before analyzing vocabulary and grammar of Twitter discourse and spoken language, one should elaborate on the data of this research. In the next section, the description of the data and their preprocessing is provided.

IV. Data Description and Preprocessing

This study compares two samples: the first includes subsamples from the ORD speech corpus of Russian everyday communication ("*Odin Rechevoj Den'*", or "One Speech Day"), and the second consists of Russian-language tweets downloaded via Twitter API and *tweepy* library [13]. The ORD Corpus captures "Russian spontaneous speech in natural communicative situations" [14] [15]. In this research 508 transcribed and annotated speech episodes are analyzed. The Twitter sample includes 152,223 tweets downloaded between February 2 and 14, 2023.

Both samples are preprocessed by removing digits, symbols, user mentions, links, and emojis. The texts are converted to lowercase and tokenized. The preprocessed samples contain 884,790 and 842,697 tokens for the speech corpus and 2,165,193 and 2,144,243 tokens for the Twitter data with and without punctuation marks, respectively. When studying the data, the samples are balanced by the number of tokens, i.e., the Twitter sample is reduced to the size of the ORD subsample.

It is assumed that the data is preprocessed throughout the analysis, even if not specifically mentioned.

V. Analysis of Lexical Features

This section compares the frequency of unigrams in tweets and speech episodes, both including and excluding stop words. Additionally, it examines the frequency of discourse words and pragmatic markers in both samples. Punctuation marks are removed from the data when conducting lexical analysis.

A. Unigrams Frequencies (Including Stop Words)

The study analyzed Russian-language tweets and transcribed everyday spoken Russian, both consisting of over 842,650 tokens. The samples were lemmatized using a Python wrapper of the Yandex Mystem 3.1 morphological analyzer [16]. The speech episodes and tweets had 862,032 and 852,864 lemmas, respectively, with 29,762 and 67,202 unique lemmas found in the ORD Corpus and Twitter sample.

The analysis found that the two samples were only partially similar in terms of their colloquial features. The first-person pronoun and negative particle "*не*" were among the top two most frequently occurring lemmas in both samples, indicating colloquial markedness. However, the chi-square test showed that the frequency of the pronoun "*я*" ("I / me") differed significantly between the two samples, while the difference in the frequency of the negative particle was statistically insignificant. Overall, the two samples were close to each other with regard to negation but not first-person pronouns.

Despite some differences in lexical diversity and colloquial markedness, the vocabulary of the Twitter sample and speech episodes is similar, with 16,915 unique lemmas found in both. This constitutes 0.96 of the ORD Corpus vocabulary and 0.87 of the Twitter sample. Additionally, the relative frequencies of these lemmas are also similar, with a Spearman correlation coefficient of 0.61 (*p* value < 0.05). Therefore, it can be concluded that while there are some differences in style and formality, the tweets resemble spoken language in terms of their vocabulary.

TABLE I. Top 10 lemmas most common to everyday spoken Russian and Russian-language tweets, including stop words

Lemma, everyday spoken Russian	**ipm**	**Lemma, Russian-language discourse on Twitter**	**ipm**
“я” (“I / me”)	34278	“и” (“and”)	28400
“ну” (“well”)	24066	“я” (“I / me”)	27323
“не” (“not”)	23896	“в” (“in / at / into”)	26581
“вот” (“here”)	23259	“не” (“not”)	23508
“да” (“yes”)	22993	“что” (“that / what”)	15907
“а” (“and / but”)	21109	“на” (“on”)	13846
“что” (“that / what”)	20358	“быть” (“be”)	12434
“и” (“and”)	19118	“это” (“it / this”)	11927
“быть” (“be”)	18543	“с” (“with”)	10919
“это” (“this”)	18430	“а” (“and / but”)	10513

B. Unigrams Frequencies (Excluding Stop Words)

After lemmatizing the tweets and speech episodes, stop words were removed using the stop words list from the nltk module [17]. The resulting ORD Corpus and Twitter data contained 410,383 and 529,307 lemmas, respectively, with 29,638 unique lemmas in the former and 67,077 unique lemmas in the latter.

One notable finding from the 10 most common lemmas in spoken Russian and Russian-language discourse on Twitter is that reaction signals and hesitation markers are more frequent in speech episodes. “*Угу*” (“yeah”) and “*э*” (“er”) rank second and fourth in the ORD Corpus, but their relative frequencies in the Twitter sample are much lower at 40 and 55 ipm, respectively. These differences are statistically significant according to the chi-square test and support Crystal's observation that reaction signals are not typical of computer-mediated discourse [5].

Verbs are more frequent in the speech episodes than in the tweets, with five out of the 10 most common lemmas being verbs related to speaking and thinking. The relative frequencies of “*говорить*” (“say / speak / tell”), “*сказать*” (“say / tell”), and “*знать*” (“know”) in the Twitter posts are significantly lower than in the episodes of daily conversation, which was demonstrated by the results of the chi-square test. This high frequency of verbs indicates colloquial markedness, as noted by Bohmann [4].

Nouns are more prevalent in the tweets than in the speech episodes, with “*человек*” (“man/person”) and “*год*” (“year”) ranking fourth and ninth among the 10 most common lemmas in the Twitter sample. No nouns appear in the list of lemmas for speech episodes. According to the chi-square test, the differences between the frequencies of these nouns in the two samples are statistically significant, which is consistent with Biber's observation that a prevalence of nouns indicates a high informational focus not typical of spoken language [9].

TABLE II. Top 10 lemmas most common to transcribed everyday spoken Russian and Russian-language tweets, excluding stop words

Lemma, everyday spoken Russian	**ipm**	**Lemma, Russian-language discourse on Twitter**	**ipm**
“это” (“this”)	38332	“это” (“this”)	19218
“угу” (“yeah”)	13391	“весь” (“all / whole”)	6341
“говорить” (“say / speak / tell”; imperfective aspect)	12107	“который” (“which”)	5772
“э” (“er”)	11492	“человек” (“man / person”)	5540
“знать” (“know”)	11110	“мочь” (“be able”)	5362
“давать” (“let / give”)	9216	“свой” (“its / his / their” etc.)	4995
“мочь” (“be able”)	8632	“просто” (“just”)	4633
“просто” (“just”)	7532	“очень” (“very”)	4474
“весь” (“all / whole”)	6777	“год” (“year”)	4189
“сказать” (“say / tell”; perfective aspect)	6394	“хотеть” (“want”)	4021

After removing stop words from Twitter data and the ORD Corpus, it is evident that both samples still share a significant number of unique lemmas. However, Twitter posts appear to be more lexically diverse than when they contained stop words. A total of 16,792 unique lemmas are found in both Twitter posts and everyday spoken Russian, making up 0.94 of the vocabulary of speech episodes and just under 0.79 of the vocabulary of Twitter posts. The Spearman correlation coefficient between the relative frequencies of the lemmas in the samples is 0.6 (*p* value < 0.05).

Excluding stop words from the data reveals that Twitter discourse demonstrates features more typical of writing and is more divergent from speech episodes in terms of vocabulary. Notably, Twitter posts rarely contain the reaction signal “*угу*” and hesitation marker “э”, which are common in everyday spoken Russian. Verbs of speaking and thinking emerge among the 10 most common lemmas in the ORD Corpus but not in the tweets. Conversely, nouns are found in the 10 most frequent lemmas in the Twitter messages but do not appear in the 10 most common lemmas in the speech episodes. Additionally, tweets are more lexically diverse.

Despite these differences, the correlation coefficient between the relative frequencies of lemmas in both samples remains statistically significant, positive, and high in absolute value.

C. Discourse Markers

Discourse markers, which are commonly associated with oral communication, are a valuable area of study when examining the similarities between Twitter messages and spoken language. These markers play a crucial role in ensuring text coherence and reflecting the interaction between the speaker and listener. While discourse markers can be found in both written and spoken language, they are more characteristic of the latter. The current study examines the relative frequencies of 18 discourse markers in both spoken Russian and Russian-language tweets, with a focus on identifying any significant differences between the two samples.

Results indicate that the discourse marker "*в действительности*" ("in fact / as a matter of fact") is absent from both the speech episodes and the Twitter posts. Among the remaining 17 markers, "*просто*" ("just") is the most frequent in both samples. However, the difference between its frequencies in the tweets and the speech episodes is statistically significant (the chi-square test, p value < 0.05). "*Вообще*" ("in general") is the second most frequent marker in the both samples, but again, there is a significant difference between the frequencies of this marker in them. "*В общем*" ("in general / altogether") and "*почти*" ("almost") are the third most frequent markers in spoken Russian and Twitter discourse, respectively. "*В общем*" is significantly more common to the speech episodes than to the tweets; the frequency of "*почти*", in contrast, is significantly higher in the Twitter messages than in the ORD Corpus. "*Вовсе*" ("at all") and "*в самом деле*" ("indeed / really") are the least common markers in the speech episodes and the tweets, correspondingly, with "*вовсе*" being less common in the former than in the latter.

Overall, while discourse markers are present in the Twitter messages, they are more common in everyday spoken Russian. The Twitter sample contains fewer markers than the speech episodes, but there is a significant positive correlation between their relative frequencies in both samples. Therefore, it can be concluded that Twitter messages are similar to speech episodes in terms of the presence of discourse markers.

D. Pragmatic Markers

When comparing Russian-language discourse on Twitter with everyday spoken Russian, it is important to analyze the use of pragmatic markers. Pragmatic markers are speech units that possess pragmatic meanings or functions, and are commonly found in spontaneous speech [18].

The relative frequencies of 62 markers from the list put forward by Bogdanova-Beglarian [19] are calculated for the tweets and the speech episodes. "*Пятое-десятое*" ("one thing and another") is the pragmatic marker that has not emerged in any of the samples. The most common markers in the ORD Corpus are "*вот* ... [any word] *вот*" (meaning and translation of this marker are highly dependent on the context), "*вот*" ("here / this"), and "*да*" ("yes"), while "*это*" ("this"), "*так*" ("so"), and "*вот* ... [any word] *вот*" appear to be the most frequent markers in the Twitter posts. The difference between the frequency of "*вот* ... *вот*" in the two samples is statistically significant.

The least common markers in the speech episodes are "*бла-бла-бла*" ("blah blah blah") and "*то-сё*" ("one thing and another"). Six markers from the list under consideration are found in none of the tweets. The least common markers in the Twitter messages are "*бла-бла-бла*", "*или как это*" ("or how [one is supposed to say]"), and "*что называется*" ("which is called"). However, the correlation coefficient between the frequencies of the pragmatic markers in the two samples is statistically significant, indicating a close relationship between them.

Overall, while the frequencies of pragmatic markers differ between Russian-language discourse on Twitter and everyday spoken Russian, there is still a significant overlap between them.

To sum up, at the level of vocabulary, the Twitter posts can be considered similar to the ORD Corpus, yet exhibiting the features different from the latter and more typical of writing. The Twitter posts are more lexically diverse than the speech episodes and do not contain many reaction signals, hesitation markers, or verbs of speaking and thinking. Nouns are also more common in the Twitter posts, which is typical of written language. However, there is still a significant overlap between the vocabulary of the Twitter posts and everyday spoken Russian, as shown by the rank correlations between the relative frequencies of certain lemmas, discourse markers and pragmatic markers.

VI. Analysis of Grammatical Features

This section presents the methodology and results of a multi-dimensional analysis of grammatical features extracted from Russian-language tweets and the ORD Corpus. It also includes a discussion on the distribution of parts of speech in the samples. Punctuation marks are not excluded from the data during the analysis of grammatical features, as they are necessary for extracting the relevant information. For instance, the frequency of wh-question words is only counted when they appear at the beginning of a sentence, after a dot.

A. Multidimensional Analysis of Grammatical Features

In this study, a multi-dimensional analysis is conducted using Biber's methodology, which assumes that linguistic features consistently co-occur in texts and that strong co-occurrence marks an underlying functional dimension [9]. Biber emphasizes that several dimensions are required to account for linguistic variation in a language, and that dimensions are continuous rather than dichotomous. Before conducting the analysis, the Twitter data is shortened and regrouped using the functions of binpacking library [20] to obtain comparable relative frequencies of grammatical features in the samples. The number of grammatical features chosen for this study is 22, involving some features not mentioned by Biber, such as imperative mood which is associated with spoken language. These include:

- 7 features related to verb morphology: past tense, present tense, future tense, perfective aspect, imperfective aspect, imperative mood, infinitive and passive participle;

- 7 features connected with pronouns: first-person pronouns, second-person pronouns, third-person pronouns, pronoun "*это*" ("this / it"), reflexive pronouns, demonstrative pronouns, indefinite pronouns;

- Wh-question words: "*когда*" ("when"), "*кто*" ("who"), "*что*" ("what"), "*как*" ("how"), "*где*" ("where"), "*почему*" ("why"), "*который*" ("which"), "*чей*" ("whose"), "*кого*" ("whom");

- Coordinating conjunctions: "*и*" ("and"), "*а*" ("and / but"), "*но*" ("but"), "*тоже*" ("also / too"), "*также*" ("also / as well"), "*однако*" ("however"), "*зато*" ("but / on the other hand"), "*или*" ("or"), "*либо*" ("or");

- Causative subordinating conjunctions: "*поскольку*" ("since"), "*ибо*" ("because"), "*потому что*" ("because"), "*так как*" ("since / because"), "*затем (,) что*" ("because"), "*оттого (,) что*" ("because"), "*вследствие того (,) что*" ("because"), "*ввиду того (,) что*" ("because"), "*благодаря тому (,) что*" ("due to the fact that");

- Concessive subordinating conjunctions: "*хотя*" ("although"), "*несмотря на*" ("in spite of"), "*невзирая на*" ("in spite of"), "*только бы*" ("if only"), "*лишь бы*" ("if only");

- Conditional subordinating conjunctions: "*если*" ("if / when"), "*ежели*" ("if"), "*кабы*" ("if"), "*раз*" ("if");

- Subordinating conjunctions of purpose: "*дабы*" ("in order to"), "*чтобы*" ("to / in order to");

- Negation markers: "*нет*" ("no"), "*не*" ("not"), "*ни*" ("not").

Once the features have been selected, their relative frequencies are calculated in each document using the spacy package by Honnibal and Montani [21]. In this research, instances per million words are computed instead of normalizing frequency counts to a text length of 1000 words as suggested by Biber. Descriptive statistics reveal that the grammatical features are more common in the ORD Corpus than in the tweets, with higher mean relative frequencies and standard deviation in the former. Coordinating conjunctions and four verb-related features are the most common in both samples, while concessive and causative subordinating conjunctions, subordinating conjunctions of purpose, reflexive pronouns, and conditional subordinating conjunctions are the least common in both samples. Indefinite pronouns have the lowest mean relative frequency in Twitter messages.

Factor analysis is performed using a principal factor analysis with Promax rotation method, which is implemented in factor_analyzer module [22]. The best number of factors is determined by examining a scree plot of the eigenvalues, which indicates the amount of variance accounted for by each factor. Only the first several factors are considered since they explain a non-trivial amount of shared variance.

Based on the scree plot obtained, it is recommended to extract five factors. This decision is supported by the fact that each of the five factors accounts for a significant amount of variance, as shown in Table III. Although there is a noticeable break between Factor 5 and Factor 6 in the scree plot, only five factors are extracted because the variance explained by Factor 6 and Factor 7 is only slightly greater than 1. The inter-factor correlations with the largest absolute values are between Factor 1 and 2 (-0.11), Factors 1 and 5 (0.14), Factors 2 and 5 (0.12), Factors 3 and 5 (-0.12), and Factors 4 and 5 (-0.11), as reported in Table IV.

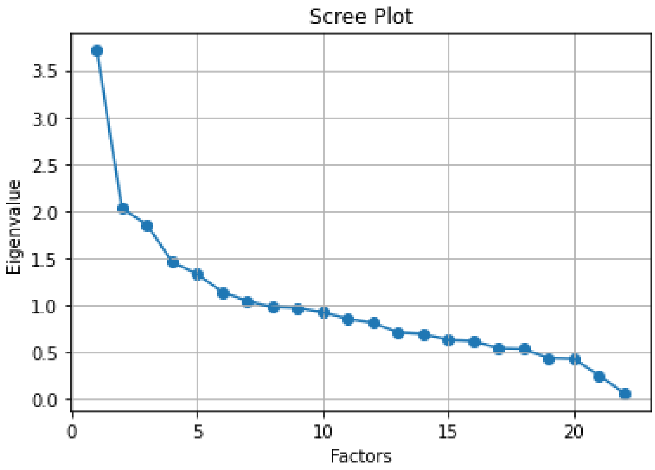

Fig. 1. The scree plot demonstrating the number of factors and the eigenvalues corresponding to them

TABLE III. THE EIGENVALUES WHICH INDICATE THE PERCENTAGE OF SHARED VARIANCE THAT IS ACCOUNTED FOR BY EACH OF 7 FACTORS

Factor	Eigenvalue
1	3.71
2	2.03
3	1.85
4	1.46
5	1.33
6	1.14
7	1.04

TABLE IV. INTER-FACTOR CORRELATIONS

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Factor 1	1.00				
Factor 2	-0.11	1.00			
Factor 3	-0.02	0.03	1.00		
Factor 4	0.01	0.06	-0.07	1.00	
Factor 5	0.14	0.12	-0.12	-0.11	1.00

After extracting the factors, it is important to identify the significant features that correspond to each factor based on their salient factor loadings. Biber explains that factor loadings indicate the degree to which a particular linguistic feature can be generalized to a factor or represents the underlying dimension of a factor. To determine the crucial grammatical features for interpreting a factor, only those with factor loadings greater than 0.30 in absolute value should be considered.

Table V shows the factor loadings of each grammatical feature for each factor. Four out of five factors have at least four significant grammatical features with loadings higher than 0.30. Factor 2 is unique in having two particularly common features: past tense and perfective aspect.

TABLE V. THE FACTORS AND THE FEATURES HAVING THE SALIENT FACTOR LOADINGS THAT CORRESPOND TO THEM AND ARE USED IN COMPUTATION OF THE FACTOR SCORES

Factor	Features	Factor loadings
Factor 1	Present tense	0.78
	Imperfective aspect	0.72
	Negation markers	0.56
	Indefinite pronouns	0.43
	Wh-question words	0.41
	Third person pronouns	0.37
Factor 2	Past tense[10]	1.07
	Perfective aspect	0.45
Factor 3	Imperative mood	0.76
	Second person pronouns	0.6
	First person pronouns	0.49
Factor 4	Infinitive	0.55
	Demonstrative pronouns	-0.5
Factor 5	Pronoun "*это*" ("this / it")	0.45
	Causative subordinating conjunctions	0.43
	Coordinating conjunctions	0.32

Next, it is necessary to standardize the relative frequencies of the grammatical features to ensure that features that occur frequently do not have a disproportionate impact on the computed factor score. This standardization process adjusts the data so that each grammatical feature is weighted based on its range of variation rather than its frequency in documents. To achieve this, the scale function from the sklearn.preprocessing package [23] is used.

Once the factor scores have been calculated for each document in the sample, the mean factor scores are determined for the different classes of data, specifically Russian-language discourse on Twitter and everyday spoken Russian. Tables VI and VII provide the mean factor scores for each class.

Regarding speech episodes, positive mean factor scores are observed for Factors 1, 3, 4, and 5. In contrast, for tweets, only Factor 2 demonstrates a positive mean factor score. This suggests that the ORD Corpus aligns closely with the dimensions represented by Factors 1, 3, 4, and 5, while only the dimension associated with Factor 2 can be identified as characteristic of the Twitter sample.

TABLE VI. DESCRIPTIVE STATISTICS OF THE FACTOR SCORES CALCULATED FOR THE ORD CORPUS

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Mean	2.9	-0.38	1.16	0.08	0.83
Standard deviation	3.79	2.28	2.81	1.65	2.58
Minimum	-10.46	-9.96	-3.82	-5.43	-8.63
25%	0.77	-1.71	-0.06	-0.84	-0.65
50%	2.97	-0.54	0.79	-0.04	0.88
75%	5.21	0.80	1.84	0.79	2.15
Maximum	24.08	13.89	31.74	14.62	12.88

TABLE VII. DESCRIPTIVE STATISTICS OF THE FACTOR SCORES COMPUTED FOR THE TWEETS

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Mean	-2.90	0.38	-1.16	-0.08	-0.83
Standard deviation	0.75	0.60	0.24	0.41	0.57
Minimum	-5.01	-1.53	-1.85	-1.42	-2.36
25%	-3.41	-0.02	-1.33	-0.35	-1.21
50%	-2.90	0.39	-1.17	-0.09	-0.85
75%	-2.35	0.73	-0.10	0.17	-0.44
Maximum	-0.86	2.47	-0.40	1.66	0.98

After obtaining the factors and factor scores for both speech episodes and tweets, the next step is to interpret the factors and identify the underlying functional dimensions. In this study, the interpretation is mostly based on the work of Biber and Bohmann.

Factor 1 is positively correlated with several features, including present tense, imperfective aspect, negation markers, indefinite pronouns, wh-question words, and third person pronouns. The functional dimension underlying this factor can be described as interactive, with a reduced form, and a minor narrative focus.

The interactivity of the discourse is demonstrated by the presence of present tense verbs, imperfective aspect verbs, wh-question words, and indefinite pronouns. According to Biber, present tense verbs refer to actions happening in the immediate context of interaction. The imperfective aspect describes ongoing actions. Wh-question words are primarily used in interactive discourse where there is a specific addressee to answer the questions. Indefinite pronouns indicate the interpersonal communicative orientation of the discourse.

A reduced discourse form is shown through the use of negation markers and indefinite pronouns. Negation markers, being analytical in this study ("*нет*" ("no"), "*не*" ("not"), "*ни*" ("not")), are associated with non-standard or fragmented presentation of information. Indefinite pronouns serve to substitute for fuller noun phrases, marking a reduced form.

The presence of third person pronouns indicates a narrative focus. These pronouns are considered markers of narrative action as they mention animate referents other than the speaker and addressee.

The salient factor loadings on Factor 2 are past tense and perfective aspect verbs, indicating a narrative dimension. Narrative discourse relies heavily on past tense and perfective aspect verbs to sequentially describe events in the past.

Factor 3 is characterized by imperative mood verbs, second-person pronouns, and first-person pronouns. This factor is similar to Factor 1 but places even more emphasis on communication with an addressee. Imperative mood verbs imply face-to-face interaction, while first- and second-person pronouns directly refer to the addressor and the addressee.

Factor 4 is the only factor with both positive and negative salient factor loadings. The factor involves infinitive verbs and demonstrative pronouns. The functional dimension underlying this factor is rather difficult to interpret. Both Biber and Bohmann find difficulty in giving a certain functional interpretation to the infinitives. However, the demonstrative pronouns having a negative factor loading may suggest the presence of less generalized content in the discourse.

Factor 5 is represented by the pronoun "*это*" ("this / it"), causative subordinating conjunctions, and coordinating conjunctions. This factor indicates discourse form reduction and the expression of personal feelings or attitudes. Causative subordinating conjunctions are associated with affective functions, serving as markers of emotions or beliefs.

The distinction between the coordinating conjunctions and their interpretation is not straightforward. They may contribute to fragmented presentation of information or mark a structurally complex, abstract linguistic style.

In terms of functional dimensions, the ORD Corpus and the Twitter messages differ significantly. The speech episodes are highly interactive, express personal feelings, and have a reduced form. On the other hand, the tweets have a clear narrative focus.

The findings provide evidence that Russian-language discourse on Twitter exhibits characteristics more typical of written language rather than spoken language. The features observed in spoken language, such as high interactivity and personal expression, are not prominent in tweets. Instead, the analysis suggests that tweets are more narrative-focused, aligning them with written language. The divergence between spoken and written language in terms of functional dimensions is significant, indicating that the discourse on Twitter differs from everyday spoken Russian.

B. Proportions of Parts of Speech in the Samples

Although the multi-dimensional analysis conducted did not include relative frequencies of part-of-speech tags, it is worth examining them when analyzing the grammatical features of Russian-language discourse on Twitter and everyday spoken Russian. For instance, adjectives are commonly used in evaluative language, which is typical of speech [4]. Adverbs, particularly time and place ones, are considered more characteristic of speech as they rely on shared physical and temporal situations [9].

To examine the distribution of parts of speech in Russian-language discourse on Twitter and everyday spoken Russian, the samples are first equalized in size. Then, a part-of-speech tag is assigned to each word in both samples using the spacy package [21]. The proportions of parts of speech in the Twitter messages and the ORD Corpus are calculated (see Table VIII). The proportions are found to be highly correlated with a Spearman correlation coefficient of 0.96 (*p* value < 0.05).

TABLE VIII. THE PROPORTIONS OF PARTS OF SPEECH IN RUSSIAN-LANGUAGE DISCOURSE ON TWITTER AND EVERYDAY SPOKEN RUSSIAN.

Parts of speech	Everyday spoken Russian	Russian-language discourse on Twitter
Adjectives	0.05	**0.07**
Adpositions	0.07	**0.09**
Adverbs	**0.11**	0.07
Auxiliaries	0.01	0.01
Coordinating conjunctions	0.04	0.04
Determiners	0.03	0.03
Nouns	0.14	**0.22**
Particles	**0.10**	0.05
Pronouns	**0.13**	0.08
Proper nouns	0.02	0.05
Punctuation	0.13	0.12
Subordinating conjunctions	0.03	0.03
Verbs	0.13	0.13

In summary, the analysis of the distributions of proportions of different parts of speech in the ORD Corpus and Russian-language discourse on Twitter reveals several significant differences. Adverbs, particles, and pronouns are more likely to occur in everyday spoken Russian, while adjectives, prepositions, and nouns are more common in tweets.

This finding is supported by previous research, which suggests that adverbs and particles contribute to coherence in fragmented speech, while pronouns are indicative of interactive and reduced speech. On the other hand, adjectives are often associated with detailed presentation of information, suggesting a connection to writing. Prepositions and nouns are also more likely to be found in writing, as they help to integrate high amounts of information and contribute to the density of the text.

Overall, these differences suggest that Russian-language discourse on Twitter exhibits more characteristics of written language than spoken language. This is further supported by the results of the multi-dimensional analysis, which has demonstrated that the narrative focus dimension, typically associated with writing, had positive factor scores for the tweets.

In conclusion, the grammatical features of Russian-language discourse on Twitter indicate a closer resemblance to writing than spoken language. Both the proportions of different parts of speech and the factor scores from the analysis of the tweets support this observation.

VII. CONCLUSION

Based on the findings, Russian-language Twitter discourse aligns closely with everyday spoken Russian in its lexical elements, though not in its grammatical structure. The tweets and spoken episodes have a significant overlap in terms of lemmas. Both datasets also exhibit discourse words and pragmatic markers, even if their prevalence is less in the Twitter posts compared to the ORD Corpus. Notably, the correlation between the two sets remains strong.

On the grammatical front, multidimensional analysis reveals that the tweets are predominantly guided by a narrative function, which is not a typical feature of oral communication. Another deviation from spoken language is the strong presence of nouns and prepositions in Twitter posts.

In essence, while the lexical aspects of the tweets lean more towards spoken forms, their grammatical elements are more reminiscent of written structures. A more exhaustive study is required for a comprehensive understanding of the nature of Twitter discourse. An extended multidimensional analysis encompassing additional grammatical or lexical features, such as discourse words and pragmatic markers, could be insightful. Furthermore, juxtaposing the lexical and grammatical traits of Twitter communication with established written language samples would offer a clearer placement of tweets on the oral-written spectrum.

This research offers insights with potential practical applications, particularly in the realms of Natural Language Processing (NLP) and linguistic education. Firstly, for developers and researchers in the NLP sector, understanding the nuanced linguistic characteristics of Russian-language

Twitter discourse can enhance the efficacy of tools designed for sentiment analysis, chatbots, or automated customer support on social media platforms. The results of this study also indicate that Twitter messages may be used as the training data for the model that is supposed to generate neither too formal (i.e., strictly written) nor too informal (i.e., extremely close to spoken language) text or speech.

Secondly, for educators and curriculum designers, the findings of this study can be instrumental in developing contemporary language teaching methodologies. Recognizing the hybrid nature of computer-mediated discourse, especially as seen on popular platforms like Twitter, can pave the way for modernized language courses that incorporate real-world, relevant examples of digital communication. This can be especially beneficial for courses aiming to teach Russian as a foreign language, where learners can benefit from understanding both the traditional spoken forms and the evolving digital nuances of the language.

Acknowledgment

The paper was prepared within the framework of the Basic Research Program at HSE University in 2023 (project “Text as big data: the modeling of convergent processes in language and speech by computational methods”).

References

[1] Bounegru, Liliana. Secondary Orality in Microblogging. 2009, https://lilianabounegru.org/2009/11/20/secondary-orality-in-microblogging/. Accessed 7 March 2023.

[2] Zappavigna, Michele. Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web. Bloomsbury, 2012.

[3] Wikström, Peter. I Tweet Like I Talk. Aspects of Speech and Writing on Twitter. 2017.

[4] Bohmann, Axel. “Situating Twitter Discourse in Relation to Spoken and Written Texts – Sprache auf Twitter in Abgrenzung zu gesprochenen und geschriebenen Texten”. Zeitschrift für Dialektologie und Linguistik, vol. 87, no. 2, 2020, pp. 250–284.

[5] Crystal, David. Language and the Internet. Cambridge University Press, 2001. http://dx.doi.org/10.1017/CBO9781139164771

[6] Herring, Susan, and Androutsopoulos, Janis. “Computer-Mediated Discourse 2.0.” The Handbook of Discourse Analysis, John Wiley & Sons, 2015, pp. 127–51, https://doi.org/10.1002/9781118584194.ch6.

[7] Abdul-Latif, Emad. “The Oralization of Writing: Argumentation, Profanity and Literacy in Cyberspace.” The Politics of Written Language in the Arab World, edited by Høigilt, Jacob, and Mejdell, Gunvor, Brill, 2017, pp. 290–308, http://www.jstor.org/stable/10.1163/j.ctt1w76vkk.17.

[8] Ferrara, Kathleen, et al. “Interactive written discourse as an emergent register.” Written Communication, vol. 8, no. 1, 1991, pp. 8–34.

[9] Biber, Douglas. Variation Across Speech and Writing. Cambridge University Press, 1988.

[10] Yates, Simeon J. “Oral and Written Linguistic Aspects of Computer Conferencing: A Corpus Based Study.” Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives, edited by Herring, Susan, 1996, pp. 29–46.

[11] Honeycutt, Courtenay, and Herring, Susan. “Beyond Microblogging: Conversation and Collaboration via Twitter.” Proceedings of the 42th Hawai’i International Conference on System Sciences, 2009, pp. 1–10.

[12] Herring, Susan. “A Faceted Classification Scheme for Computer-Mediated Discourse.” Language@Internet, vol. 4, no. 1, 2007, pp. 1–37.

[13] Roesslein, Joshua. “Tweepy: Twitter for Python!” Github, 2020, github.com/tweepy/tweepy. Accessed 5 January 2022.

[14] Asinovsky, Alexander, et al. “The ORD Speech Corpus of Russian Everyday Communication “One Speaker’s Day”: Creation Principles and Annotation”. Text, Speech and Dialogue, 12th International Conference, 2009, pp. 250–257, https://doi.org/10.1007/978-3-642-04208-9_36.

[15] Sherstinova, Tatiana, et al. “Sistema annotirovaniya v zvukovom korpuse russkogo yazyka “Odin rechevoy den'” (Annotation System in the ORD Speech Corpus).” Mat-ly XXXVIII mezhdunarodnoy filologicheskoy konferentsii, 2009, pp. 66–75.

[16] “nlpub/pymystem3.” Github, https://github.com/nlpub/pymystem3. Accessed 18 March 2023.

[17] Bird, et al. Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc., 2009.

[18] Bogdanova-Beglarian, Natalia, Filyasova, Yulia. “Discourse vs Pragmatic Markers: A Contrastive Terminological Study.” 5th International Multidisciplinary Scientific Conference on Social Sciences and Arts, SGEM 2018, vol. 5, iss. 3.1, 2018, pp. 123–130.

[19] Bogdanova-Beglarian, Natalia (eds.). Pragmaticheskiye markery russkoy povsednevnoy rechi: slovar'-monografiya (Pragmatic Markers of Russian Everyday Speech: Dictionary-Monograph). Nestor-Istoriya, 2021.

[20] Maier, Benjamin. “binpacking 1.5.2.” PyPI, 30 Nov. 2021, https://pypi.org/project/binpacking/. Accessed 25 February 2023.

[21] Honnibal, Matthew, and Montani, Ines. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017.

[22] “Factor-analyzer 0.4.1.” PyPI, https://pypi.org/project/factor-analyzer/0.2.2/. Accessed 5 May 2023.

[23] Pedregosa, F., et al. “Scikit-learn: Machine Learning in Python.” Journal of Machine Learning Research, vol. 12, 2011, pp. 2825–2830.