Russian Journal of Linguistics
ISSN 2687-0088 (print), ISSN 2686-8024 (online)

2023 Vol. 27 No. 3 543–569
http://journals.rudn.ru/linguistics

https://doi.org/10.22363/2687-0088-33757
EDN: MRDMOZ

Research article / Научная статья

The words that make fake stories go viral: A corpus-based approach to analyzing Russian Covid-19 disinformation

Alina G. MONOGAROVA[1], Tatyana A. SHIRYAEVA[1]✉ and Elena V. TIKHONOVA[2]

[1]Pyatigorsk State University, Pyatigorsk, Russia
[2]MGIMO University, Moscow, Russia
✉shiryaevat@list.ru

Abstract
Since the outbreak of the Covid-19 pandemic in 2020, the spread of the new virus has been accompanied by the growing infodemic that became a dangerous prospect for Internet users. Social media and online messengers have been instrumental in making fake stories about Covid-19 viral. The lack of an efficient instrument for classifying digital texts as true or fake is still a big challenge. Deceptive content and its specific characteristics attract attention of many linguists, making it one of the most popular contemporary topics in corpus-based research. This paper explores the language of viral Covid-related fake stories and identifies specific linguistic features that distinguish fake stories from real (authentic) news using quantitative and qualitative approaches to text analysis. The study was conducted on the material of the self-compiled diachronic corpus containing Russian misleading coronavirus-related social media posts (a target corpus of 897 texts) which were virally shared by Russian users through social media platforms and mobile messengers from March 2020 to March 2022 and the reference corpus containing genuine materials about the virus. First, we compared two corpora using an interpretable set of features across language levels to find whether there is evidence of significant variation in the language of fake and real news. Then, we focused on frequency profiling to extract other over-represented groups of words from both corpora. Finally, we analyzed the corresponding contexts to indicate whether these features can be considered as linguistic trends in Russian Covid-related fake story making. Findings regarding the role of these over-represented groups of words in fake narratives about coronavirus revealed efficiency of frequency profiling in indicating lexical patterns of the language of deception.
Keywords: *Covid-19, fake story, infodemic, disinformation, frequency profiling*

For citation:
Monogarova, Alina G., Tatyana A. Shiryaeva & Elena V. Tikhonova. 2023. The words that make fake stories go viral: A corpus-based approach to analyzing Russian COVID-19 disinformation. *Russian Journal of Linguistics* 27 (3). 543–569. https://doi.org/10.22363/2687-0088-33757

Язык вирусных фейковых новостей: корпусный подход к анализу русскоязычной дезинформации о Covid-19

А.Г. МОНОГАРОВА[1], Т.А. ШИРЯЕВА[1]✉, Е.В. ТИХОНОВА[2]

[1]Пятигорский государственный университет, Пятигорск, Россия
[2]МГИМО МИД России, Москва, Россия
✉shiryaevat@list.ru

Аннотация
С самого начала пандемии Covid-19 в 2020 году распространение нового вируса сопровождалось нарастанием инфодемии, в результате которой Интернет-пользователи получали огромное количество ложной и потенциально опасной информации. Социальные сети и онлайн-мессенджеры сыграли важную роль в транслировании различных фейковых сообщений о Covid-19. Отсутствие эффективного инструмента обнаружения текстов, содержащих дезинформацию, по-прежнему является серьезной проблемой. Интересным видится рассмотрение специфических характеристик подобного контента с позиций корпусной лингвистики. Цель настоящей статьи – на основе изучения русскоязычных текстов вирусных фейковых историй о Covid-19 определить ключевые языковые черты, отличающие подобные истории от аутентичных новостей, а также выявить лексические особенности языка фейков. Исследование проводилось на материале составленного авторами диахронического корпуса русскоязычных фейков о Covid-19 (целевой корпус, состоящий из 897 текстов), распространяемых российскими пользователями через социальные сети и мобильные мессенджеры в период с марта 2020 по март 2022 года, а также референтного корпуса, в текстах которого представлена подтвержденная факт-чекинговыми организациями информация о коронавирусе. В качестве первого шага мы сравнили представленность различных лингвистических особенностей в целевом и референтном корпусах. Кроме того, мы извлекли из целевого корпуса несколько высокочастотных групп слов и проанализировали соответствующие контексты ложных нарративов, чтобы сделать вывод о том, можно ли рассматривать данные лексические группы в качестве специфических характеристик языка фейковых новостей. Полученные результаты позволяют выделить ключевые лексико-грамматические и стилистические различия фейковых историй и верифицированных новостей о Covid-19, а также демонстрируют эффективность корпусного подхода к выявлению лексических паттернов языка дезинформации.
Ключевые слова: *Covid-19, фейк, инфодемия, дезинформация, анализ частотности*

Для цитирования:
Monogarova A.G., Shiryaeva T.A., Tikhonova E.V. The words that make fake stories go viral: A corpus-based approach to analyzing Russian Covid-19 disinformation. *Russian Journal of Linguistics.* 2023. V. 27. № 3. P. 543–569. https://doi.org/10.22363/2687-0088-33757

1. Introduction

Social media's power to spread deceptive content instantly has become one of the key factors in the development of the Covid-19 digital infodemic that fueled a lot of conspiracy theories and misinformation about the new virus in 2021–2022 (Kopytowska & Krakowiak 2020, Gisondi et al. 2022, Pavlina 2022). Unfortunately, anxiety over distressing fake news is not the biggest impact of the infodemic. The recent study (Islam et al. 2020) claims that in the first year of the

Covid-19 pandemic more than 5,800 people around the world were admitted to hospitals after following fake medical recommendations virally shared on social networks.

Academic research on the nature of fake news may contribute to overcoming the challenges and dangers offered to the public by the viral dissemination of deceptive content. Although the term 'fake news' has already entered scholarly discourse (Tandoc & Lim 2017), there is still no unambiguous and simple definition of the phenomenon. Grieve and Woodfield (2023) explained this by pointing to the shifts in fake news due to the explosive growth of social networks and media. Habgood-Coote (2019) stated that the term 'fake news' does not have a stable publicly accepted meaning and is used to undermine the credibility of the media. We believe, however, that the key difference between fake news and simply untrue information is the driving force behind them, namely, fake stories and news are written deliberately to spread a false message. Therefore, we follow Allcott and Gentzkow (2017) who define 'fake news' as news articles that are intentionally and verifiably false and could mislead readers. It is most likely that motivation for creating fake news is either commercial (viral messages attract attention of potential customers) or ideological (false stories can be used to promote a candidate or to ruin a reputation). Thus, when analyzing linguistic features of fake news, it is necessary to focus on the fact that fake news providers may use different strategies to appeal to different readers.

However, we must note the ambiguous role of fake news in highlighting some significant social issues. According to Beckett (2017), fake news gives mainstream quality journalism the opportunity to show that it has value based on expertise, ethics, and experience. Besides, fake news can provoke a meaningful debate. During the Covid-19 pandemic, fake news has surprisingly contributed to the development of the discussion about public health measures and vaccines. Many people frightened by fake stories about vaccines and tests began to study these issues more deeply. However, fake news broadcasts false information, trust in which can lead readers to wrong conclusions and, as a result, wrong decisions. Therefore, an in-depth study of these materials it is necessary to detect false information more effectively.

The language of the narratives deliberately created to mislead people is of great interest to contemporary linguistics (e.g., Gjylbegaj 2018, Sutu 2020, Ahmed et al. 2018). Corpus technologies can be implemented to mine quantitative and qualitative information about content structure and style of electronically stored data. Frequency profiling proved to be useful in uncovering some techniques used in fake story making (Zhang & Ghorbani 2020). Recent works on the language of fake news feature extraction techniques such as detection of unreliable news using *n*-grams and application of semantic similarity metrics (Ahmed 2017). Besides, corpus-based quantitative data analysis is used to identify systematic nuances between fake and fact-checked news with the focus on exploring the social context (Mahyoob 2021).

This study aims to identify specific linguistic features that distinguish Russian viral fake stories about Covid-19 from real news and to display some lexical patterns frequently used in coronavirus-related fake story making. The methodology used in this paper involves: 1) building a corpus of viral fake stories that circulated on social media during the first two years of the Covid-19 pandemic (March 2020 – March 2022), and compiling a reference corpus containing news collected from reliable sources (websites of fact-checking organizations); 2) indicating differences in the target and reference corpora using a set of interpretable linguistic features; 3) analyzing three groups of words over-represented in the corpus of Russian Covid-related fake stories (references to influential social actors, coronavirus-related neologisms and dysphemisms); 4) exploring corresponding contexts to indicate whether the use of these groups of words can be considered as lexical trends in Russian coronavirus-related fake story making.

2. Corpus-based approach to text analysis

The growing application of the corpus-based approach to text analysis (e.g., Chen et al. 2020, Muslimah 2020, Lu et al. 2021) can be attributed to the fact that it offers a number of efficient tools for exploring large amounts of digital data, searching for lexical and structural units and evaluating their statistical significance (Kytö 2010). However, the choice of the tools for performing corpora in-depth investigation depends on the research agenda, and we should first outline the objectives that are achievable within the framework of the study.

Fake stories about Covid-19 began spreading almost immediately after the first reports of the new virus. Over time, the number of viral misleading narratives grew, as did the number of plots around which fake stories were created. A diachronic corpus can facilitate the analysis of the word frequency distribution in different periods of the pandemic. Counting frequencies of specific units diachronically is one of the methods of historical corpus linguistics that offers researchers a set of instruments for mining evidence of language change (Baron et al. 2009). According to (Curzan 2009), historically organized data captures stages of linguistic development over time providing linguists access to contrastive or comparative studies of the language. Our research adopts this approach as we focus on the diachronic evolution of the language of deception. It allows us to capture peaks and troughs in collected data (Brezina 2018), in other words, to indicate the periods when analyzed tokens were used more or less frequently, which is important for understanding whether rises and drops in the word frequencies are determined by social context.

The new pandemic has produced a number of corpus-based studies (e.g., Christopher & Simon-Vandenbergen 2021, Goddard & Wierzbichka 2021, Ponton 2021, Lun et al. 2022, Peng & Hu 2022) that use frequency profiling as an effective tool in diagnosing Covid-related digital content. In particular, quantitative and qualitative collaboration method can be applied to explore a wide range of issues from discourse characteristics of different text types to influence of Covid-19 on

language patterns. Muslimah revealed high frequency of critical strategies in digital content about the new coronavirus based on the frequency of the corresponding tokens representing indirect criticism (Muslimah 2020).

A number of statistical techniques can be applied to compare distributions of specific groups of words and to determine the words that can be found in the corpus significantly more or less frequently than expected (Baron et al. 2009). When the word count indicates notable changes in frequencies of the units, which generally have a stable distribution, it may provide significant information on the types of text being studied (Sinclair 1991).

This work can be viewed as an application of Rayson's approach to qualitative corpus-based research (Rayson 2019) which involves comparing a target corpus with a reference corpus to discover differences in the language. In our case, we need to evaluate the differences between the corpus containing fake stories about Covid-19, and the corpus containing reliable materials about the same range of topics. However, differences observed when comparing corpora may be purely accidental. Therefore, the extracted features must be tested for significance using the log-likelihood coefficient (computing a *p-value)* (Rayson & Garside 2000).

The proposed method can be extended to comparative diachronic studies to track the transformation of language strategies used for covering a topic during different time periods (Essam & Abdo 2021). Analyzing English and Chinese Covid-19 discourse, Yu compares Covid-related news before and after the lockdown (extracting frequently used vocabulary and *n*-grams from self-built corpora) demonstrating transformation of "Covid-19 descriptions in the UK media into a more objective and neutral one than before" with an increased use of expressions of restriction and social conflicts (Yu et al. 2021). Therefore, a corpus-based approach can be beneficial in indicating distinctive features of fake news.

3. Materials and methods

3.1. Methodology

This study employs mixed methods adopting quantitative and qualitative approaches to text analysis. The overall research strategy involves six main tasks: building a target corpus (Corpus 1 and Corpus 2) of Russian Covid-related fake stories for each year of the pandemic (with distribution of texts by months) and compiling a reference corpus of the materials published by reliable fact-checking organizations; evaluating and comparing representation of specific linguistic features in Russian coronavirus-related fake stories and in real news by applying QDA Miner to raw corpora; further data preprocessing (lemmatization, stop words removal, lowercasing); making a word frequency list for each corpus and checking the lists for other over-represented categories (or groups of words) which were not captured with QDA screening; testing the significance of the observed differences by calculating log-likelihood values and sorting the words by the significance score; analyzing corresponding contexts (actual fake stories and news about Covid-19

form Corpus 1 and Corpus 2) to indicate whether these features (observed differences) can be considered as lexical trends in Russian coronavirus-related fake story making.

3.2. Data collection and preprocessing

For our previous study of fake narratives about Covid-19 (see Monogarova et al. 2021), we compiled a corpus of false Covid-19-related stories that had been virally shared by Russian social media users from March 2020 to March 2021 (hereinafter referred to as Corpus 1). However, over the next year, as the pandemic continued, the accompanying infodemic did not slow down either. Therefore, new data were added to Corpus 1, and it was expanded by a collection of digital texts (hereinafter referred to as Corpus 2), representing the same types of fake stories – deliberately false texts, virally shared by Russians via Telegram, Viber, WhatsApp, Vkontakte, Odnoklassniki, Facebook and Instagram[1] from March 2021 to March 2022 (Table 1). When compiling Corpus 2, we relied on the same principles of building specialized text corpora, which were described in detail in our study of topic change in the Covid-19 disinformation (Ibid, p. 87). To ensure the balance of Corpus 2 we only included texts that meet the following criteria: verifiability (reliable fact-checking organizations proved that the stories are false); viral popularity (within the framework of this study a text is considered viral if it has more than 50,000 unique digital views); maximum character limit of 2000 characters.

Notably, Corpus 1 is characterized by a larger genre diversity of viral texts, which is associated with the fading public interest in this topic in mid-2021. By comparison, Corpus 2 contains only 3 scripted audio messages with fake announcements, while this genre was popular during the first year of the pandemic (with 77 scripts included in Corpus 1). In this case, we believe that a slight register variation in the data structure is acceptable, since the purpose of this work is to indicate the distinctive linguistic features of all the fake stories about Covid-19 that gained viral popularity and most likely were perceived as reliable by many Russian social media users (judging by the high numbers of reposts). Corpus 1 was registered with the Russian Federal Service for Intellectual Property as a database[2], and Corpus 2 is in the process of obtaining a certificate at the time this paper is being prepared. When presenting examples of fake narratives in this paper, we refer to the episode number under which the stories are found in these two databases.

To compare the linguistic features represented in fake stories and real news about Covid-19, we built the reference corpus containing actual fact-checked news. When compiling the reference corpus, we consider its balance and

[1] Facebook and Instagram are social media services, parts of Meta Platforms Inc., added to the register of extremist organizations and banned in the Russian Federation.

[2] Monogarova, Alina & Alexander Bagiyan. 2021. Russian text bank of fake news and their linguistic features. Database #2021621693, registered with the Federal Service for Intellectual Property of the Russian Federation 08/14/2021.

representativeness, including a comparable number of texts on similar coronavirus-related topics from trusted sources (materials on Covid-19 published by the fact-checking organization StopFake and translated articles published by Covid Infodemic Europe and Coronavirus Facts Alliance). As a result, we prepared two corpora to be compared – the target corpus consisting of two collections of texts representing fake stories of the first (Corpus 1) and the second (Corpus 2) years of the Covid-19 pandemic and the reference corpus containing real news about Covid-19.

Table 1. **Corpus Structure before and after Data Cleansing**

			Before data cleansing		***After data cleansing***	
Corpus ID	***Time periods covered***	***Number of episodes***	***Total words***	***Unique word forms***	***Total words***	***Unique word forms***
Corpus 1 (fake stories)	March 2020-March 2021	491	45,205	23,552	26,964	16,002
Corpus 2 (fake stories)	April 2021-March 2022	406	39,966	20,193	22,261	14,984
Reference Corpus (real news)	March 2020-March 2022	825	76,017	38,931	39,895	21,011
Total		**1722**	**161,188**	**82,676**	**89,120**	**51,997**

Further corpora transformations were determined by the specifics of the following analytical operations. To evaluate the distribution of linguistic features in the language of fake stories and genuine coronavirus-related materials, we applied QDA miner to the raw corpora, only removing such elements as graphic materials, dates, timing, and numbers of episodes. However, frequency profiling used to discover other distinctive features of the Russian fake stories about Covid-19 was carried out on the preprocessed data. We preprocessed the target and the reference corpora performing text lemmatization and stopwords removal (which is extracting and deleting the words such as pronouns, prepositions, and conjunctions that do not give significant information about analyzed discourse) using Natural Language Toolkit (NLTK) written in Python. Further data cleansing involved lowercasing and removal of punctuation.

3.3. Exploring differences between Russian fake stories and real news about Covid-19

The proposed method of the linguistic analysis of fake and real coronavirus-related news is based on the investigation of differences between the target and the reference corpora using an interpretable set of linguistic features for identifying meaningful distinctive characteristics between deceptive and non-deceptive content. However, the choice of linguistic attributes for analyzing differences between fake news and real news is a very challenging task. In this regard, we

should mention Grieve & Woodfield's (2023) detailed analysis of the real and fake articles by Jayson Blair of the *New York Times* based on Multidimensional Analysis Tagger. This approach proved to be effective in comparing patterns of grammatical variation in Blair's real and fake news as it allowed researchers to identify differences in the values of forty-nine features measured across every article in two corpora making interesting conclusions about the variations in frequencies of nouns and redicative adjectives in fake news. Mahyoob et al. (2021) compare collections of real and false news from social networks based on a bundle of discriminating linguistic features and attributes which are chiefly stylistic features of news (e.g., reported speech, quotation, proper nouns). As part of the lexical approach to text comparison, various lexical resources (e.g., LIWC) are applied to real and fake news articles (see Rashkin et al. 2017). This method may be insightful in revealing specific features of the lexicon used in fake news.

To compare representation of the linguistic features in Russian coronavirus-related fake stories and reliable news, we make use of a QDA Miner, which is research software for coding and analyzing qualitative data. The matrix of linguistic features used in this study is based on a set of linguistic attributes that could representatively reveal the distinguishing features of fake and real news. For example, we compared the use of superlative, comparative and subjective adjectives because creators of deceptive content tend to use a lot of subjective words as they dramatize or sensationalize a news story, while authentic news items use more comparative adjectives (Raskin et al. 2017). We also compared the use of stative verbs, passive voice, and modal verbs, as according to recent studies (see Kuzmin et al. 2020), misleading texts are often characterised by frequent use of these verbs and verb-forms. We also included the first, second and third person pronouns in the set of features, because their frequent occurrences are traditionally attributed to the language of deception (Pisarevskaya 2017).

Using QDA Miner, we assigned codes to a set of features across language levels and applied them to annotate the data in both corpora. These features include first person pronoun, second person pronoun, third person pronoun, stative verb, modal verb, passive voice, proper noun, abstract noun, adverb of manner, conjunctive adverb, comparative adjective, superlative adjective, subjective, sentence length (short < 10 words; long > 20 words), reported speech, quotation, negation, interrogative, exclamation, and terminology (Figure 1).

Then we investigated information about distribution of these 21 linguistic features in both corpora and extracted linguistic characteristics using the QDA clustering. After that we retrieved relative frequencies of 21 codes/linguistic features from both collections of texts and tested for differences in the relative frequencies of 21 linguistic features between the target and the reference corpora. The automatic count performed with the UCREL log-likelihood wizard was used to show how frequently a linguistic feature appears in fake stories and in real news (significance testing based on log-likelihood values is described in more detail in 3.3).

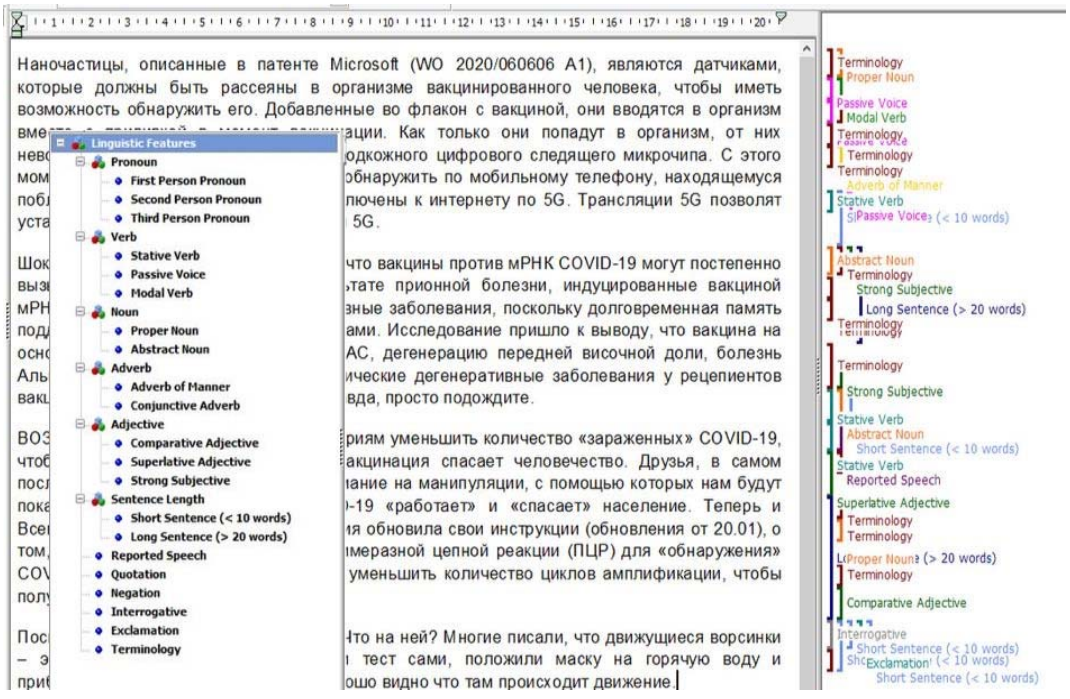

Figure 1. Sample of the target corpus annotated with the QDA Miner codes assigned to the linguistic features

3.4. Testing significance of other observed differences

After comparing patterns of lexico-grammatical variation in the target and the reference corpora, we performed frequency profiling to check for other significant language patterns in Russian fake news stories that were not captured by QDA screening. After generating two frequency lists displaying raw frequencies of all the words in both corpora, we found two groups of words (coronavirus-related neologisms and dysphemisms) that were over-represented in the target corpus. Frequency lists also revealed that although the numbers of overall occurrences of proper names in the two corpora are comparable, fake stories often used certain proper nouns to refer to some influential social actors, but these names were significantly under-represented in real news. To make sure that the observed differences are not just a random deviation, we tested the significance of these three groups of words (proper nouns referring to real-life personalities, neologisms and dysphemisms) using the log-likelihood (hereinafter referred to as LL) test. In other words, we compared frequencies of the words in the target corpus with the frequencies of the same words in the reference corpus taking account of the sizes of both corpora.

LL values help us determine discourse significance of the words (Baron et. al. 2009) to see if a significant difference in the frequency of use of the same words in the target and the reference corpora can flag some lexical trends and give us considerable information about the way deceptive content is organized (Ahmed,

2017). We calculated the log-likelihood statistics for each of the words in the two lists using the UCREL log-likelihood wizard, created by Paul Rayson, and sorted the words by significance score (Table 2) establishing significance at the $p < 0.05$ level as a cut-off point. That means that the results displaying LL < 3.84 are considered not significant and the observed differences between corpora are most likely accidental (Rayson & Garside 2000).

Table 2. **Observed Differences sorted by LL values (generated with the UCREL log-likelihood wizard)**

Word	***Observed frequency in the target corpus***	***%1***	***Observed frequency in the reference corpus***	***%2***	***LL***
гейтс/gates	185	0.38	14	0.04+	140.82
рокфеллер/rockefeller	73	0.15	0	0.00+	86.66
ротшильд/ rothschild	73	0.15	0	0.00+	86.66
пландемия/plandemic	34	0.07	0	0.00+	40.36
намордник/muzzle	31	0.06	0	0.00+	36.80
сорос/soros	26	0.05	0	0.00+	30.87
фаучи/fauci	44	0.09	5	0.01+	27.98
хондзе/honjo	11	0.02	0	0.00+	13.06
гебреисус/ghebreyesus	55	0.11	23	0.06+	7.66
радзуэлина/rajoelina	14	0.03	3	0.01+	5.60
…					
%1 and %2 – observed frequencies in normalized (percentage) form **+ sign** indicates that the word is more frequent, on average, in the target corpus					

According to the results of the LL test, three main differences observed while comparing the target and the reference corpora (the frequent use of certain names, neologisms and dysphemisms) are statistically significant. The big LL values of these words are determined by their zero or minimum representation in the reference corpus. For instance, some key names (Rothschild, Soros, Rockefeller), around which many viral fake coronavirus-related stories were created, were not found at all in the reference corpus. Thus, we assume that the frequent use of this vocabulary in intentionally deceiving narratives may indicate the use of certain linguistic strategies by the authors of fake stories. Although frequent use of proper names is traditionally attributed to authentic news (Mahyoob et al. 2021), over-represented references to some influential personalities found in fake narratives about Covid-19 might be an interesting feature for an in-depth linguistic analysis. In the Results section (4.2. and 4.3.), we will take a closer look at this vocabulary and the way it is represented in the Russian coronavirus-related fake story making.

A number of Voyant tools (open-source application developed by S. Sinclair and G. Rockwell for analyzing digital texts) were also applied to the preprocessed target corpus to extract collocates of the words of interest and to visualize relative frequencies of individual words diachronicaly depicting the distribution of a word's occurrence in every month of the analyzed period – from March 2020 to March 2022 (graphs in sections 4.2., 4.3. are generated with Voyant tools).

4. Results and discussion

4.1. Linguistic features reflecting differences between Russian coronavirus-related fake stories and real news

In this section, we present the quantitative results, testing for differences in the relative frequencies of 21 features across language levels between 897 fake and 825 real news stories about Covid-19, and discuss discovered distinctive features.

Table 3 displays differences between fake and real news indicated by differences in LL values and classifies variables as showing large ($p < 0.0001$; critical value = 15.13), medium ($p < 0.001$; critical value = 10.83), small ($p < 0.01$; critical value = 6.63), or insignificant effects ($p < 0.05$; critical value = 3.84). We sorted LL values so that the largest LL value is placed at the top of the list representing the most distinctive feature of fake stories as compared to real news.

Table 3. **Significance testing for the differences in the relative frequencies of 21 features between the coronavirus-related fake and real news**

Codes/ linguistic features	***O1***	***%1***	***O2***	***%2***	***LL***	***Ratio***	***Effect***
Exclamation	319	4.45	17	0.21+	374.97	4.43	large
Interrogative sentence	387	5.39	154	1.87+	138.83	1.53	large
Conjunctive adverb	105	0.13	315	0.41-	126.87	-1.70	large
Short sentence (< 10 words)	853	11.89	548	6.64+	116.28	0.84	large
Reported speech	673	9.38	438	5.31+	88.26	0.82	large
Passive verb	305	0.37	514	0.68-	71.44	-0.87	large
Second person pronoun	511	0.62	263	0.35+	62.75	0.85	large
Comparative adjective	206	0.24	347	0.46-	54.17	-0.92	large
Abstract noun	619	0.75	815	1.07-	44.30	-0.51	large
Stative verb	931	1.13	628	0.83+	38.04	0.46	large
First person pronoun	105	0.13	32	0.04+	35.51	1.60	large
Subjective adjective	132	0.16	60	0.08+	22.36	1.03	large
Terminology	875	1.15	769	0.94+	17.58	0.30	large
Proper noun	519	0.63	601	0.79-	14.09	-0.32	medium
Quotation	347	4.84	298	3.61+	13.69	0.42	medium
Superlative adjective	157	0.19	94	0.12+	11.46	0.63	medium
Modal verb	501	0.61	370	0.49+	10.90	0.32	medium
Long sentence (> 20 words)	785	10.94	824	9.99+	3.32	0.13	insignificant
Adverb of manner	291	0.35	304	0.40-	2.20	-0.18	insignificant
Third person pronoun	5298	6.45	4978	6.55-	0.62	-0.02	insignificant
Negation	364	5.07	405	4.91+	0.21	0.05	insignificant

O1 – observed frequency of the feature in the target corpus
O2 – observed frequency of the feature in the reference corpus
%1 – relative frequency of the feature in the target corpus
%2 – relative frequency of the feature in the reference corpus
+ sign indicates that the word is more frequent, on average, in the target corpus
- sign indicates that the word is more frequent, on average, in the reference corpus
Ratio refers to how frequently the feature appears in fake stories as compared to real news

As observed in Table 3, such linguistic features as third person pronoun (LL 0.62), adverb of manner (LL 2.20), long sentence (LL 3.32) and negation (LL 0.21) do not reflect any substantial linguistic differences in real or fake stories. However, retrieved quantitative data provides strong evidence of significant variation in other 18 lexico-grammatical and stylistic patterns used in viral coronavirus-related fake stories and real news about Covid-19.

The most distinctive feature between Russian coronavirus-related fake and real news is the use of ***exclamation***, which is very common in the fake stories but is significantly under-represented in the real news. Most likely, this is due to the desire of fake news providers to emphasize personal attitude to the problem expressing shock, surprise or other strong emotions. This finding is in line with the results of the recent research on political misinformation (Oehmichen et al. 2019) which show statistical significance of the differences in syntactic style of misleading and reliable news. Writers of misleading posts do not generally avoid emotional statements, otherwise, they tend to be "distinctive in their use of language" making greater use of exclamation marks and capitalization (Ibid). While news writers try to appear unbiased (Rashkin et al. 2017) focusing on unemotional presentation of unbiased stories, fake stories providers tend to overuse the patterns of spoken language.

We found four common patterns of use of exclamation marks in coronavirus-related fake stories (see Table 4). We see that exclamation is not only used as an intensity marker, but also indicates the bits of information that should be the focus of the reader's attention. Sociolinguists point to the shifts in use and perception of exclamation marks in informal communication on social networks. According to McCulloch (2019), exclamation is being used not as an "intensity marker, but as a sincerity marker", showing politeness and softening polite requests. This observation may contribute to the understanding of pragmatic reasons for over-representation of exclamatory sentences in the language of fake stories. Thus, we can assume that the frequent use of the indicated exclamation patterns is due to the writer's intention to appeal more personally to the reader.

Table 4. **Relative frequencies of the exclamation patterns in fake news**

Pattern	***Examples***	***%***
Exclamation mark in a declarative sentence	*(1a) <...> Ко всем будут ходить врачи с полицейскими. Отказывайтесь, от любых тестов на вирус. ЭТО ВАШЕ ПРАВО ОТКАЗАТЬСЯ**!!!** <...>* [*<...> Doctors and policemen will visit everyone. Refuse to take any tests for the virus. IT'S YOUR RIGHT TO REFUSE**!!!** <...>*] (episode #73, April 2020)	47.02
Exclamation + imperative verb	*(1b) Ни за что **не делайте** ПЦР-тест**!** В Германии врач провел под микроскопом исследования теста ПЦР на COVID-19. И обнаружил на кончиках тестов, металлические скобы, которые реагируют на волны 5G... <...>* *[**Do not take** a PCR test**!** In Germany, a doctor examined a PCR test for COVID-19 under the microscope. And he found metal staples that respond to 5G waves on the tips of the tests … <...>]* (episode #42, March 2020)	28.21

Exclamation + interrogation	*(1c) ВОЗ признала самоизоляцию граждан вредной для борьбы с COVID-19. И зачем тогда нас посадили на эти карантикулы**?!** Глава Всемирной организации здравоохранения (ВОЗ) Тедрос Гебреисус признал, что самоизоляция граждан … <…>* *[The WHO acknowledged that self-isolation hampers the fight against COVID-19. Why do we need this lockdown then**?!** The head of the World Health Organization (WHO), Tedros Ghebreyesus, admitted that the self-isolation of citizens ... <...>* (episode # 97, March 2020)	18.81
Exclamation mark as intensifier inside a sentence	*(1d) Испанские исследователи обнаружили, что вакцина Pfizer содержит 99: оксид графена **(!)** и практически больше ничего. <...>* *[The Spanish researchers found that Pfizer's vaccine contains 99% graphene oxide **(!)** and practically nothing else. <...>]* (episode #303, December 2020)	5.96

Similarly, ***interrogative sentences*** tend to be substantially more common in fake news. Browsing the context for this feature in the fake news corpus allowed us to discover that interrogative sentences are frequently embedded into the beginning of a fake story as a means of formulating and defining the topic (2a), and into the conclusion (2b) giving the readers "food for thought" and leading them to certain conclusions. In our dataset 148 coronavirus-related stories began either with a general or with a special question (using interrogative adverb *почему (why*), and 128 fake news articles used disjunctive question as a closing remark.

(2a) ***Почему*** *доктора в Германии начали массово писать увольнительные? Докторам предлагают 12000 евро в месяц за участие в геноциде – проведении массовой вакцинации. <...> [**Why** have the doctors in Germany started quitting their jobs recently? The doctors are offered 12,000 euros a month for participating in a genocide – mass vaccinations. <...>]* (episode #330, December 2020)

(2b) *<...> Все, что мы видим, означает, что есть специальные поддельные шприцы, чтобы обмануть общественность. В свою очередь, это имеет вряд ли какой-то смысл, если вакцинация безвредна**, не так ли?** [<...> We see that there are special fake syringes to deceive the public. In turn, this doesn't make any sense if the vaccination is so harmless, **does it?]* (episode #339, December 2020)

We suppose that in addition to structuring the narrative, interrogative sentences also have a pragmatic meaning. As can be seen from the examples above, questions help the authors of fake news to dialogize the narrative. A similar conclusion was reached by Ivanova (2020) who stated that interrogatives make the argumentation more emphatic, and solicitate active commitment to issues, feedback and empathy from the audience.

Another feature that exhibits a non-negligible difference between Russian fake and real news about Covid-19 is the use of ***ultra-short sentences*** which are more common for fake news. 482 sentences in the fake news corpus are just one or two words, and in 285 cases one of the words is an imperative verb (3a, 3b) or a noun (3c):

(3a) <...> ***Остановитесь!*** *В вакцине от COVID-19 есть вещества, повреждающие мозг! Эта упаковка от «атиковидной» вакцины фармакомпании Астра Зенека, которой будут вакцинировать британцев, я разобрала состав <...> [<...>* ***Stop it****! There are substances in the COVID-19 vaccine that damage your brain! This package is from Astra Zeneca's "Aticoid" vaccine, which will be used to vaccinate the British people, I studied the ingredients ... <...>]* (episode #314, December 2020)

(3b) *Этот вирус не был выделен. Положительный результат теста может получить кто и что угодно. Даже курица или апельсин.* ***Просто прочитайте!*** *За последние 55 лет <...> [This virus has not been isolated. Anyone can get a positive test result. A chicken or an orange. Just* ***read this****! Over the past 55 years <...>]* (episode #375, January 2021)

(3c) <...> Итак*, графеновая ковидная жижа от Pfizer содержит: ХЛОРИСТЫЙ КАЛИЙ, ОДНООСНОВНЫЙ ФОСФАТ КАЛИЯ, ХЛОРИД НАТРИЯ, ФОСФАТ НАТРИЯ.* ***Внимание!*** *Липидные наночастицы, защищающие РНК <...> [<...> So, Pfizer's graphene covid slurry contains: POTASSIUM CHLORIDE, POTASSIUM MONONE PHOSPHATE, SODIUM CHLORIDE, SODIUM PHOSPHATE.* ***Attention!*** *Lipid nanoparticles protecting RNA <...>]* (episode #508, June 2021)

Interestingly, the number of long sentences (>20 words) in both datasets is not significantly different. However, long sentences in fake news differ from the long sentences in real news in terms of structure. Most of them are simple, or compound sentences joined with coordinating conjunctions. Real news is characterized by a much greater variety of complex sentences. This explains the fact that reliable news tends to use more conjunctive adverbs, as seen in Table 3.

The use of ***terminology*** is the most distinctive lexical feature under analysis. Most notably, terminology is used at substantially higher rates in fake news. In addition to basic Covid-related terms, e.g., *коронавирус (coronavirus), ПЦР-текст (PCR test), вакцинация (vaccination*), fake news tends to contain more frequent use of specific medical terminology. Besides, fake stories often use domain-specific terms which do not appear in real news (Figure 2). This finding, however, does not corroborate previous work by Torabi & Taboada (2019) who stated that on average fake news articles use overly emotional language, while frequent use of terminology was indicative of reliable news. However, the fact that terminology appeared to be a significantly presented lexical group in the fake news about Covid-19 may be determined by the nature of the topic (disease, its symptoms, safety measures). Register and genre variations affect the distribution of terminology in fake stories. Social media posts are less likely to contain terms, while deceptive articles often use domain-specific terminology. This result is best explained by functional theories of language use (Biber & Conrad 2019) according to which differences in communicative purpose and context are reflected in linguistic structure.

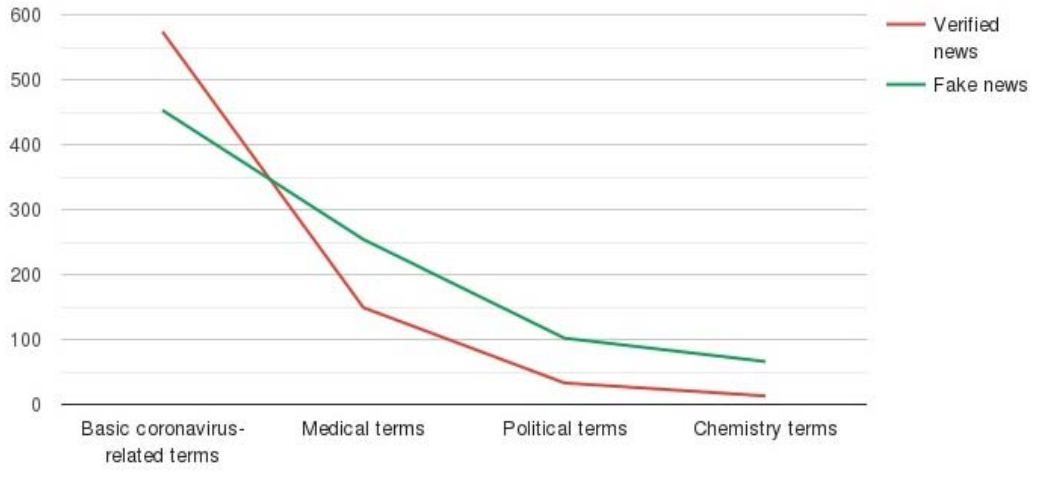

Figure 2. Distribution of the terminology in fake and real news about Covid-19

Overall, we find that, on the one hand, Russian fake news about Covid-19 is characterized by more frequent use of various lexico-grammatical and stylistic features associated with emotional discourse. On the other hand, fake news tends to use a lot of terminology, reported speech and quotations which indicates the intention of fake story writers to create well-structured, highly informative texts. We will further illustrate and discuss other distinctive features of fake and real news in Sections 4.2 and 4.3. using the material of some fake stories virally shared during the Covid-19 pandemic.

4.2. Real-life people as central characters of fake stories

As noted in 3.4, references to some real-life personalities which were over-represented in the Russian Coronavirus-related fake news, appeared much less frequently in real news or did not occur at all. In this section, we will take a closer look at how viral fake stories about some influential people are organized in terms of composition and style, and attempt to evaluate the role of the frequent use of these names in the language of coronavirus-related disinformation.

The most common proper noun within Corpus 1 and Corpus 2 is Bill Gates (token *гейтс* (*gates*) has 185 occurrences as a reference to a person and 15 occurrences as part of the phrase *Фонд Билла и Мелинды Гейтс*/*The Bill and Melinda Gates Foundation)*. The following three positions are occupied by the names of billionaires–Rockefeller (token *рокфеллер* (*rockefeller)* ranks 134th with 73 occurrences), Soros (token *сорос* (*soros*) appeared 26 times in Corpus 1), and Rothschild (token *ротшильд* (*rothschild*) is mentioned 73 times). The name of the Madagascar President Rajoelina (token *радзуэлина (rajoelina*) appeared in fake stories 15 times during the spring and summer of 2020. The name of the Japanese scientist Honjo (token *хондзе*) appeared 11 times in August 2020, becoming the most popular proper noun of this period.

The context analysis of the episodes containing corresponding names showed that these famous persons are turned into either protagonists or antagonists of fake stories–either villains and organizers of the pandemic (Gates, Rockefeller, Rothschild, Soros), or truth-tellers exposing secret information about the WHO (Rajoelina), and the Wuhan laboratory (Honjo). During the second year of the Covid-19 pandemic, the proper nouns *gates, rothschild, rockefeller* were also frequently used in viral fake stories (Figure 3). In addition, the texts of conspiracy theories involving the current Director-General of the WHO Tedros Adhanom Ghebreyesus (*гебреисус* (*ghebreyesus*) – 55 occurrences within Corpus 2) and the American infectious disease specialist Anthony Fauci (*фаучи* (*fauci*) – 44 occurrences within Corpus 2) began to gain popularity during the period from April 2021 to December 2021 (Figures 3 and 4).

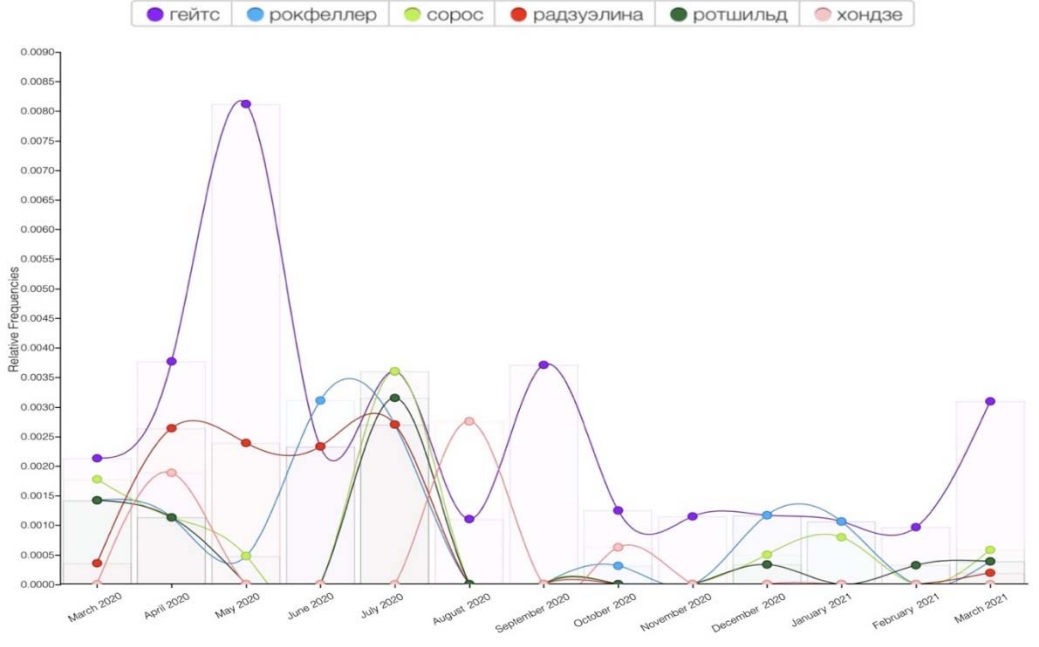

Figure 3. Frequencies of the references to influential social actors in fake stories across Corpus 1 (March 2020-March 2021)

Notably, most episodes (97 out of 118) involving real-life personalities represent "international" fake stories. Russians actively shared translations of English-language texts or retellings of conspiracy theories. Public figures from Russia are practically absent in viral fakes. Although the names of some Russian politicians and experts are found in Corpus 1 and Corpus 2 (e.g., *мясников (miasnikov)* – 14 occurrences, *юдин (yudin)* – 10 occurrences, *гаряев (garyaev)* – 7 occurrence), none of them is the central character of a separate story.

Browsing for contexts in the target corpus indicated several compositional patterns used by fake news writers. As can be seen in Table 3, coronavirus-related fake news often used reported speech, references to reputable sources, citations of

field experts to appear more credible. The high frequency of quotes in fake texts seems to be an interesting finding: 113 of 118 untruthful texts with references to influential people, officials or field experts contained quotes or reported speech. A similar pattern of results was obtained by Mahyoob et al. (2021) who found that fake news articles tend to use more quotes and reported speech than reliable news. However, we are also interested in the way quotations were embedded in the fake stories. The writers of Russian Covid-related fake news articles most often chose the 'distorted quote' strategy, shortening the real quote, changing its meaning. The transformed quote was often placed in the headline of fake news to attract more public attention. The next strategy involves extracting a quote outside of the original context, namely embedding the real quote in a fake story to give it more credibility. Less commonly, a completely made-up text was attributed to a well-known expert in the field (Table 5).

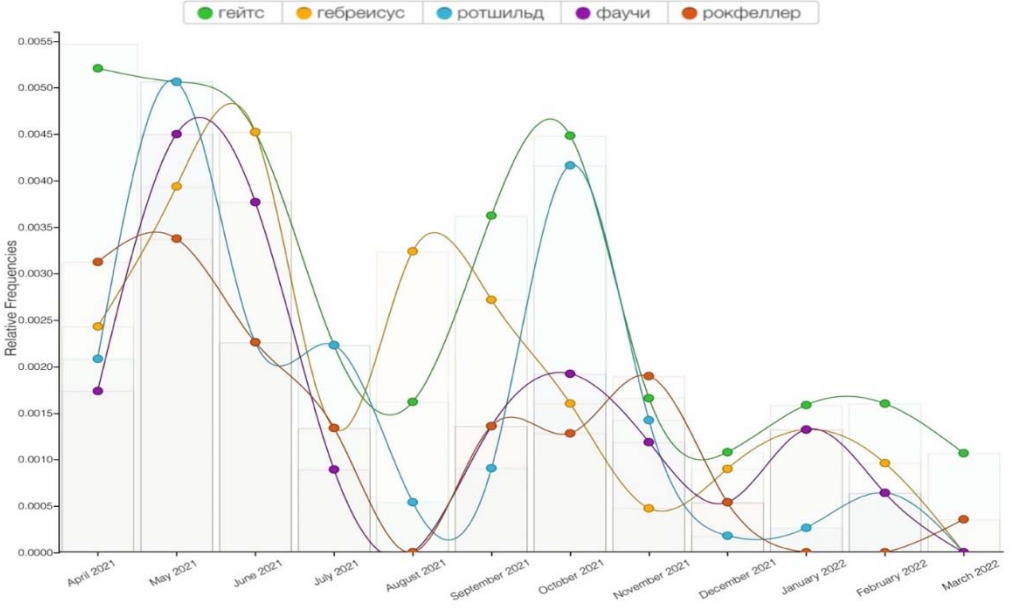

Figure 4. Frequencies of the references to influential social actors in fake stories across Corpus 2 (April 2021-March 2022)

Table 5. **Patterns of embedding quotations, reported speech and references into fake stories**

Pattern	***Number of episodes Corpus 1***	***Number of episodes Corpus 2***
Distorted quotation	37	21
True quotation embedded in a fake story	25	16
Fake story falsely attributed of an expert in the field	9	5
Total	**71**	**42**

The first strategy can be illustrated with a fake story about President of Madagascar Andry Rajoelina (4). In 2020 Rajoelina became the central character in a series of international viral fake stories about coronavirus as his name occurred

in 11 episodes attributing to him the words that he never said. In all the stories united by this term, Rajoelina appears as a whistleblower disclosing the information about WHO's proposal to poison the COVID-19 medication developed in Madagascar. The following text that was actively shared by Russian Internet users from March to July 2020 promotes the most popular Rajoelina-related storyline claiming that the WHO offered the President of Madagascar a $ 20 million bribe to add poison to local experimental coronavirus medicines. The example shows that the reported quotation was significantly distorted by inserting the words *взятка (bribe) and отравление (poisoning)* which originally did not appear in Rajoelina's statement.

(4) *Президент Мадагаскара Андри Радзуэлина* ***заявил, что ВОЗ предложила ему взятку в размере 20 миллионов долларов за отравление используемого в стране лекарства от COVID-19 под названием «COVID-19 Organics», изготовленное из артемизии.*** *Мадагаскар попросил прикрыть эту лавочку из аферистов и призвал все страны выйти из ВОЗ. [Madagascar President Andry Rajoelina* ***stated that the WHO had offered him a $20 million bribe to poison the country's artemisia-based COVID-19 drug called COVID-19 Organics.*** *Madagascar called for the dissolution of this organization of swindlers and called on all countries to withdraw from the WHO]* (episode #161, July 2020).

The third pattern does not use a quotation but the name of an influential expert, embedding it in a fake story. This can be illustrated with a virally shared story about the Japanese scientist-immunologist, the Nobel Prize winner in physiology or medicine Tasuku Honjo, who was turned into one of the major ideologists of Covid-19 dissidence in Russia by fake story makers in April 2020 (5). According to a text widely shared on Russian social media, Honjo allegedly states that Covid-19 is an artificial virus that leaked from a Chinese laboratory. This false statement, which the scientist never made, exists in several variations with minor transformations. This example also demonstrates another distinguishing feature of fake materials which is frequent use of the first-person pronouns.

(5) *Шок!!! Японский профессор физиологии и медицины, доктор Тасуку Хондзе, вызвал сегодня сенсацию в средствах массовой информации, заявив, что коронавирус не является естественным. «<...>* ***Я работал уже четыре года в Уханьской лаборатории в Китае и знаю весь персонал этой лаборатории. Я позвонил им всем после появления информации о коронавирусе, но все их телефоны были отключены уже не менее трех месяцев*** *<...>. [Shocking!!! Japanese professor of physiology and medicine, Dr. Tasuku Honjo, caused a media sensation today by saying that the coronavirus is not a natural virus. "<...>* ***I have been working for four years at the Wuhan laboratory in China and I know all the staff of this laboratory. I called them all when the news about the coronavirus came out, but all their phones had been switched off for at least three months*** *<...>"]* (episode #176, August 2020).

The choice of these strategies by Russian fake story writers can be explained in both social and psychological terms. In April 2020, due to the increasing adverse effects of the infodemic, the Russian authorities began an active fight against fakes through the release of official refutations of fake news and materials debunking the conspiracy theories. The level of public skepticism towards unverified information grew. This was accompanied by the employment of new strategies in fake story making. Since anonymous statements on Covid-related topics appeared to be less credible than the words of well-known scientists, fake news often contained references to reputable sources and citations of field experts. This idea is in line with the recent work by Khan et al. (2021) who pointed to the use of references to influential persons to elevate credibility of deceptive content.

The majority of Covid-19 fake news posts and articles involving businessmen Bill Gates, George Soros, the Rockefellers, and the Rothschilds are based on the claim that the Covid-19 pandemic was planned by billionaires in cooperation with the WHO to "turn people into slaves" by vaccinating them against an "invisible virus" that does not exist. Since the beginning of the pandemic, Bill Gates has been the main target of multiple conspiracy theories spread on social media. The BBC even called him "the voodoo doll of Covid conspiracies" (Wakefield 2020). In Russian viral narratives during the first year of the pandemic, Bill Gates was often referred to as *«создатель дивного нового мира» (the creator of a new brave world), «стоящий за пандемией» (a person behind the pandemic), «главный вакцинатор» (the main vaccinator).*

A common feature of the fake news articles involving references to Bill Gates is the excessive use of subjective adjectives (*шокирующий, невероятный, гениальный, необъятный, катастрофический / shocking, incredible, ingenious, immense, catastrophic*) and superlatives (*хитрейший, мощнейший, богатейший/ the smartest, the most powerful, the richest*) as well as abstract nouns related to the semantic category of LIE (*обман, грабеж, надувательство, афера/ deceit, robbery, swindle/ fraud*). This feature can be illustrated by a very popular fake story (6), according to which the Covid-19 pandemic was planned and funded by Bill Gates back in 2012 to be executed in 2019 to make money from the coronavirus vaccine.

(6) *<...> Если кто не в курсе, то все* драконовские меры*, принятые во многих странах, были разработаны* ***хитрейшим Биллом*** *Гейтсом при прогоне тренировочных действий при Всемирной пандемии в октябре 2019 года. <...> Поэтому, сразу после объявления пандемии, ВОЗ дала правительствам план реагирования, который и заключался в принятии,* ***глупейших*** *в научном плане и* ***катастрофических*** *в экономическом плане, мер. <...> [<...> Just for your information, all the draconian measures taken in many countries were developed by* ***the most cunning*** *businessman Bill Gates during the training activities during the World Pandemic in October 2019. <...> Therefore, immediately after the announcement of the Covid-19 pandemic, the WHO gave governments the plan, which involved taking*

> ***the stupidest*** *and* ***the most disastrous*** *measures <...>]* (episode #89, April 2020)

After analyzing the relevant contexts, we conclude that the over-representation of the names of real influential persons is more related to the specifics of the Covid-19 conspiracy making but is not characteristic of fake news.

4.3. Neologisms, dysphemisms and negative opinion shaping

According to the frequency distribution data, other lexical trends in Russian Covid-19 fake story making is the use of neologisms and dysphemisms. Most of the neologisms found in the analyzed narratives dated March 2020 – March 2021 are nouns formed as a result of morphological and syntactic word composition. They do not give names to new objects or emerging realities but are used to devalue and discredit the phenomena that already have names (*пандемия/пландемия – pandemic/plandemic*; *коронавирус/барановирус* – no English equivalent, rough translation – *a virus that only sheep (stupid people) believe in*; *прививка/прибивка* – no English equivalent, rough translation – *a vaccine that will kill you*).Although there is a small variety of forms of dysphemisms in the corpus (7 tokens) (Figures 4 and 5), their actual frequency (204 occurrences per 26,964 words) is high compared to zero number of occurrences in the reference corpus.

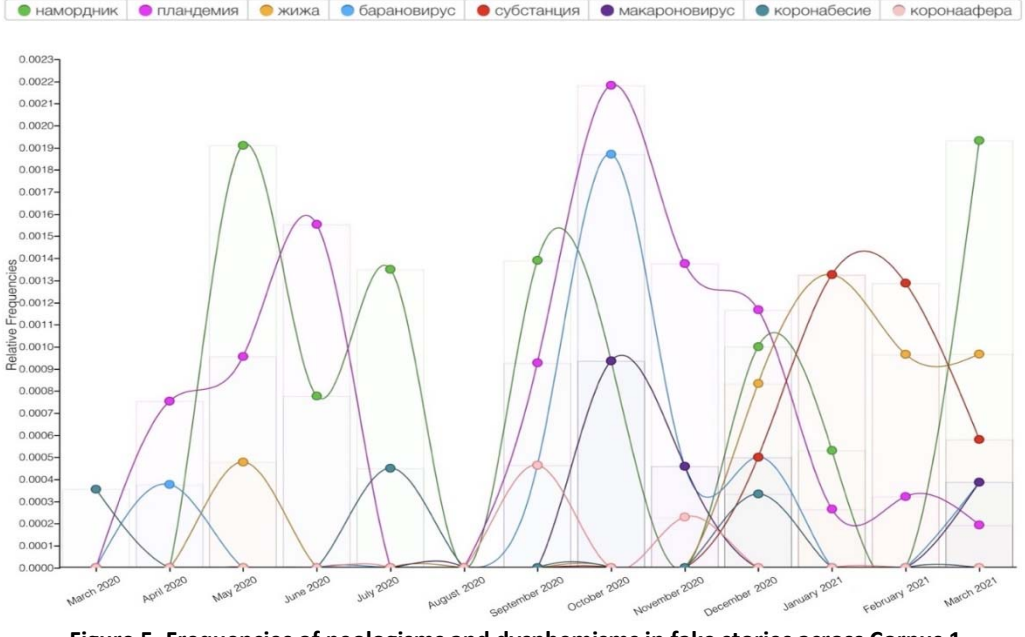

Figure 5. Frequencies of neologisms and dysphemisms in fake stories across Corpus 1 (March 2020-March 2021)

Analysis of the corresponding episodes containing neologisms shows that in all contexts these words have negative connotations, and their use in texts is associated with the author's desire to criticize the new rules dictated by the

pandemic (wearing masks, vaccination, testing for covid, etc.). The most common corpus neologism *пландемия/plandemic (*план+пандемия/ plan+pandemic) ranking 115th with 34 occurrences, is used as a substitute for the Russian term *пандемия* (*pandemic*) and is associated with the conspiracy theory about the coronavirus pandemic being planned by world's elite in cooperation with the WHO a decade ago.

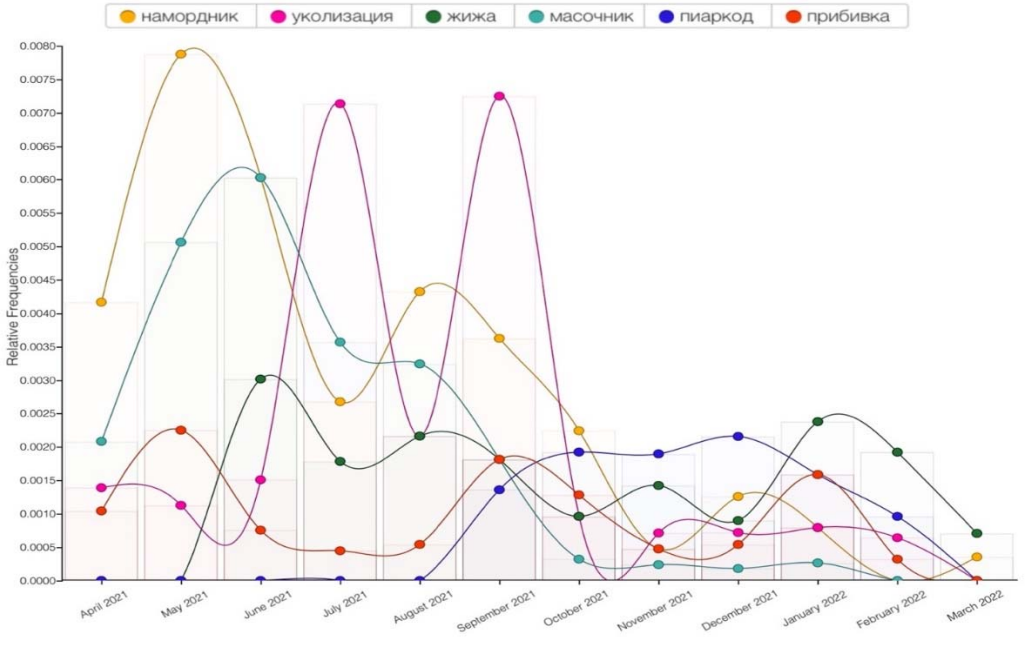

Figure 6. Frequencies of neologisms and dysphemisms in fake stories across Corpus 2 (April 2021-March 2022)

The frequency distribution list shows 5 newly coined lexemes denoting *Covid-19* which were often used in viral texts calling for vaccine refusal or denying the existence of the virus as an effective language tool for negative opinion shaping. These neologisms are formed by replacing one of the stems: *корона (corona)* or *вирус (virus*) in the compound term *коронавирус* (*coronavirus*). The word *барановирус* (баран+вирус/ ram+virus) is based on a comparison of people who believe in Covid-19 with sheep/rams. Another popular neologism *макаронавирус* (макароны+вирус/ noodles+virus) compares the new virus with deceiving the population. It is based on the Russian set expression "вешать лапшу на уши" which is equivalent to the English idiom "to hang noodles on one's ears" meaning "to fool or mislead someone". Other new words are based on the analogy of the coronavirus with madness – *коронабесие* (корона+бес, бешенство / corona+ madness), analogy with fraud – *коронаафера* (корона+афера / corona+fraud) and comparison with paranoia: *коронапараной*я (корона+параной/ corona+paranoia).

The new nouns denoting groups of people are significantly less represented in both Corpus 1 and Corpus 2: *масочник* (a person wearing a mask) (10 occurrences), *безмасочник* (a person not wearing a mask) (14 occurrences), *ковидиот* (covid+idiot) (5 occurrences). Interestingly, the term *ковидиот* is used in the texts of fake stories in two opposite meanings – to name the people who deny the pandemic and its danger, neglect precautions, as well as the people who, on the contrary, are very afraid to get infected and panic severely. The surge in new words during the Covid-19 pandemic is in line with the thesis by Al-Salman & Haider (2021) who stated that linguistic change and creativity as a universal property of language reflects global social changes.

The neologisms listed above are found in clusters with different terms, as they are used in various fake stories covering a range of topics. However, all the episodes where new evaluative substitutions for the term *Covid-19* were found (e.g., 7) are united by the idea of Covid-dissidence (denial of the fact that this virus exists).

(7) *Не прошло и полтора года, как CDC признал, что PCR-тесты не подходят для тестирования на* ***барановирус****, поскольку не могут отличить его от других болячек, находят в пепси-коле и арбузе. Лабораториям дан срок до 31 декабря, после чего им нужно будет перейти на другие способы тестирования <...> [In less than a year and a half, the CDC recognized that PCR tests are not suitable for testing for* ***baranovirus*** *(Russian neologism – ram+virus), as they cannot distinguish it from other diseases, the tests find the virus in pepsi cola and watermelon. Laboratories are given a deadline of December 31, then, they will need to switch to other testing methods <...>]* (episode #212, September 2020)

Dysphemisms are marked word forms which differ from neutral vocabulary as they are motivated by either fear or hatred or humor and expresses an author's attitude towards the subject (Terry, 2020: 59). In Russian fake Covid-19 stories, the aggressive potential of these words was used to criticize masks and vaccines. In 31 episodes, the term *маска* (*mask*) is substituted with a word *намордник* (*muzzle*). From December 2020 to March 2021 COVID-19 vaccine was frequently called *жижа* (*slurry*), *бульон* (*broth*) and *субстанция* (*substance*). In fake narratives, means of protection are framed as instruments of control and deception of the population.

Periods of the increasing popularity of the texts that use the dysphemism *намордник* (*muzzle*) to aggressively convince the readers of the futility of face masks coincide with the introduction of the requirement to wear masks and gloves in public places in Moscow (May 12, 2020) and active public debate on the mandatory masks in schools in September 2020 (the beginning of the school year in Russia) and in December 2020 – January 2021 (the end of distance learning for schoolchildren). The word *бульон* (*broth*) as a substitution for the term *вакцина* (*vaccine*) was used several times before the start of mass vaccination in Russia.

However, from November 2020 to March 2021, the nouns *жижа* (*slurry*) and *субстанция* (*substance*) occupy the leading positions among dysphemisms in the corpus. We assume that a large number of texts that became viral during this period contribute to the rapid "fading" of dysphemisms, namely, when a frequently used derogatory word ceases to produce the desired effect on the reader, and a new expressive replacement is required.

5. Conclusions

Significant growth of deceptive content on the Internet has exposed the urgent need for further development of automatic text analysis in order to classify data as fake (misleading) or factual (reliable). Using experimental procedures described above, we demonstrated the application of corpus technologies to determining lexical patterns typical of a deliberately misinforming narrative.

This study provides clear evidence of lexico-grammatical and stylistic variation in the language of Russian coronavirus-related fake and real news. We found that 18 of 21 analyzed linguistic features indicate significant differences between misleading and reliable data. The most distinctive features are the use of terminology and subjective adjectives, systematic stylistic nuances between fake and real news include the use of exclamation, interrogation, ultra-short sentences, quotations and reported speech.

Frequency profiling helped us determine the trends in the use of particular groups of words achieving specific goals of the authors of fake narratives. The high frequency of references to influential celebrities is related to the fact that a series of fake Covid-19 stories are based on false storylines where famous real-life people were associated with the actions that they did not perform or the words that they did not say. The names of real people in Covid-related fake narratives are used either to make disinformation sound more convincing (e.g., information on behalf of scientists or experts), or to fuel conspiracy theories (e.g., stories about famous businessmen being involved in the spread of the virus). A higher-than-expected frequency of neologisms and dysphemisms across the target corpus points to the desire of fake story makers to shape negative attitudes towards the objects of the new reality for promoting Covid-dissidence. Fake news writers tend to introduce many substitution words with negative connotation, e.g., *маска (mask) – намордник (muzzle)* or *вакцина (vaccine) –жижа (slurry*) to discredit the WHO-approved recommendations for the prevention and treatment of the new virus.

Significant differences in the use of coronavirus-related neologisms and dysphemisms in fake and real news might flag specific linguistic strategies used by Russian fake story providers. However, we suppose that outside the context of the Covid-19 pandemic, the use of these words is just a feature of expressive speech and cannot be considered as a reliable factor in fake news detection.

Unfortunately, new fake stories about Covid-19 and related aspects still emerge. Therefore, future research endeavors can be focused on including a larger corpus in the study by adding new false narratives which have gone viral since

March 2022. We also hope to expand on the observations made in the present paper by discovering linguistic trends on the level of collocations.

REFERENCES

Ahmed, Hadeer. 2017. *Detecting Opinion Spam and Fake News Using n-Gram Analysis and Semantic Similarity*. University of Ahram Canadian.

Ahmed, Hadeer, Issa Traore & Sherif Saad. 2018. Detecting opinion spams and fake news using text classification. *Security and privacy* 1 (1). 1–15. https://doi.org/10.1002/spy2.9

Allcott, Hunt & Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31 (2). 211−236. https://doi.org/10.1257/jep.31.2.211

Al-Salman, Saleh & Ahmad S. Haider. 2021. COVID-19 trending neologisms and word formation processes in English. *Russian Journal of Linguistics* 25 (1). 24–42. https://doi.org/10.22363/2687-0088-2021-25-1-24-42

Baron, Alistair, Paul Rayson & Dawn Elizabeth Archer. 2009. Word frequency and key word statistics in historical corpus linguistics. *Anglistik: International Journal of English Studies* 20 (1). 41–67.

Biber, Douglas & Susan Conrad. 2019. *Register, Genre, and Style*. Cambridge University

Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics: A Practical Guide.* Cambridge University Press. https://doi.org/10.1017/9781316410899.008

Chen, Lian-Ching, Kuei-Hu Chang & Hsiang-Yu Chung. 2020. A novel statistic-based corpus machine processing approach to refine a big textual data: An ESP Case of COVID-19 News Reports. Applied Sciences 10 (16). 5505. https://doi.org/10.3390/app10165505

Christopher, S. Butler & Anne-Marie Simon-Vandenbergen. 2021. Social and physical distance/distancing: A corpus-based analysis of recent changes in usage. *Corpus Pragmat* 5 (4). 427–462. https://doi.org/10.1007/s41701-021-00107-2

Curzan, Anne. 2009. Historical corpus linguistics and evidence of language change. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics: An international handbook*, 1091–1109. De Gruyter Mouton. https://doi.org/10.1515/9783110213881.2.1091

Essam, Bacem A. & Muhammad S. Abdo. 2021. How do Arab tweeters perceive the Covid-19 pandemic? *Journal of Psycholinguistic Research* 50. 507–521. https://doi.org/10.1007/s10936-020-09715-6

Gjylbegaj, Viola. 2018. Fake news in the age of social media. *International E-Journal of Advances in Social Sciences* 4 (11). 383–391. https://doi.org/10.18769/ijasos.455663

Goddard, Cliff & Anna Wierzbicka. 2021. Semantics in the time of coronavirus: "Virus", "bacteria", "germs", "disease" and related concepts. *Russian Journal of Linguistics* 25 (1). 7–23. https://doi.org/10.22363/2687-0088-2021-25-1-7-23

Grieve, Jack & Helena Woodfield. 2023. The Language of fake. *News Series: Elements in Forensic Linguistics,* https://www.cambridge.org/core/elements/language-of-fake-news/7B37014A5C0768AEE806167E8ADD5897. (accessed 11 January 2023).

Habgood-Coote, Joshua. 2019. Stop talking about fake news! *Inquiry* 62. 1033–1065.

Ivanova, Irina. 2020. Pragmatic functions of interrogatives in media texts. *Media Linguistics* 7 (4). 501–515.

Islam, Md Saiful, Tonmoy Sarkar, Sazzad Hossain Khan, Abu-Hena Mostofa Kamal, S M Murshid Hasan, Alamgir Kabir, Dalia Yeasmin, Mohammad Ariful Islam, Kamal Ibne Amin Chowdhury, Kazi Selim Anwar, Abrar Ahmad Chughtai & Holly Seale. 2020. Covid-19–Related infodemic and its impact on public health: A global social media analysis. *American Journal of Tropical Medicine and Hygiene* 103 (4). 1621–1629.

Khan, Ali, Kathryn Brohman & Shamel Addas. 2021. The anatomy of 'fake news': Studying false messages as digital objects. *Journal of Information Technology* 37 (2).

Kopytowska, Monika & Radosław Krakowiak. 2020. Online incivility in times of Covid-19: Social disunity and misperceptions of tourism industry in Poland. *Russian Journal of Linguistics* 24 (4). 743–773. https://doi.org/10.22363/2687-0088-2020-24-4-743-773

Kuzmin, Gleb, Daniil Larionov, Dina Pisarevskaya & Ivan Smirnov. 2020. Fake news detection for the Russian language. In *Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM).* 45–57.

Kytö, Merja. 2010. Data in historical pragmatics. In Jucker Taavitsainen & Irma Taavitsainen (eds.), *Historical pragmatics*. Berlin/New York: Walter de Gruyter Handbooks of Pragmatics https://doi.org/10.1515/9783110214284.2.33

Lun, Wong Wei, Mazura Masture Muhammad, Muhamad Fadzllah Zaini, Rahimy Damit, Carrine Teoh-Ong, Charanjit Kaur Swaran Singh & Norhayati Yusoff. 2022. Analysis of Covid-19 related phrases using corpus-based tools: Dualisms language & technology. *Journal of Positive School Psychology* 6 (3). 5034–5044.

Mahyoob, Mohammad, Jeehaan Algaraady & Musaad Alrahaili. 2021. Linguistic-based detection of fake news in social Media. *International Journal of English Linguistics* 11 (1). 99–109. https://doi.org/10.5539/ijel.v11n1p99

McCulloch, Gretchen. 2019. *Because Internet: Understanding the New Rules of Language*. Riverhead Books.

Monogarova, Alina, Tatiana Shiryaeva & Nadezda Arupova. 2021. The language of Russian fake stories: a corpus-based study of the topical change in the viral disinformation. *Journal of Language and Education* 7 (4). 83–106. https://doi.org/10.17323/jle.2021.13371

Muslimah, Ryza Wahyu. 2020. A corpus-based analysis of critical strategies in Covid-19 corpora. *Journal of Linguistics and Literature* 4 (2). 258–268. https://doi.org/10.33019/lire.v4i2.89

Oehmichen, Axel, Kevin Hua, Julio Amador Diaz Lopez, Miguel Molina-Solana, Juan Gómez-Romero & Yike Guo. 2019. Not All Lies Are Equal. A Study Into the Engineering of Political Misinformation in the 2016 US Presidential Election. *IEEE Access* (99) 1–1. 1–6.

Pavlina, Svetlana. 2022. Pragmatic and stylistic perspectives on British and American COVID-19 cartoons. *Russian Journal of Linguistics* 26 (1). 162–193. https://doi.org/10.22363/2687-0088-27107

Peng, Zhibin & Zhiong Hu. 2022. A bibliometric analysis of linguistic research on COVID-19. *Frontiers in Psychology* 13. https://doi.org/10.3389/fpsyg.2022.1005487

Pisarevskaya, Dina. 2017. Deception detection in news reports in the Russian language: Lexics and discourse. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism.* 74–79.

Ponton, Douglas M. 2021. "Never in my life have I heard such a load of absolute nonsense. Wtf." Political satire on the handling of the COVID-19 crisis. *Russian Journal of Linguistics* 25 (3). 767–788. https://doi.org/10.22363/2687-0088-2021-25-3-767-788

Rashkin, Hannah, Eunsol Choi, Jin Yea Jang, Svitlana Volkova & Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* 2931–2937. https://doi.org/10.18653/v1/D17-1317

Rayson, Paul. 2019. Corpus analysis of key words. In Carol A. Chapelle (ed.), *The encyclopaedia of applied linguistics*, 1–7. Oxford: Wiley-Blackwell.

Rayson, Paul & Roger Garside. 2000. Comparing corpora using frequency profiling. *In The Workshop on Comparing Corpora. Hong Kong, China. Association for Computational Linguistics.* 1–6. https://doi.org/10.3115/1117729.1117730

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford University Press.

Sutu, Rodica Melinda. 2020. Fake news, from social media to television case study of the Romanian presidential elections 2019. *Styles of Communication* 11(2). 81–92.

Tandoc, Edson & Zheng Wei Lim. 2017. Defining "Fake News": A typology of scholarly definitions. *Digital Journalism* 6 (3). 1–17. https://doi.org/10.1080/21670811.2017.1360143

Torabi Asr, Fatemeh & Maite Taboada 2019. Big Data and quality data for fake news and misinformation detection. *Big Data & Society* 6 (1).

Yu, Hangyan, Huiling Lu & Jie Hu. 2021. A corpus-based critical discourse analysis of news reports on the COVID-19 pandemic in China and the UK. *International Journal of English Linguistics* 11 (2). 36. https://doi.org/10.5539/ijel.v11n2p36

Zhang, Xichen & Ali A. Ghorbani. 2020. An overview of online fake news: Characterization, detection, and discussion. *Information processing and management* 57 (2). https://doi.org/10.1016/j.ipm.2019.03.004

Other sources

Beckett, Charlie. 2017. 'Fake news': The best thing that's happened to Journalism at Polis. (http://blogs.lse.ac.uk/polis/2017/03/11/fake-news-thebest-thing-thats-happened-to-journalism/) (accessed 11 January 2023).

How Bill Gates became the voodoo doll of Covid conspiracies (6 June 2020). BBC News. (https://www.bbc.com/news/technology-52833706) (accessed 25 October 2022).

Article history:
Received: 07 March 2023
Accepted: 20 August 2023

Bionotes:

Alina G. MONOGAROVA is Assistant Professor of the English Language and Professional Communication Department at Pyatigorsk State University, Russia. Her research interests embrace corpus linguistics, text mining and text analysis, as well as standardization of developing terminologies.
e-mail: alinach12@yandex.ru
https://orcid.org/0000-0003-4098-0341

Tatyana A. SHIRYAEVA is Professor of Linguistics, Head of the English Language and Professional Communication Department at Pyatigorsk State University, Russia. She is Editor-in-Chief of the research journal *Professional Communication: Top Issues of Linguistics and Teaching Methods*. Her research interests focus on discourse analysis, sociocognitive linguistics with particular emphasis on professional discourse studies, theory and practice of intercultural professional and business communication, English for special purposes, genre analysis and pragmatics. She is author and co-author of over 200 publications. Several research articles were published in high ranking journals, including *Heliyon*, *Humanities and Social Sciences Reviews*, *International Journal of Arabic-English Studies*, *Journal of Language and Education*, among others.
e-mail: shiryaevat@list.ru
https://orcid.org/0000-0001-5508-8407

Elena V. TIKHONOVA is Associate Professor at the Department of Foreign Languages of MGIMO University, Moscow, Russia. She is also Deputy Editor-in-Chief of the *Journal of Language and Education*. Her areas of interest include discourse analysis, sociocognitive linguistics, and psycholinguistics. She conducts research in the field of English for specific purposes, genre analysis, pragmatics, and academic writing. She has authored numerous

articles in high-impact international journals. She is a member and lecturer of the Association of Scientific Editors and Publishers (ASEP).
e-mail: etihonova@gmail.com
https://orcid.org/0000-0001-8252-6150

Сведения об авторах:

Алина Геннадьевна МОНОГАРОВА — доцент кафедры английского языка и профессиональной коммуникации Пятигорского государственного университета. Ее исследовательские интересы включают корпусную лингвистику, анализ текста, стандартизацию терминологий развивающихся сфер.
e-mail: alinach12@yandex.ru
https://orcid.org/0000-0003-4098-0341

Татьяна Александровна ШИРЯЕВА — профессор, заведующая кафедрой английского языка и профессиональной коммуникации Пятигорского государственного университета, главный редактор научно-исследовательского журнала «Профессиональная коммуникация: актуальные вопросы языкознания и методики обучения». Ее научные интересы сосредоточены на дискурс-анализе, социокогнитивной лингвистике, в особенности на исследованиях профессионального дискурса, теории и практики межкультурного профессионального и делового общения, английского языка для специальных целей, жанрового анализа и прагматики. Она является автором и соавтором более 200 публикаций, среди которых статьи в высокорейтинговых журналах, включая *Heliyon*, *Humanities and Social Sciences Reviews*, *International Journal of Arabic-English Studies*, *Journal of Language and Education* и др.
e-mail: shiryaevat@list.ru
https://orcid.org/0000-0001-5508-8407

Елена Викторовна ТИХОНОВА – доцент кафедры иностранных языков МГИМО МИД России. Является заместителем главного редактора международного научно-исследовательского журнала *Journal of Language and Education*. Сфера ее научных интересов – дискурс-анализ, социокогнитивная лингвистика, психолингвистика. Реализует исследования в сфере английского языка для специальных целей, жанрового анализа, прагматики, академического письма. Опубликовала ряд статей в высококвартильных международных журналах. Является членом и лектором Ассоциации научных редакторов и издателей (АНРИ).
e-mail: etihonova@gmail.com
https://orcid.org/0000-0001-8252-6150