National Research University Higher School of Economics
Journal of Language & Education Volume 7, Issue 4, 2021

Monogarova, A., Shiryaeva, T., & Arupova, N. (2021). The Language of Russian Fake Stories: A Corpus-Based Study of the Tpical Change in the Viral Disinformation Spread During the First Year of the Covid-19 Pandemic. *Journal of Language and Education,* 7(4), xxx-xxx. https://doi.org/10.17323/jle.2021.11371

The Language of Russian Fake Stories: A Corpus-Based Study of the Tpical Change in the Viral Disinformation Spread During the First Year of the Covid-19 Pandemic

Alina Monogarova[1], Tatyana Shiryaeva[1], Nadezhda Arupova[2]

[1]Pyatigorsk State University

[2]Moscow State Institute of International Relations (MGIMO University)

Correspondence concerning this article should be addressed to Nadezhda Arupova, Moscow State Institute of International Relations (MGIMO University), 76, Prospect Vernadskogo Moscow, 119454, Russian Federation. E-mail: arupova.n.r@inno.mgimo.ru

The spread of disinformation during the Covid-19 pandemic is largely associated with social media and online messengers. Viral disinformation disseminated in 2020–2021 was related to a wide range of topics that caused panic among people. Many false narratives emerged and attracted public interest over time, which mainly reflected the general public's utmost belief in these topics. Text mining can be used to analyze the frequencies of keywords and topic-related vocabulary in order to track the changing focus of the public concerning online disinformation. In this paper, we present the results of a corpus-based study of Russian viral fake stories circulating during the first year of the Covid-19 pandemic. We propose a method for analyzing the central topics and dynamics of topical change in the context of the Russian Covid-19-fake story. In order to accomplish this objective, we make use of a set of tools to extract keywords, count their frequencies and analyze corresponding contexts. We apply these tools to the compiled specialized diachronic corpus of Russian viral false Covid-19-related stories. The obtained data is evaluated to determine the dynamic of topical shifts by tracking the changes in keyword frequencies as well as the use of other high-frequency corpus words. The findings of the work concerning topical fluctuations in the Russian viral Covid-19 disinformation agenda as well as given explanations for the identified drifts in public interest in the topics during the first year of the pandemic can contribute to developing effective strategies for combating the spread of fakes in the future.

Keywords: disinformation, Covid-19, fake story, infodemic, viral text, specialized corpus, word frequency, topic change

Introduction

The Covid-19 pandemic has proven that online messengers and social networks have a great potential to make disinformation go viral. Huge amounts of unverified materials written on the web about the Covid-19 pandemic during 2020 led to the coinage of a new term "infodemic," meaning "too much information including false or misleading information in digital and physical environments during a disease outbreak"[1]. In Russia, viral texts about Covid-19 and related issues did rounds on social networks and messaging platforms such as WhatsApp, Telegram and Viber. Many of the circulating texts contained false information about the symptoms and treatment of the new virus, the numbers of Covid-19 cases, the state of hospitals and upcoming government-imposed restrictions. Due to overwhelming public interest and trust in such texts, the dissemination of disinformation was criminalized in Russia, and eventually, WhatsApp imposed a strict limit on the number of messages that can be forwarded as a measure to stop the viral spread of disinformation regarding Covid-19.

Malicious use of social media during the Covid-19 pandemic has been instrumental in the creation of misleading news on key topics related to Covid-19. According to a study on virus-related infodemic and its impact on public health (Islam et al. 2020), more than 5,800 people around the world were admitted to hospital and at least 800 people died in 2020 as a result of false information on social media. Since the Covid-19 infodemic evidently poses a real danger to human lives, it is necessary to study the texts containing the most

[1] WHO (2021) Public health research agenda for managing infodemics. World Health Organization.

popular fake stories in order to understand why certain narratives were successful in fueling conspiracy theories and initiating panic among people.

The growing application of corpus-based research (e.g., Webber & Stroud, 2013; Koplenig, 2017; Karjus et al. 2020) can be attributed to the fact that representative corpora are useful in obtaining quantitative data on the units of analysis and answering a number of questions about the texts, storylines and key topics. In this paper, we propose a quantitative method to measure the topical change in the Russian Covid-19 disinformation spread by counting word frequencies in diachronic collections of target data. The diachronic analysis of viral Covid-19-related fake stories spread in Russia during the first year of the pandemic involves identifying key topics of the fake news and tracking topical shifts in public interest over time. According to Karjus et al. (Karjus et al. 2020, 86), such shifts are reflected in a specialized diachronic corpus in changing frequencies of the topic-related vocabulary. Thus, we expect the frequency distribution of the corpus terms to capture the changes in the disinformation agenda over time.

Previous works (Faust, 2018; Pesta et al., 2018; Mariani et al., 2019) based on the keyword research model, also focus on capturing peaks and troughs in diachronic data to track topical changes in a certain field; however, they do not study the factors that determine fluctuations of public interest in different topics. In addition, the above studies aimed at determining topical drifts use specialized corpora containing academic or media texts as target material. Our work is, so far as we are aware, the first occasion on which word frequencies have been used to analyze the changes in the disinformation agenda (namely, the texts that deliberately mislead the reader).

Describing the dynamics of topical change in the disinformation agenda can contribute to the development of strategies to fight the spread of disinformation about Covid-19 and related issues. The derived knowledge on why public interest on some topics is short, while others remain consistently popular among the public may shed light on what factors make certain fake narratives go viral.

The paper is structured as follows: In the Methodology section we introduce the related work in the literature and describe the proposed method. The Results and Discussion sections present the obtained results and analyze the frequencies of keywords and topic-related vocabulary along with the periods of time when different topics on the Russian disinformation agenda gained or lost popularity while depicting topics that retained their popularity throughout the first 13 months of the pandemic.

Methodology

Building Specialized Corpora

This work is inspired by the prior corpus-based research built on the premise that collecting and analyzing large amounts of discourse samples is an effective tool for understanding the way people communicate (e.g., Biber et al., 1998; Stubbs, 2001; McCarthy & Carter, 2001; Budge & Pennings, 2007; Grimmer & Stewart, 2013). When dealing with specialized corpora targeted towards a particular text type, genre or subject field, vocabulary sorted by frequency can provide very illustrative information (Ngula, 2018). Our research model is based on the assumption that analyzing word frequencies within a corpus with a narrowed text focus (Covid-19-related fake stories) can assist in identifying some patterns in the language of Russian viral fake stories.

When people make up fake stories with the intention of promoting disinformation, they alter their language to not only fit the agenda but also play to the audience's feelings (e.g., fears, desires, interests, etc.) in order to draw their attention to the story and make it go viral. Thus, the vocabulary used in such stories may reflect patterns that allow us to draw conclusions about the general linguistic characteristics of fake narratives. Thus, before discussing the used methodology in detail, we should outline the research agenda by answering three main questions. What types of texts are suitable to address the specific objectives of the study? What is the unit of analysis? What type of corpus (raw or annotated) better suits the research goals?

The design of the specialized corpora complies with the standards of a linguistic investigation if collected texts, representing a particular domain of use, are balanced and sampled. Biber and Jones (Biber & Jones, 2009, p.

1288) note that the term "corpus composition" refers to the text categories that are included the corpus structure. Texts selection is based on a sampling method according to which the collected data should represent the target text categories and the texts should be of similar size. Register variation is also an important issue in the framework of this research, as we intend to analyze Covid-19-related text and audio messages containing fake narratives that went viral during the first year of the pandemic in Russia. These texts are inhomogeneous in terms of style and strategic organization, and their authors have used different registers. Therefore, the corpus should be designed to represent all used registers.

In contrast to discourse analysis, which focuses on a detailed discussion of a few texts, corpus studies analyze large volumes of texts, indicating that this method gives high importance to corpus size. Large corpora are used in historical research for tracing lexical and semantic changes. Such studies require a larger number of units (words and collocations) to be processed, and these units may occur in texts with low frequencies. Therefore, statistical procedures should be applied to very large text collections (Leech & Fallon, 1992; Scott, 2001; Baron et al, 2009). The number of texts sufficient enough to analyze narrow-profile subjects depends on a variety of text categories (genre, format, etc.) that characterize a particular discourse. For example, British Academic Written English (BAWE) is a specialized corpus that consists of good-quality student assignments across disciplines (from first-year to master's students), with a total of 2896 independent texts and 6,514,776 words. Air Traffic Control (ATC) corpus is represented by 70 hours of recorded conversations between controllers and several aircraft in three major airports of the United States. These collections are representative of a special-purpose corpus that offer numerous possibilities of examining and interpreting data from particular areas. However, more targeted corpora (e.g., Louvain Corpus of Native English Essays that contain essays written by British pupils and university students with a total of 324,304 words) also provide sufficient evidence for specified studies related to lexicology or grammar (Curzan, 2009).

Corpora structures can also be determined by the type of units chosen for analysis. Stefanowitsch and Gries (2009, p. 933) mention that, for a long time, lexical issues have been the focus of most corpus-based research, which means they were considered the primary unit of investigation in corpus linguistics. Before the first corpora containing grammatical annotation were built, words and word forms were the only available tools for assessing corpora, and therefore, they were of particular interest to the researchers in this field (Stefanowitsch, 2006, p. 62). However, advances in text processing have broadened the research horizons to include higher-order structural units of analysis. According to Biber and Jones (2009, p. 1289), three main units of analysis in modern corpus studies are each occurrence of a linguistic feature, each individual text or the entire corpus. These units are used in different research designs to either compare texts and define the differences between them or to analyze the use of the linguistic structure. The nature of the units of investigation is different since each represents a different level of language. Some of them do not even have quantitative characteristics. For example, while grammar categories are not numeric and cannot be subjected to quantitative analysis, texts have quantitative characteristics, and numerical procedures can be applied to them (Biber & Jones, 2009, p. 1290). In this work, the units of analysis are keywords and topic-related vocabulary of every part of speech regardless grammar category. Therefore, we preprocess the corpus by lemmatizing it. This would ensure that each word is represented in its original form.

The next research question is whether to annotate the corpus or to use a raw collection of texts. Annotated corpora include additional interpretative linguistic information that could easily be separated from the raw corpora in order to avoid causing any loss of information. The type of information that could be used as an annotation depends on the study objective. For example, POS-tagging is considered to be the most common annotation. It refers to tagging words in a corpus in order to indicate the part of speech they belong to (Leech, 2005, p. 17). There is no consensus among scholars on which corpus, raw or annotated, is best suited for analyzing linguistic phenomena. Sinclair (Sinclair, 2005, p. 5) specifically prefers to work with a raw corpus since it is a "pure" experimental material that most accurately reflects the language features a linguistic study can focus on. Some researchers (Leech, 2005; Wilbur et al., 2006; Kim et al., 2008), while discussing annotation issues, have stated that it is a tool for enriching the original corpus through value addition and may be useful not only for the annotator but also for other linguists who may apply it in their work. Meurers (2005, p. 1620) asserts that large corpora containing million tokens or more can only be annotated automatically. Today, however, semi-automatic annotation procedures are applied even for smaller corpora.

Counting frequencies of specific units in diachrony is one of the methods of historical corpus linguistics. 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, p. 41). Outlining possible research directions, Sinclair (Sinclair, 1991, p. 31) mentions that 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. Since this work focuses on tracing the dynamics of topic change in manipulative e-communication during a certain period, we apply the technique of keyword analysis, which is one of the most widely-used methods for determining significant words based on comparing the frequencies of words in a target corpus with frequencies of the same words in a reference corpus (Baron et al., 2009, p. 41). Through quantitative analysis, researchers get a list of words sorted based on their raw or relative frequencies in texts. The results may be unexpected with the flagging of some linguistic trends that are not obvious without statistical data.

Procedure

The overall framework of the proposed methodology can be described in three main steps. First, we build a corpus of Russian viral fake messages (13 sections, one for each month of the first year of the pandemic, including March 2021) and preprocess the corpus by removing punctuation and graphic elements and normalizing all texts to lowercase letters in order to avoid word doubling in the statistics caused by lowercase or uppercase differences. Next, the texts are lemmatized, stopwords are removed from them and absolute synonyms, different forms of the same word or words, are replaced with one lemma. Then, we use a network graph showing topic clusters and term frequencies across all sections of the corpus to track the dynamics of topic change in Russian fake stories being circulated during the first year of the Covid-19 pandemic (March 2020 to March 2021). We also analyze the most common non-Covid-19-related vocabulary and determine contexts surrounding high-frequency words and collocations.

Methods and Instruments

Input Data. Corpus Structure

From the very beginning of the Covid-19 pandemic, users have been posting/reposting on their social media accounts and forwarding to each other via messengers a lot of false content; however, though these texts could be characterized as fake narratives, not all of them went viral. In this section, we introduce the principles of text selection developed to ensure that the compiled corpus is representative and well balanced.

What is a viral fake story? A Covid-19-related fake narrative refers to text containing information about the recent pandemic caused by Covid-19 (and its related aspects) that was officially announced as false by federal agencies, scientific organizations and officials. Such narratives were spread by many users through popular virtual informal communication channels, primarily social media platforms (such as Instagram, Facebook, Vkontakte, Odnoklassniki, etc.) and mobile messengers (such as WhatsApp, Telegram, Viber, etc.) regardless of whether the intention was to deceive or not. Within the framework of this research, a text is considered viral if it has more than 50,000 unique views. Counting the number of reposts and views in messengers is not always possible, as people send texts to personal and group chats, which are not accessible. Therefore, when selecting study materials, we accounted for the number of views for the forwarded texts via open channels, profiles, bots, public chats and groups. The authors of fake content are usually unknown, and the information in their stories is either completely made up or presented in a significantly distorted manner.

Data source. We began gathering fake narratives about Covid-19 from March 2020 when the WHO made an official announcement of the pandemic. The sources of the texts are popular Russian social networks and mobile messengers. When collecting data, viral audio messages were transcribed. However, video fakes were not considered unless they went viral in the form of texts spread on the web. It is difficult to establish the original source of some texts and the date of their first appearance since most of the authors are anonymous and the texts have been reposted many times. In addition, on April 1, 2020, the President of Russia passed a law[2] imposing criminal punishment for spreading disinformation. Many of the texts that had already been

[2] Federal Law of April 1, 2020 N 100-FZ "On Amendments to the Criminal Code of the Russian Federation and Articles 31 and 151 of the Criminal Procedure Code of the Russian Federation". Collected Legislation of the Russian Federation of April 6, 2020 № 14 (Part I) Art. 2030.

added to the corpus were subsequently deleted from the web sources at the request of Rospotrebnadzor (Russian Federal Service for Surveillance on Consumer Rights). Therefore, in the appendix to the compiled corpus, against the deleted texts, we have placed links to the resources dealing with Covid-19 myth debunking since they post original texts. The corpus (in raw and preprocessed versions) was registered with the Russian Federal Service for Intellectual Property as a database[3]. In this work, we have specified the episode number in this database when examples of fake narratives are introduced.

Data verification. Covid-19-related disinformation covers wide-ranging topics (restrictions, lockdown, healthcare advice, vaccination, etc.), and there is a need to prove that all of the considered texts contain false information. For each text added to the corpus, one of the official institutions (health authorities, ministries, and agencies and federal media) has refuted the claim made in the story on their websites or verified pages on social networks. We also used the following resources to fact-check the information from the collected narratives: COVID Infodemic Europe, Poynter's International Fact-Checking Network and Coronavirus Facts Alliance.

Types of false narratives. The corpus of Russian viral fake stories about Covid-19 is stylistically heterogeneous, containing texts across all genres and categories that were popular among Russian users in the first year of the pandemic ("insider" information from reliable sources (officials, doctors, scientists); warnings; guidelines (how to wear a mask, how to behave during restrictions); pseudo-medical advice/ recipes; information from medical personnel working in "the red zone"; fake stories disguised as news or scientific article; fragments of a pseudo-research; conspiracy theories; and fake documents).

Character limit and the number of episodes. The typological diversity of included texts entails setting a maximum limit of 2000 characters and a minimum of 200 characters (an average fake narrative consists of about 600 characters). This is an important step for data balancing since if both ultra-short texts and large narratives are included in the corpus, the raw frequencies of keywords will not give accurate information about the dynamics of topic change and public interest in a certain topic. These limits enable us to eliminate the disproportion between the number of episodes and the number of keywords extracted from each episode (short texts exhibit lower occurrences of keywords, while the same number of large episodes contain a lot of keywords related to the same topic). It must be noted that the distribution of texts by month is uneven. The number of fake Covid-19 stories peaked for the first time in March 2020 (52 episodes) and April 2020 (48 episodes) during the period of self-isolation in Russia. During the summer of 2020, only 66 texts that meet the established research principles went viral on the web. Since the fall of 2020, there has been an increase in the number of viral fake stories, with statistically significant peaks in November 2020 (48 episodes), December 2020 (54 episodes) and March 2021 (54 episodes), after the start of vaccine testing in Russia (see the Table below for more details on episode statistics).

One of the major obstacles to corpus balancing is the so-called "chain fakes," i.e., texts significantly distorted due to paraphrasing and specifications as a result of multiple reposts. Thus, the question arises as to what the original text looked like and when it first appeared. For example, there is a popular fake story of doctors and police officers coming to people's homes and forcing them to get vaccinated («*Если придут врачи с полицейскими. Отказывайтесь от любых тестов на вирус. Потом заставят пройти повторный тест, а он уже покажет наличие вируса. Как вирус подтвердится.* <…>» (Episode #56, April 2020). This story was first posted on Instagram in March 2020 and was refuted by the Ministry of Internal Affairs that stated that the information was false and no vaccine has been found yet. However, the story began circulating in messengers again in January 2021 and looked even more convincing than the original text since vaccination had already started in Russia at that time. In such cases, we add a text next to the earliest recorded date in the corresponding chronological section of the corpus.

Corpus Preprocessing. Lemmatization and Data Cleansing

To enable accurate word frequency analysis and extraction of keywords and topic-related vocabulary, the original text corpus should be preprocessed, i.e., all forms of one word are converted to its base form so they are analyzed as a single item. Components that do not carry additional meaning are eliminated from the texts.

[3] Russian-language bank of electronic texts containing verbal markers of linguistic manipulation. Database #2021621693, registered with the Federal Service for Intellectual Property of the Russian Federation 08/14/2021. Authors: Monogarova, A. G., Bagyan, A. Yu

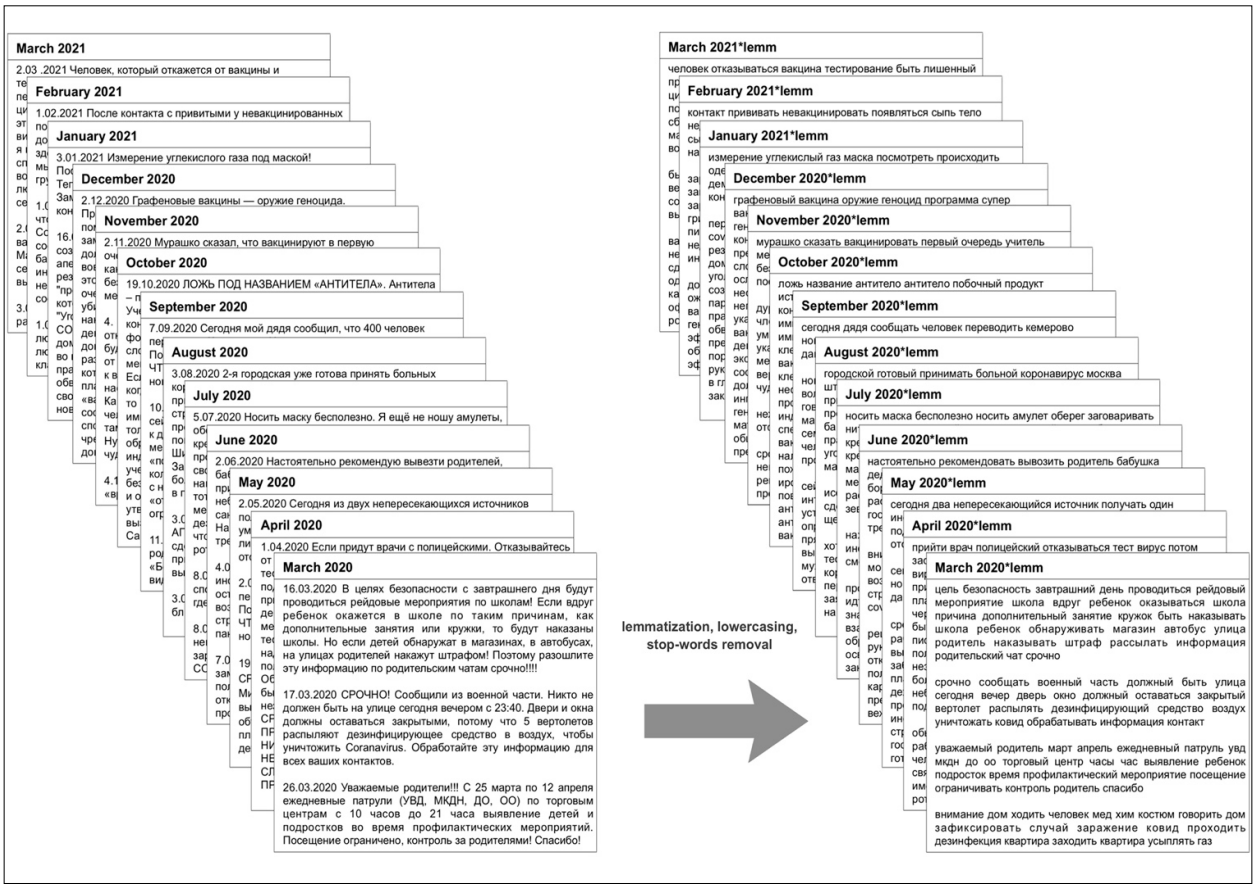
These elements primarily include stop words, absolute synonyms and punctuation marks. Viral Covid-19-related fakes are often characterized by emotional presentation, and many e-texts contain graphic elements (pictograms, logograms, ideograms and smileys). Moreover, uppercase letters are used to draw the audience's attention to particular bits of information.

As the first step to corpora preprocessing, we performed *lemmatization* and *denoising* (lowercasing and the removal of punctuation and graphic elements) using Natural Language Toolkit (NLTK) in Python. After lemmatization, all the words in the corpus are represented with their regularized forms, which helps avoid retaining "multiple terms representing the same word or phrase and thus decrease the vocabulary size" (Sarica et al., 2020). The number "19" was removed from *ковид-19/Covid-19* so that the word frequency counter recognizes each of these terms as a single word form.

Then, we continue to clean the data by removing *stopwords* (commonly used words such as most pronouns, prepositions, conjunctions and quantifiers that do not hold much meaning) from the corpus. These groups of words are statistically significant since they appear in texts multiple times as a part of the sentence structure in order to make narratives coherent, but they give no considerable information about the analyzed bits of discourse. To exclude these words from frequency index, we customize a NLTK pre-existing stopwords list that already contains most Russian function words by expanding it with some common adverbs (*вообще, совсем, просто, еще, только, явно, тоже, настолько, насколько, вовсе, почти, очень*), all forms of possessive and demonstrative pronouns and some interjections that are commonly used in expressive fake narratives (*ну, ага, ого, увы, ура, вау, эх, ой, ох, господи, боже мой*). The overall denoising procedure has been illustrated in Fig. 1.

Figure 1

Corpus preprocessing



Collected samples of discourse contained four groups of words that hampered text normalization since their occurrence in the corpus could distort the results of frequency-based statistical analyses. These groups are as

follows: synonyms that can be substituted in all possible contexts; multi-word terms that are found in fake texts in both full and shortened forms (acronyms, initialisms, etc.); words that have several spelling variations within the corpus; and terms containing numbers or symbols. To avoid any ambiguity in decoding results of frequency analysis, we normalize the texts by *replacing similar words with one lemma*.

First, absolute synonyms are replaced with one semantically equivalent word form. For example, in the case of synonymous words for the coronavirus, (*ковид*, 256 occurrences in the corpus before word replacement), *коронавирус* (159 occurrences), *covid* (94 occurrences), *корона* (48 occurrences), three less common words are filtered out and replaced with the most common form, *ковид*. The word *прививка* (106 occurrences) is substituted with the term *вакцина* (316 occurrences). Likewise, the term *вакцинированный* replaces the word *привитый*. Note that the adjectives *дистанционный* and *удаленный* were not replaced, since, within this discourse, they are not total synonyms; the word *дистанционный* refers to online education, and *удаленный* refers to working in a virtual format.

Most multi-word terms in the corpus are names of organizations and toponyms that are more often used in an abbreviated form (ВОЗ/Всемирная организация здравоохранения, МВД/Министерство внутренних дел, США/Соединенные Штаты Америки/Америка/штаты, КНР/Китайская Народная Республика/Китай/Поднебесная, РФ/Российская Федерация/Россия и т.д). The choice of the substitute depends on the following conditions. If there is a one-word variant of such a term (e.g., *Россия, Италия, Китай*), then all forms in the row are replaced with it. When a term has only multi-word variations (e.g., *полимеразная цепная реакция/ПЦР*), we choose the abbreviation as the substitute form. These will ensure that unit searching and frequency counting are easier as a one-word abbreviation is easier to automatically find in a text than a combination of two or more words.

Different spellings of the same terms are also an obstacle to the analysis of word frequencies as each form has a separate frequency index. This can result in serious ambiguity in the statistics. In different texts across the corpus, the names of vaccines are spelled either in Latin or Cyrillic (*Pfizer/Файзер, AstraZeneca/АстраЗенека, Moderna/Модерна, Sputnik/Спутник)*. In this case, the term with higher frequency was chosen as a replacement. Furthermore, the adjective *коронавирусный* is misspelled (*короновирус*) by the authors of viral stories in 58 episodes. Therefore, the wrong spelling is replaced with the correct one. The term *CDC* appears in fake narratives in three different forms, namely, in Latin, Cyrillic and abbreviations *(CDC/ЦКПЗ/ Центр по контролю и профилактике заболеваний)*. We choose the English abbreviation CDC as a replacement, as it occurs in the corpus (31 occurrences) more often than the other forms.

Most automatic word frequency counters recognize the numbers or symbols in the structure of a term as separate units. For example, in almost all original texts, the term denoting the fifth-generation mobile network is represented by the abbreviated form *5G*. Therefore, the frequency analyzer we use in this work (Voyant Tools) recognizes this compound term as two words. Therefore, we replace it with a Russian word *пятьджи*, which appeared in the original texts only a few times. However, this replacement enables us to receive accurate statistical data.

By employing the word replacement procedure, we compress the vocabulary without losing the word's meaning. In normalized texts, all semantically identical words are replaced with one lemma (with most satisfying the requirements of the chosen analysis toolkit). Thus, all occurrences of these words are attributed to one term providing us a clear picture of topic development and topic change in Russian fake story-making during the first year of the Covid-19 pandemic. Therefore, the material of this study represents a corpus of 13 sections (12 first months of the pandemic + March 2021), 491 viral texts (episodes), with each containing at least 200 and at most 2000 characters. After data preprocessing, the corpus contains 26,964 words, with 16,002 unique word forms (see the table below for more detailed information).

Statistics for the Corpus before and after Denoising and Lemmatization

			Before Denoising and Lemmatization			*After Denoising and Lemmatization*	
Corpus ID	***Number of Viral Fake Episodes***	***Average Words Per Episode***	***Vocabulary Density***	***Total Words***	***Unique Word Forms***	***Total Words***	***Unique Word Forms***
March 2020	52	85.9	0.483	4,470	2,160	2,593	1,269
April 2020	48	80.1	0.513	3,845	1,972	2,373	1,243
May 2020	36	88.1	0.540	3,173	1,712	1,966	1,121
June 2020	22	93.2	0.553	2,052	1,134	1,205	717
July 2020	24	90.3	0.588	2,169	1,275	1,007	804
August 2020	20	87.8	0.591	1,757	1,039	1,014	695
September 2020	25	85.3	0.624	2,134	1,331	1,010	791
October 2020	30	104.5	0.537	3,135	1,682	1,812	1,115
November 2020	48	94.1	0.498	4,517	2,250	2,640	1,817
December 2020	54	110.1	0.481	5,950	2,859	3,604	2,008
January 2021	41	90.8	0.537	3,724	1,998	2,110	1,315
February 2021	37	85.1	0.527	3,151	1,662	2,351	1,201
March 2021	54	94.9	0.483	5,128	2,478	3,275	2,006
Total	**491**			**45,205**	**23,552**	**26,964**	**16,002**

Analyzing the Dynamics of Topic Change using Word Frequencies

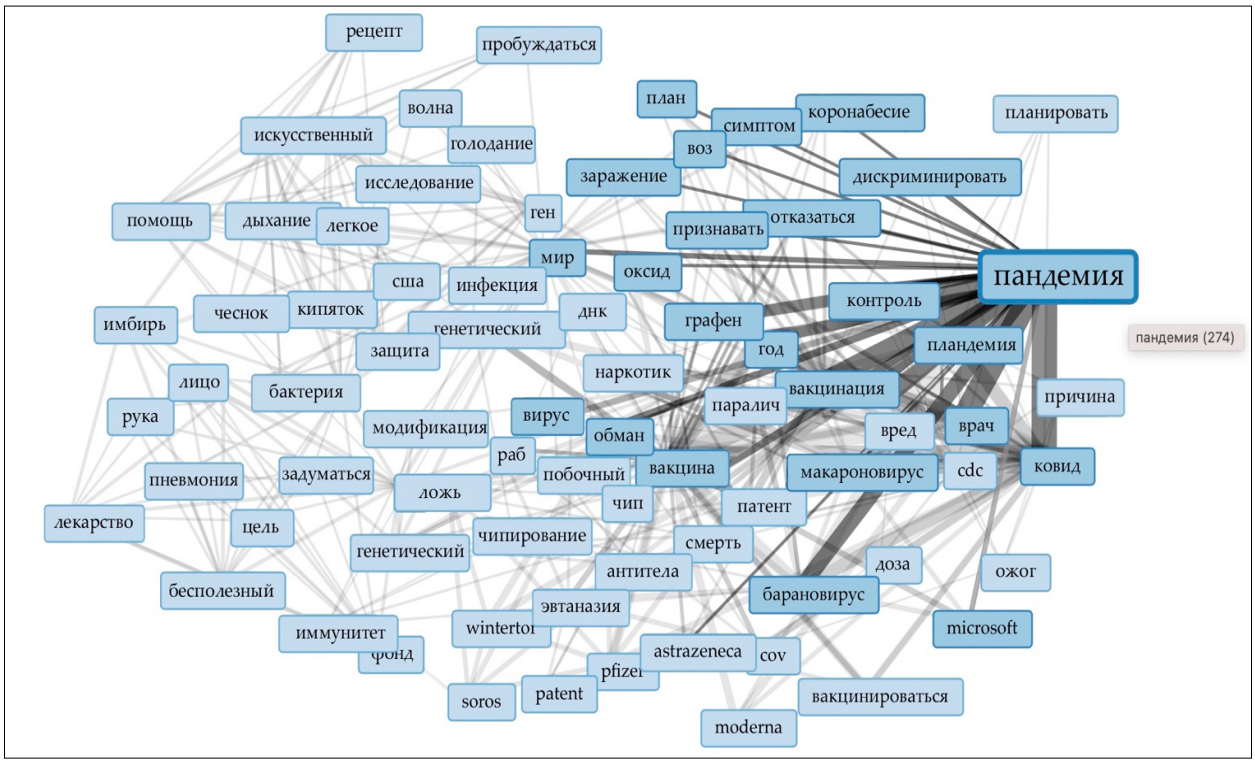
Counting frequencies of words in the corpus and studying the contexts surrounding the keywords and topic-related vocabulary were performed using Voyant Tools, developed by S. Sinclair and G. Rockwell, which is a web-based application for performing text analysis. Preprocessed collection of digital texts is evaluated with tools such as Trends (that depicts the distribution of a word's occurrence across a corpus), Corpus Collocates (that represents keywords and terms occurring in close proximity to them), Collocates Graph (that visualizes the semantic proximity of words within the corpus based on a distribution semantic model) and Contexts (that shows each occurrence of a keyword with a bit of surrounding text) (Sampsel, 2018, p. 153). All of the graphs presented in the Results section were also made using Voyant Tools.

Network Graph and Term Frequencies

When the corpus is uploaded to Voyant Tools, all 13 sections, representing data from March 2020 when the Covid-19 pandemic was officially announced till March 2021, are allocated chronologically. We then generate a network graph from the preprocessed data, where nodes are terms, and edges represent similarities between them. A fragment of the graph is shown in Fig. 2. This is a graphic representation of the higher-frequency terms that appear in close proximity, and we use the clusters of keywords and collocates to extract topics and topic-related vocabulary. Keywords are extracted based on their raw frequencies (the total number of occurrences in the corpus) and collocates are mined according to their frequencies in the context of the related keywords.

Figure 2

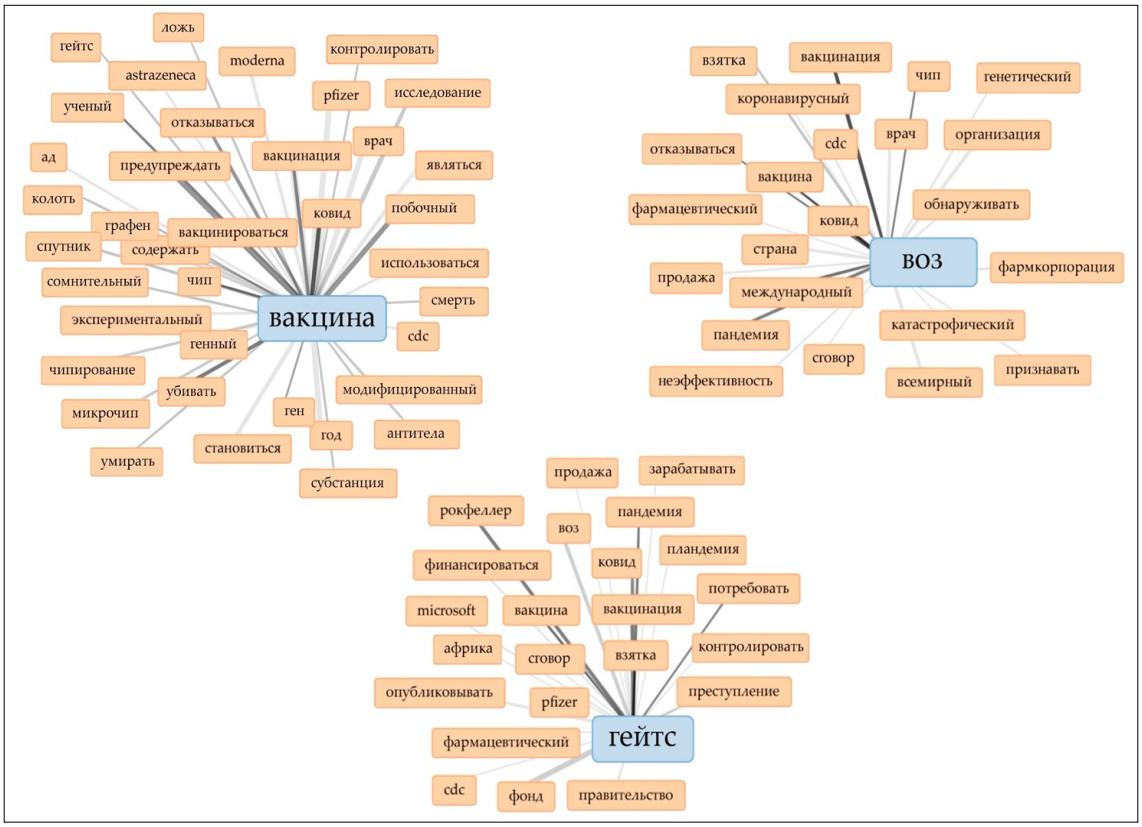
Fragment of the network graph generated from fake Covid-19 narratives



Graph clustering simulates flow within a graph, "promoting flow where the current is strong, and demoting flow where the current is weak" in order to show only the collocates that frequently occur in high proximity to a keyword (Jurisica & Wigle, 2005, p. 122). Clusters represent small topics within a larger set of fake texts and contain the vocabulary used to represent made-up storylines in fake narratives (some cluster structures are illustrated in Fig. 3). Although frequencies of keywords alone allow the tracking of the dynamics of topic change in disinformation agenda, counting the frequency of topic-related vocabulary can provide more illustrative data. The distribution of keyword frequencies depicts an increase or decrease in user interest in the topic over time, while frequencies of the topic-related vocabulary show the attitude of the authors towards a topic. For example, in most Russian mass media texts, the term *вакцина* "vaccine" has the following lexical environment—*ковид, вирус, инфекция, новый, коронавирусный, действовать, эффективность, защищать, антитела, иммунитет, уровень, компонент, близкий, вакцинация, распространение, коллективный* (based on texts in the news about Covid-19 vaccine during September 2020 – Mar 2021, 50 episodes in total). In fake stories, the same term is most often found in connection with the following words: *ковид, ложь, побочный, ученый, контролировать, исследование, cdc, врач, вакцинация, отказываться, предупреждать, ад, колоть, графен, спутник, вакцинироваться, pfizer, astrazeneca, moderna, содержать, гейтс, сомнительный, чип, экспериментальный, генный, чипирование, микрочип, убивать, умирать, ген, становиться, год, субстанция, антитела, модифицированный, смерть, использоваться.*

Figure 3

Topic clusters (a keyword and collocates)



Using cluster data, we compiled a list of key topics that constitute the disinformation agenda in the first year of the pandemic and also extracted the topic-related vocabulary. Then, we applied the Trends Tool to determine the frequency of each keyword in each of the 13 months. A diachronic study of the compiled corpus allowed us to track user interest drifts over these months. The dynamics of increase and decrease of user interest in different Covid-19-related topics are demonstrated in the distribution graphs in the Results section.

N-grams and Surrounding Contexts

The search for the most frequent word combinations in the corpus is based on n-grams extraction (unigrams not considered) for which we made use of the Corpus Collocates tool. In the framework of this research, we refer to an n-gram as a collocation composed of a contiguous (linear) sequence of n-units (two or more terms; bigrams, trigrams, etc.) that occur together more often than expected by chance. The extracted phrases are further arranged into a rank-ordered list, indicating their occurrences throughout the first 13 months of the pandemic.

For interpreting the data received from frequency distributions, we employ the Contexts Tool to determine the context surrounding high-frequency keywords and topic-related vocabulary. The tool helps us select and group false narratives together based on the shared keywords. This allowed us to cluster the narratives into a wide range of scenarios (major fake storylines, false rumors and conspiracies) that misinformed people about different Covid-19-related issues.

Results

A diachronic analysis of the length of the texts showed that, over time, fake narratives became longer and longer. From March 2020 to May 2020, viral fake texts mainly sent out short warnings, news and announcements of up to 500 characters. However, in October 2020, 86% of the texts contained at least 700 characters, with the

average increasing to more than 750 characters in March 2021. We have determined the two most probable reasons for the increase in the length of an average text episode. First, during the first year of the Covid-19 pandemic, Russian users received (via social networks and messengers) a lot of misinformation about Covid-19 and related aspects, and many of these fake stories were repeatedly refuted by the media, Russian authorities and reputable sources in the blogosphere. This naturally increased the level of information filtering by users; in other words, the audience became more skeptical about any Covid-19-related information. Thus, by the end of the year, to convince readers of the veracity of a theory or news, authors of fake stories needed to provide more detailed arguments, and this, we believe, resulted in the increase of average character length. The second assumption is based on a change of a key subject in the disinformation rhetoric. Prior to mid-autumn 2020, fake story creators focused on topics such as quarantine, restrictions, masks and statistics on Covid-19 cases. This type of disinformation was mostly represented in the form of short alarming texts and announcements. However, since the beginning of the winter of 2020, the majority of fake stories revolved around vaccines and their associated dangers. The greater part of these narratives involves argumentation mimicking scientific evidence, numerous statements of pseudoscientists and fake discoveries.

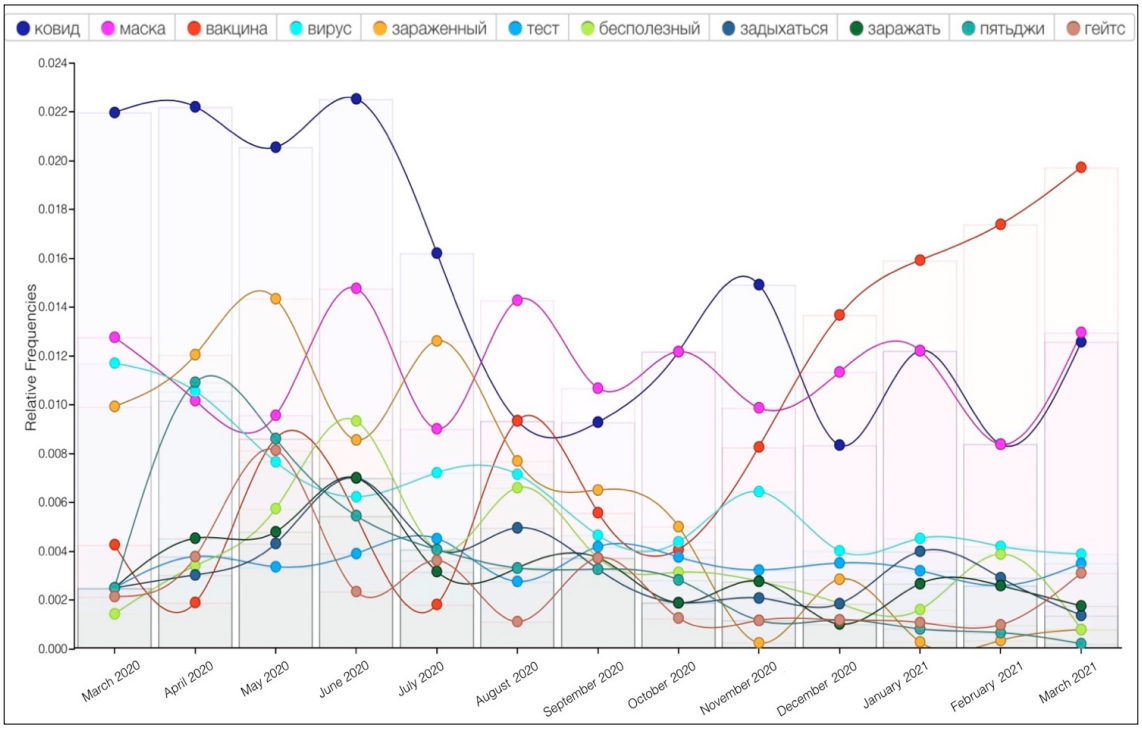
The evolution of the fake news agenda over time is also associated with an unequal ratio of neutral vocabulary and terminology in the corpus. In the first half of the year, terminological density was relatively low (except for keywords such as *ковид (covid), пандемия (pandemic), пневмония (pneumonia),* etc.). However, the density increased significantly from December 2020 to March 2021 since specialized terms from the fields of biology, medicine and physics frequently began occurring in the texts during this period (e.g., *оксид (oxide), графен (graphene) гематоэнцефалический (hematoencephalic), рандомизированный (randomized), валентный (valence),* etc.). Likewise, most of the fake narratives in the second half of 2020 and early 2021 tend to stylistically mimic scientific articles to appear more convincing to readers.

The words with the highest raw frequencies in the corpus are *ковид* (*covid*, 557 occurrences), *маска* (*mask*, 460), *вакцина* (*vaccine*, 422), *вирус* (*virus*, 240), *зараженный* (*infected*, 197), *тест* (*test*, 138), *вакцинация* (*vaccination*, 131), *бесполезный* (*useless*, 121), *врач* (*doctor*, 115), *задыхаться* (*suffocate*, 115), *заражать* (*infec*t, 110), *5G/пятьджи* (110), *скрывать* (*conceal*, 109). Interestingly, some of the top words of 2020[4] (*самоизоляция* (*self-isolation*), *удаленка* (*remote work*), *дистанцирование* (*distancing*), according to the Institute of the Russian Language. A.S. Pushkin, appear in the corpus only a few times. Moreover, some of the main Russian neologisms of 2020, *ковидарность, коронакризис, зумиться, карантиниться* (Ivanenko & Zhuravleva, 2020, p. 66), are not present in the corpus at all. Since the word *ковид* (*covid*) is a key term outlining a range of topics related to the pandemic, it appeared in almost all episodes, and therefore, we will not consider its frequency separately within the analysis of topic change. However, it must be noted that for the first months of the pandemic in Russia, *ковид* ranks first in the frequency list (March 2020 – 62 occurrences, April 2020 – 59 occurrences and May 2020 – 43 occurrences). This is due to the extreme demand for information about this new virus, as well as its nature, origin and symptoms. Later, the public interest switched from Covid-19 as an independent topic to other Covid-19-related issues. Moreover, starting from August 2020, there has been a decrease in the frequency of *ковид* in the corpus (August 2020 – 17 occurrences, September 2020 – 20 occurrences and October 2020 – 39 occurrences). August and September show the growing interest in the term *маска* (mask) (August 2020 – 26 occurrences and September 2020 – 23 occurrences), and after December 2020, the term *вакцина* (*vaccine*) dominated the fake news agenda (December 2020 – 82 occurrences, January 2021 – 60 occurrences, February 2021 – 54 occurrences and March 2021 – 102 occurrences). (Fig. 4 and other graphs show the relative frequencies of the terms, while the text presents raw frequencies.)

[4] Osadchy, M.A. (2021). Results of the study "Word of the Year" by State Institute of the Russian Language named after A.S. Pushkin. https://www.pushkin.institute/news/detail.php?ID=27341

Figure 4

Frequencies of 11 most common terms across the corpus



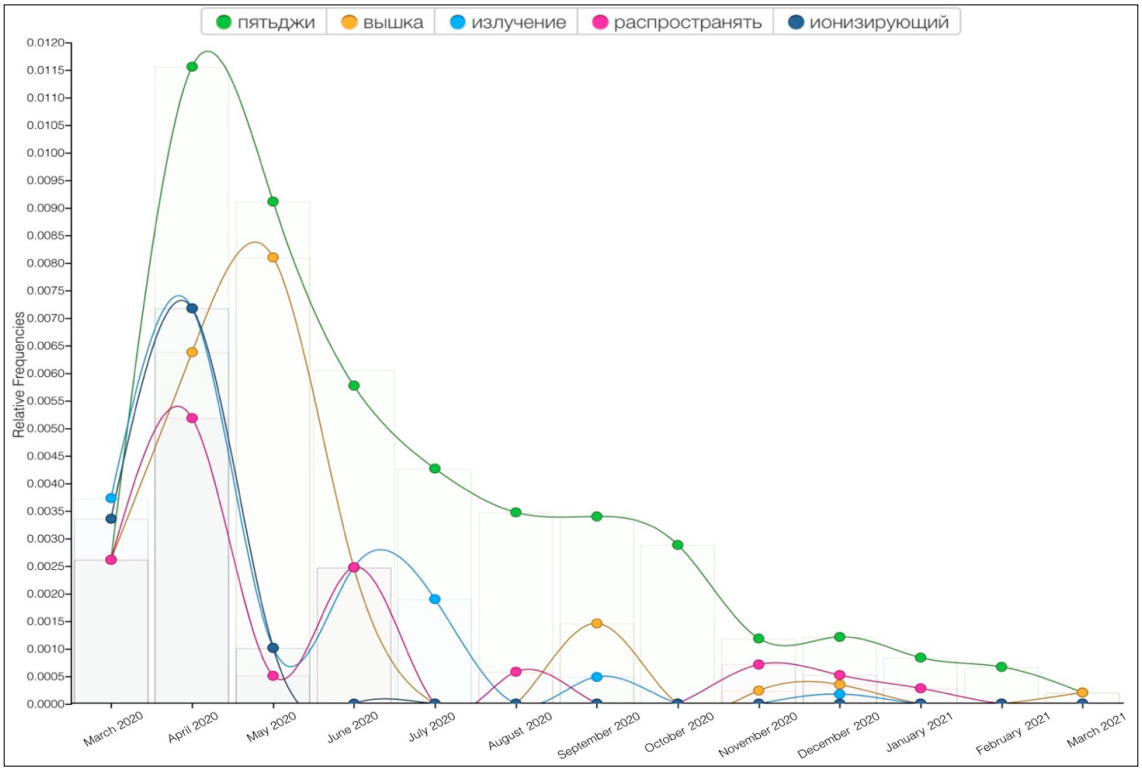
A sharp rise followed by a similarly sharp drop in the frequency of some keywords and topic-related vocabulary during the year suggests that user interest in the corresponding topics only *boomed for a short period of time*. Most of these topics were essentially conspiracy theories that appear, rapidly gaining popularity due to their claim of revealing "secret information," and then fade away as quickly along with public interest. For example, the conspiracy theory about *5G* (*пятьжди*) causing Covid-19 was popular from April to May 2020 (*5G/пятьджи*, April 2020 – 29 occurrences and May 2020 – 18 occurrences). Later, after numerous public refutations from experts, the interest of the Russian public in this topic declined (Fig. 5). In July 2020, the term 5G was mentioned only nine times, and in December 2020, the number of occurrences decreased significantly to four as compared to the April peak (of seven occurrences). This number went down further, with only two occurrences in February 2021. On the network graph, the keyword *5G/пятьджи* is found in a cluster with the topic-related vocabulary, *вышка, сеть, излучение, распространять, ионизирующий, радиочастотный, контроль, ковид, контролировать, управлять, вакцина, тест, китай, смертельный, подчинять, сгорать, электромагнитный*.

Top fake stories about Covid-19 and 5G

1. 5G towers spread Covid-19 infection: *«Карантин – обман народа. Все придумано для внедрения цифровизации общества. Коронавирус распространяется из-за 5G. Первый город, где появилась вышка 5G – китайский Ухань, а после тестирования там новой технологии организм человека дал иммунный ответ, что привело к коронавирусу»*. (Episode #84, April 2020)
2. Authorities/some "secret organizations" use 5G waves to control people through PCR tests and vaccines: *«ВНИМАНИЕ! ПОЧЕМУ НЕЛЬЗЯ ДЕЛАТЬ ПЦР-ТЕСТЫ! В Германии врач провел под микроскопом исследования теста ПЦР на Covid-19. И обнаружил на кончиках тестов, металлические скобы, которые реагируют на волны 5G. <…> Их таким образом вводят далеко в нос. Скобы прикрепляются в носу и реагируют на сигналы 5G. А связи с тем, что тесты заставляют делать по любой причине. Но такие скобы могут накопиться, и последствия управления волнами 5G будут на лицо»*. (Episode #42, March 2020)

Figure 5

Frequencies of the keyword "пятьджи" (5G) and topic-related vocabulary



Another series of fake texts based on the chemtrail conspiracy theory interested many Russians in November 2020 (*химтрейл*, November – 42 occurrences). According to chemtrail theorists, some organizations sprayed chemicals from passenger planes and infected the population with Covid-19. However, February 2021 shows a sharp decline in the frequency of *химтрейл*, with the term being mentioned only three times and all within one episode. The increase and decrease in the popularity of this term and topic-related vocabulary are shown in Fig. 6. The terms in close proximity to the keyword *химтрейл* in the corpus are *химикат, геноцид, распылять, депопуляция, распространять, белый, ковид, самолет, след, дорога, отравлять, лететь, высота*.

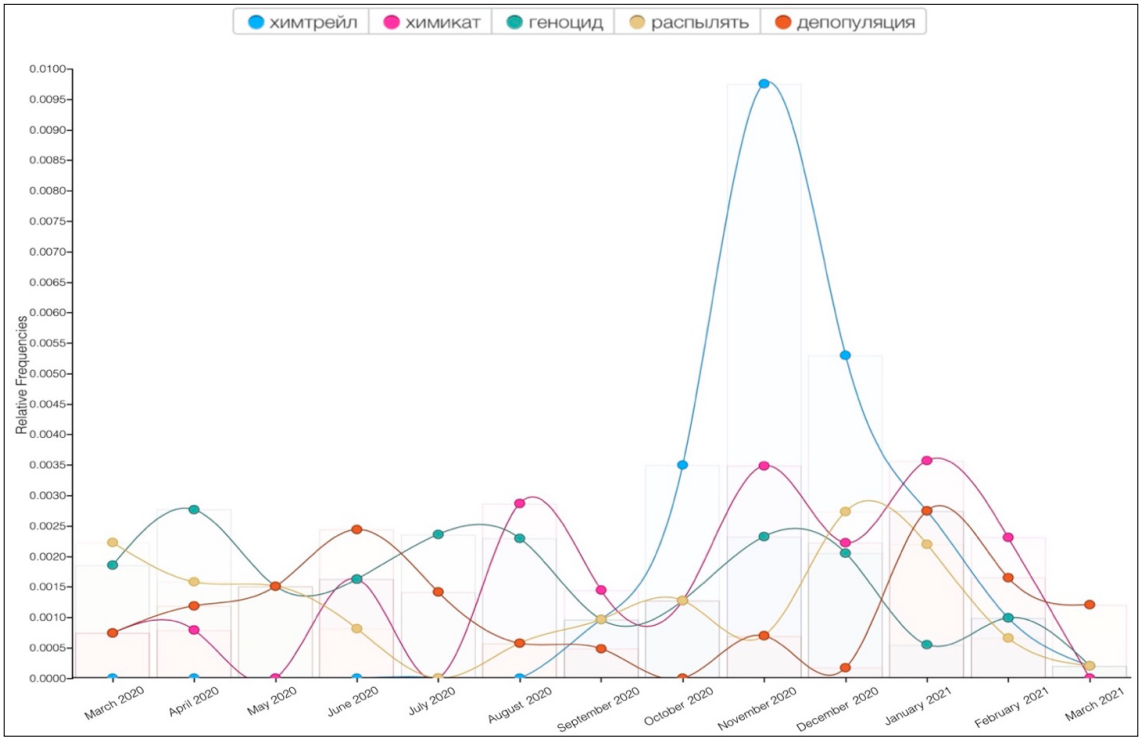
We indicated one popular fake storyline that was represented in various forms in 15 viral episodes about chemtrails. All these episodes were united by the idea that Covid-19 victims were poisoned by chemicals sprayed from the sky. We illustrate this fake story with the most popular text: *«То, что нам выдают за ковид в тяжелых случаях, когда заражены легкие, на самом деле химический ожог легких или химический пневмонит, начинается с першения в горле и кашля. Это реакция дыхательной системы на химическое отравление. Нас травят с воздуха химтрейлами, и это уже не секрет, видимо эффективней в разы, если еще вышки 5G настроить на определенную частоту»*. (Episode #81, November 2020)

User interest in a number of topics throughout the year had a wave-like character. These topics include misinformation on restrictions during the self-isolation period, fines and passes required to move freely around the city/region/country, etc. Although no official quarantine (карантин) was announced in Russia (only a mandatory self-isolation regime), many Russians used this term to refer to restrictive measures. Fake narratives associated with quarantine and related issues occupied one of the leading positions among misleading viral texts from March to May 2020 when the restrictions were lifted in Russia (карантин, March 2020 – 10 occurrences, April 2020 – 14 occurrences and May 2020 – 7 occurrences). Then, the term карантин practically disappeared from the agenda, only occurring nine times in a few fake announcements throughout the summer of 2020. Later, however, due to the spread of rumors about new lockdowns around the world, false warnings about upcoming quarantine measures in Moscow and some large regions in Russia began to be shared through messengers. The revival of this topic can be traced by the high frequencies of the keyword карантин in

September 2020 (17 occurrences) and November 2020 (12 occurrences). We also indicated the increasing frequencies of the topic-related vocabulary (локдаун, штраф, пропуск, протокол, полицейский, ковид, разрешение, запрет, незаконный, италия, необходимость, первый, день, отказаться, право, комендантский, час, сажать) during this period (Fig. 7).

Figure 6

Frequencies of the keyword "химтрейл" (chemtrail) and topic-related vocabulary

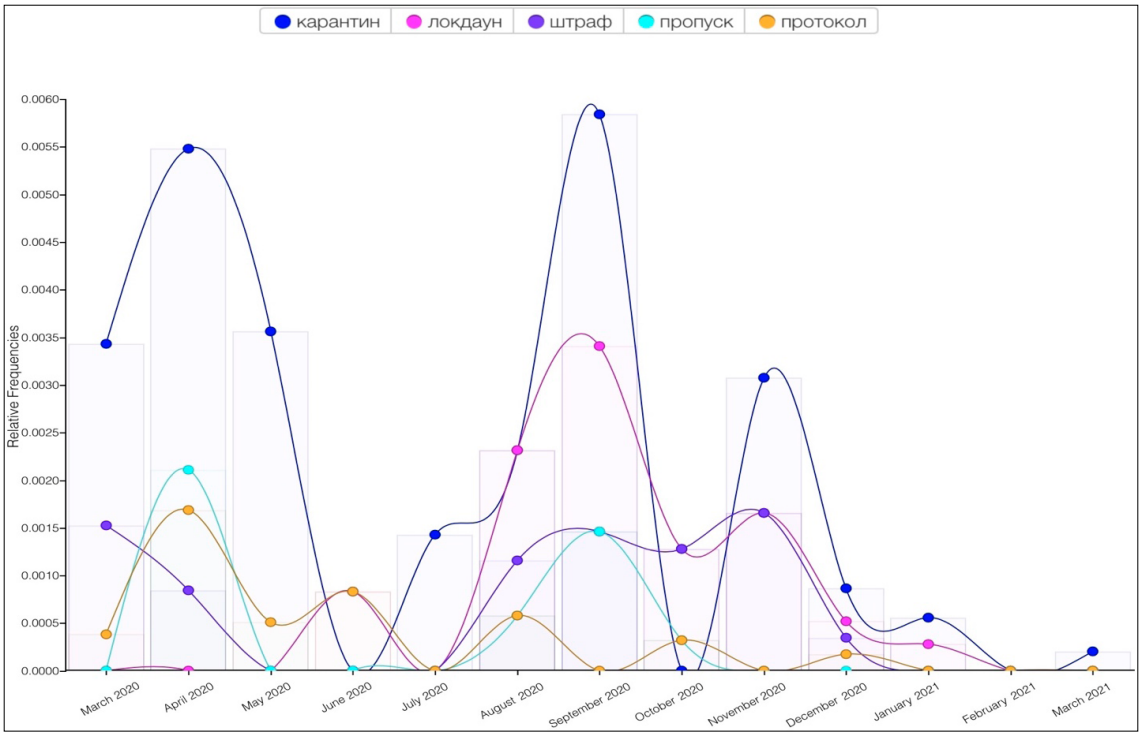


Top fake announcements about quarantine and lockdown from "officials" and "insiders"

1. Disinformation about restrictions/passes/freedom of movement during the period of self-isolation: «*С понедельника 13 апреля 2020 года спецпропуска старого образца прекращают свое действие и тем, кому необходимо в условиях коронавируса передвигаться пешком или на автомобиле по городу, необходимо получить новый пропуск.* <…> *Оформить пропуск можно в отдельном разделе на сайте мэрии – сказать заявку и получить заполненный пропуск на свою электронную почту*». (Episode #64, April 2020)
2. The new quarantine in Russia/around the world will last from three to six months: «*Инсайд: Германию сажают на цепь с 30 августа. 31 июля немецкий сайт liebeisstleben.de сообщает, что через месяц, с 30 августа 2020 года в Германии начинаются повторные карантинные блокировки, которые продлятся от ТРЕХ до ШЕСТИ месяцев. То есть до 1 марта 2021-го года*». (Episode #167, July 2020)
3. Quarantine and other restrictive measures are illegal: «*Апелляционный суд Португалии признал ПЦР-тестирование фейком, а карантин незаконным. И вот теперь уже доказанный на практике факт, что ПЦР-тестирование никак не годится для выявления каких бы то ни было инфекций, в частности ковид-19* <…>». (Episode #287, November 2020)

Figure 7

Frequencies of the keyword "карантин" (quarantine) and topic-related vocabulary



Some topics, on the contrary, aroused *stable interest* from the public throughout the first year of the pandemic. These topics are primarily part of fake narratives that provide "arguments" against using essential matters of the new COVID era, including masks and coronavirus viral and antibody tests (164 out of 491 episodes). A significant number of occurrences of the term *маска* (*mask*) was first witnessed in March 2020 (36 occurrences) after the announcement of the compulsory use of masks in all Russian regions. These guidelines introduced a mask regime in Moscow on May 12, 2020, which resulted in a prolonged boom in fake stories about protective face masks, lasting throughout the next 10 months of the pandemic (Fig. 8). The last month considered in this research is March 2021, and disinformation about protective masks was still hugely popular on the web. More and more erroneous theories were arriving throughout these months. For example, summer fake stories were more focused on numerous cases of the development of heart disease and hypoxia (a condition caused by oxygen deprivation) allegedly from wearing masks. This trend can be reflected in the high frequencies of the term *задыхаться* (*suffocate*) (spring 2020 – 24 occurrences and summer 2020 – 27 occurrences). The arrival of the new trend in misinformation about face masks was marked by growing frequencies of the term *червь* (*worm*) that often occurred in the same contexts as the term *маска* (*mask*) (*червь*, January 2021 – 25 occurrences and March 2021 – 23 occurrences). After analyzing this term in original texts using the Contexts Tool, we found out that its frequent use during that period resulted from the spread of a false story about worms being placed on face masks by some pharma companies to harm public health. The typical narrative is the presentation of the results of an alleged experiment of examining a mask under a microscope and discovering thousands of "moving worms" on it.

Top misleading narratives about protective face masks

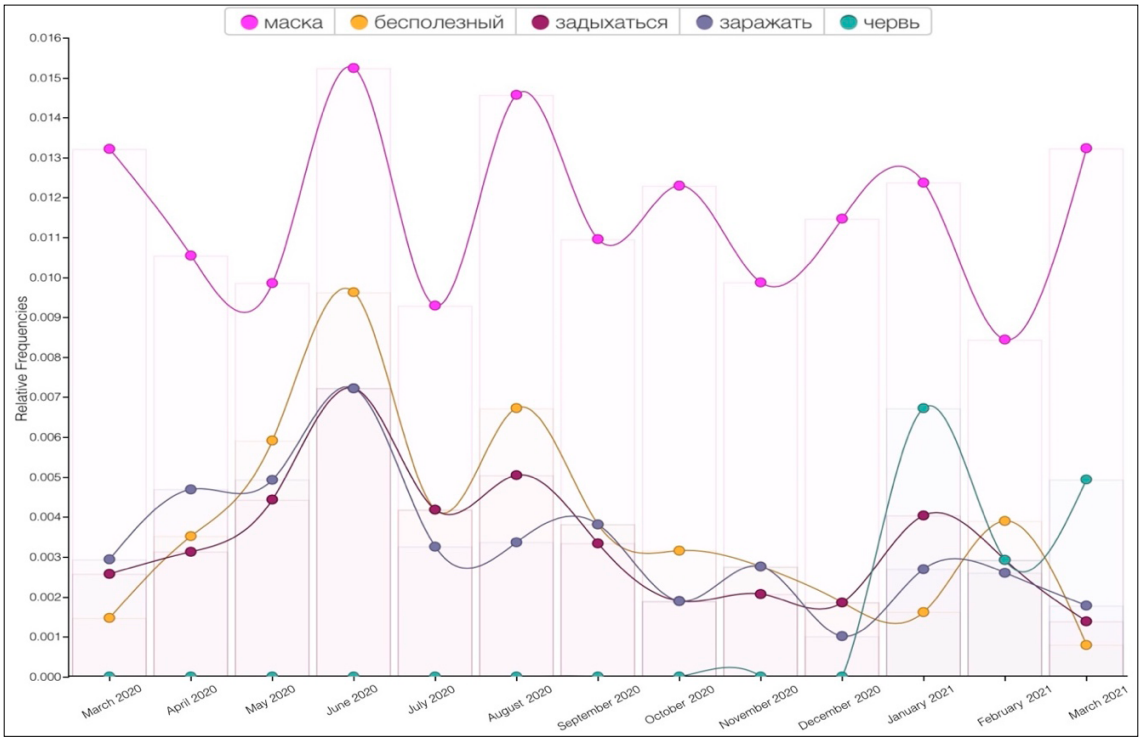
1. People who wear face masks are more likely to get infected by Covid-19 and other diseases: «*ВРЕД МАСОК И ПЕРЧАТОК. ПАМЯТКА Копируйте и распространяйте! Маски нужны для того, чтобы люди засорили себе легкие! Им нужны больные, а маска будет очень скоро провоцировать инфекционные заболевания, в том числе и туберкулез!* <…> *НАС ТРАВЯТ!!!*» (Episode #130, June 2020)
2. Some people, allegedly representatives of a political party, are distributing free face masks soaked in chemicals/drugs/substances infected with Covid-19: «*Новый вид криминального ПРЕСТУПЛЕНИЯ сейчас.*

Ходят «волонтеры» от различных партий… от двери к двери, раздавая маски. Они говорят, что это инициатива местных властей. Просят вас «примерить» маску… если «подойдет» то «дадут» 3–5 штук… в зависимости от количества членов семьи… МАСКА ПРОПИТАНА химикатом с наркотическим действием… Человек просто «отключается» (Episode #194, September 2020)

3. Stories about people who died or were injured as a result of wearing a face mask; for example: a) «<…> *Установлено, что ношение маски №95 часами может СНИЗИТЬ ОКСИГЕНАЦИЮ крови на 20%, что приводит к ПОТЕРЕ СОЗНАНИЯ. Именно эта ситуация случилась с несчастным парнем, который ездил на своей машине в маске №95. В итоге он потерял сознание, попал в ДТП и получил травмы*». (Episode #139, June 2020); b) «<…> *В Германии в результате ношения маски 13-летняя девочка потеряла сознания в школьном автобусе и вскоре умерла в больнице! Ребенку просто ЗАПРЕТИЛИ СНИМАТЬ МАСКУ. Сколько еще это будет продолжаться?!*» (Episode #233, October 2020)
4. Viruses and moving worms were found on face masks: «*Мы положили маску на горячую воду. Видно, как там что-то двигается.* <…> *Это нереально, вот такие черви в этих масках*» (Episode #490, March 2021)

Figure 8

Frequencies of the keyword "маска" (mask) and topic-related vocabulary



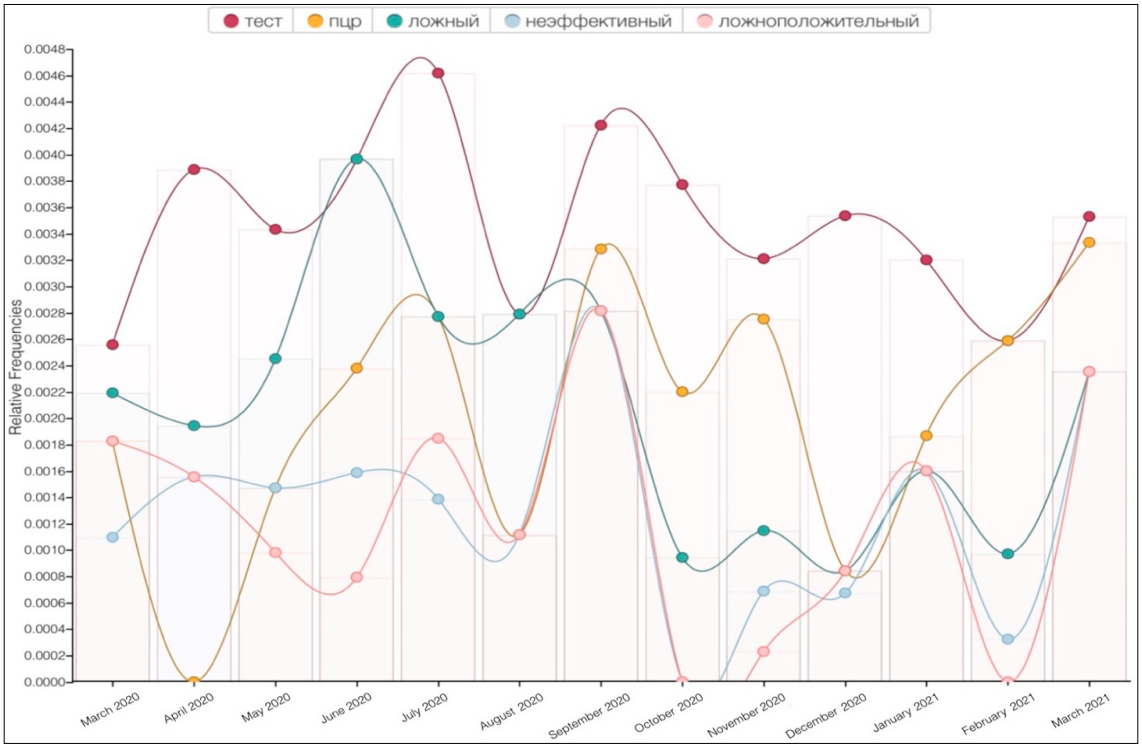
PCR tests, as the most popular method for detecting the virus, have also been the subject of numerous "truth-revealing" stories throughout the year. Every month (albeit in different numbers), new stories about the dangers of PCR testing were going viral. The term *тест* (*test*) occupied the 6th rank in the corpus, with higher frequencies in July 2020 (13 occurrences) and September 2020 (10 occurrences) and the lowest frequencies in August 2020 (6 occurrences) and February 2021 (5 occurrences). Despite the inconsistencies in the number of occurrences, the topic remained popular throughout the study period (Fig. 9). In most viral fake stories, PCR tests are defined as "useless sticks" that are incapable of distinguishing the flu virus from the coronavirus. The adjectives *бесполезный* (*useless*) and *неэффективный* (*ineffective)* occur 28 and 19 times, respectively, in high proximity to the keyword *тест* (*test*). The words *ложный* (*false*) and *ложноположительный* (*false positive*) also appear 22 and 18 times in the same contexts with the keyword, respectively, as part of the arguments supporting the idea that PCR tests are constructed to indicate positive results in order to increase the number of infection cases. Other words found in the closest proximity to the term *тест* (*test*) in the corpus include *пцр, антитела, вред, положительный, опасность, зараженный, опасный* and *отказываться*.

Top fake stories about Covid-19 antibody and viral tests

1. The real purpose of PCR testing is to destroy people's immune defenses: «*Эта информация исходит от моих коллег-врачей. Вред ПЦР-тестирования. Доступ к мозгу осуществляется через нос. Тест имеет другую цель. Введения тестового стержня глубоко в нос вызывает повреждение гематоэнцефалического барьера и даже эндокринных желез. Цель этого глубокого нарушения состоит в том, чтобы сломать барьер и создать вход в мозг для каждой инфекции. <…> Это НЕ от ковида. Это посягательство опасно!*» (Episode #41, March 2020)
2. PCR tests are ineffective in diagnosing Covid-19; for example: a) The tests give the flu for Covid-19: «*Тестирование на коронавирус – это глобальный фейк. Тест НЕ СПОСОБЕН показать, что человек — носитель именно ковида-19, так как содержит лишь три гена-маркера. <…> По сути гены-мишени теста ПЦР не являются специфичными для Covid-19, и поэтому подобный тест не способен определить, что человек является носителем именно SARS-CoV-2*». (Episode #165, July 2020); b) Lemon/coca-cola/ chicken fillet tested positive for Covid-19: «*Посмотрите! Тест на макаронавирус у лимона положительный!!!!! Экспресс-тесты на Covid-19 дают положительный результат, если просто капнуть на них сок лимона. <…>*». (Episode #264, November 2020)
3. The tests infect people with prions/Morgellons disease: «*<…> Тест доставляет прионы прямо в цель. Смертность – 100% в течение 10 лет. Никто не проверяет тесты, и вообще никогда никто не проверяет заражение прионами. Вы знаете, что такое прионы? Наверняка – нет. <…>*». (Episode #465, March 2021)

Figure 9

Frequencies of the keyword "тест" (test) and topic-related vocabulary



Other topics popular among Russian Internet users throughout the first year of the Covid-19 pandemic are pseudo-medical advice and homemade medication. The keywords *рецепт* (*recipe*) (26 occurrences), *лекарство* (*medication*) (25 occurrences) and *совет* (*advice*) (17 occurrences) are found on the network graph, forming intersecting relations with the following topic-related vocabulary: *дыхание, дыхательный, гимнастика, чеснок, имбирь, кипяток, глотка, лимон, перец, бикарбонат, израиль, японский, ученый, исследование, сода, луковица, лук, проверять, больница, проверенный, статья, нагревать, принимать, день, тест, подавлять, блокировать, лечить, уходить*. Malicious narratives describing folk recipes against Covid-19 can be found in all 13 sections

of the corpus (i.e., in each month). Interestingly, in March 2020, April 2020 and May 2020, there was an increased interest in recipes and recommendations, which allegedly come from foreign (non-Russian) doctors. The most common digrams collected from the corpus in a close proximity to the keywords such as *рецепт (recipe), совет (advice)* and *рекомендация (recommendation)* during this period are collocations noun + adjective, indicating the origin of the expert who gave advice; for example, *японские ученые* (*Japanese scientists*, 4 occurrences), *тайваньские ученые* (*Taiwanese scientists*, 2 occurrences), *израильские ученые* (Israeli scientists, 2 occurrences), *израильский рецепт* (Israeli recipe, 2 occurrences). However, after September 2020, this trend gradually faded giving way to another—anonyms recommendations and advice from some unnamed "experts" and "doctors."

Top pseudo-health-care advice and homemade recipes for treating Covid-19

1. Home testing for Covid-19: «<…> *Тайваньские эксперты советуют простую самопроверку на коронавирус, которую мы можем проводить каждое утро: сделайте глубокий вдох и задержите дыхание более, чем на 10 секунд. Если вы успешно завершите его, не кашляя, не испытывая дискомфорта, заложенности, стеснения, и т.п., то это доказывает, что в легких нет фиброза, и это указывает на отсутствие инфекции*». (Episode #28, March 2020)
2. Homemade "interferon": «*Домашний интерферон – защита от короны. Алгоритм приготовления домашнего интерферона: взять 2 сырых белка от куриных домашних яиц в эмалированную узкую кастрюлю, добавить сок половины лимона* <…>». (Episode #108, May 2020)
3. Pepper / onion / garlic / ginger / lemon / soda / boiling water as a treatment for Covid-19. For example: a) «*И еще один СЕРЬЕЗНЫЙ ПРЕВОСХОДНЫЙ СОВЕТ от японских докторов, лечащих случаи Covid-19: вы должны постоянно обеспечивать увлажнение своего рта и горла, не давая им пересыхать.* <…> *питье воды или других жидкостей смоет его через пищевод в желудок. И уже в животе … ваша желудочная кислота убьет вирус». (Episode#20, March 2020); b) «ИЗРАИЛЬСКИЙ РЕЦЕПТ ПРОТИВ КОВИДА. В Израиле не смертей о Covid-19. Было найдено лекарство от вируса COVID19 или способ его устранения. Рецепт простой: лимон, бикарбонат. Смешивайте и пейте как горячий чай каждый день,* <…>». *(Episode #92, April 2020); c) «У КОРОНАВИРУСА аллергия на ИМБИРЬ!!! К таким заключениям пришли специалисты. Что необходимо для того, чтобы не дать вирусу проникнуть и «завладеть» вами? 3 раза в день ЖЕВАТЬ ИМБИРЬ.* <…>». (Episode#81, October 2020)

The analysis of the frequency distribution of the corpus vocabulary also revealed that some Covid-19-related topics started growing rapidly as a result of the emergence of new realities. For example, fake stories about vaccines, green passes and mandatory vaccination exploded at the end of 2020 and continued to gain popularity among Russian users at the end of the study period (March 2021).

Although the keyword вакцина (vaccine) ranked third in the corpus with 422 occurrences throughout the year, vaccines practically remained off the disinformation agenda until November 2020. Vaccination was critically discussed in general, as Covid-19 vaccines had not yet been tested and were undergoing development. However, vaccines (вакцина) became a key subject of fake narratives in December 2020 after the start of vaccination in large Russian cities (вакцина, December 2020 – 82 occurrences, January 2021 – 60 occurrences, February 2021 – 54 occurrences and March 2021 – 102 occurrences) (Fig. 10). Recent studies published on the impact of official refutations on the spread of misinformation found that refutations reduce demand and sharing of fake stories (MacFarlane et al 2021: 248). Thus, the promotion of revealing argumentation against fake stories through reputable sources can reduce user interest in that topic. Our data show that although conspiracy theories about vaccines have remained steadily popular since the appearance of the first fake story on this topic, viral texts about vaccines did not "live" long, as every month a new theory replaced the previous one. The analysis of the corpus show that argumentation against Covid-19 vaccines is based on the following three false ideas. In December, users actively shared texts about vaccination modifying modify people's genes, affecting their heredity aspect (ген (genes), December 2020 – 16 occurrences, модифицированный (modified) – 6 occurrences). In January, arguments against vaccines were based on alleged huge amounts of graphene in vaccines (графен (graphene), January 2021 – 18 occurrences). In February and March 2021, the word чип (chip/microchip) appeared 10 and 22 times, respectively, in close proximity to the keyword вакцина (vaccine) in fake stories claiming that vaccines were being used by the WHO and world elites to control the population by implanting microchips in people's bodies.

Top fake stories about Covid-19 vaccines

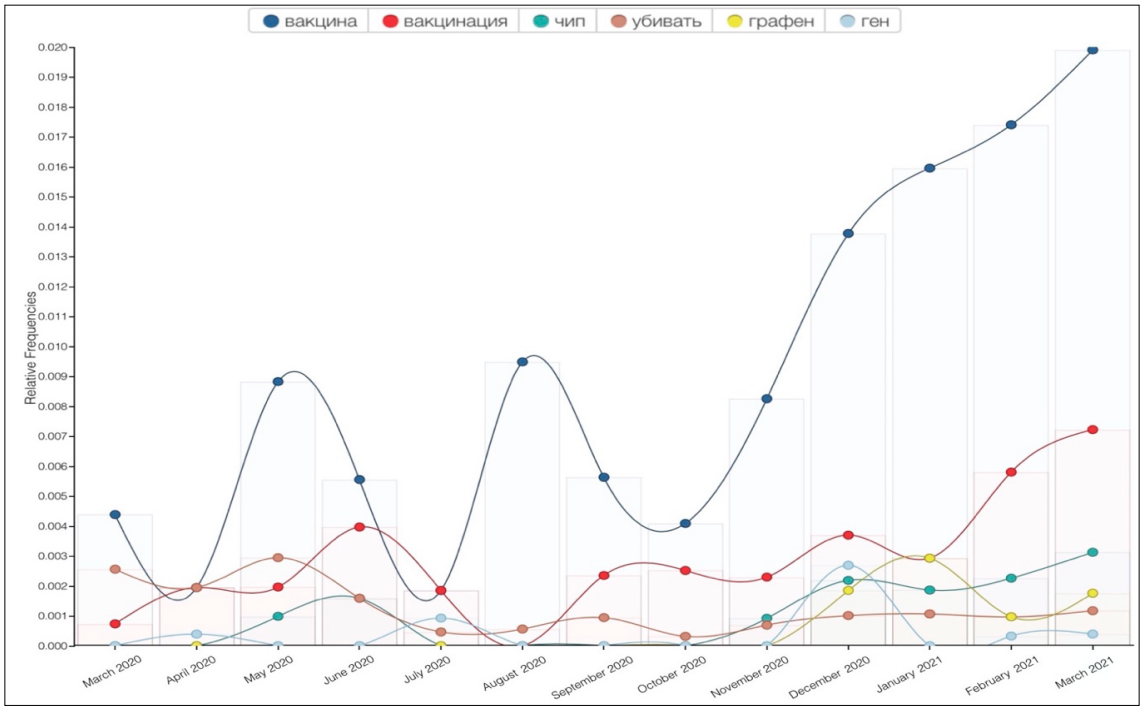
1. Hazardous toxic substances / abortive material / graphene / luciferin are secretly placed in vaccines: «*Сенсация! Испанские исследователи обнаружили в электронный микроскоп, что вакцина Pfizer содержит 99% оксида графена и практически больше ничего!* <…> *Этот продукт не был разработан, чтобы избежать инфекции, вызванной вирусом. Истинная цель этого продукта остаётся скрытой. Токсичность оксида графена - достаточная причина, чтобы остановить глобальную программу вакцинации*». (Episode #303, December 2020)

2. Vaccines are a means of inserting microchips in people and establishing control over them: «*Наночастицы, описанные в патенте Microsoft (патент США WO 2020/060606 A1), являются датчиками, которые должны быть рассеяны в организме вакцинированного человека, чтобы иметь возможность обнаружить его. Добавленные во флакон с вакциной, они вводятся в организм вместе с прививкой в момент вакцинации. Как только они попадут в организм, от них невозможно избавиться, в отличие от подкожного цифрового следящего микрочипа. С этого момента привитых людей можно будет обнаружить по мобильному телефону, находящемуся поблизости*. <…>». (Episode #371, January 2021)

3. Stories about mass deaths caused by vaccination: «*Почему никто не говорит о Гибралтаре?* <…> *На 6 января у них было всего 10 смертей типа от новомодного гриппа. 10 января на остров привезли субстанцию под названием «вакцина». К 20-му января погибло 53 человека. Сегодня WOLDOmeter показывает, что погибло 70 человек. СЕМЬДЕСЯТ!!!*» (Episode #391, February 2021)

Figure 10

Frequencies of the keyword "вакцина" (vaccine) and topic-related vocabulary



Some Covid-19-related topics cannot be extracted based on one keyword. For example, misleading information about the number of Covid-19 cases was circulating on the web throughout the year. The purpose of these fake texts was to either increase the level of panic among the public by exaggerating the number of Covid-19 cases and related deaths (it was reported that there were no places in hospitals) or spread doubts in people's minds by substantially understating the numbers of infection cases. When no keyword can extract the topic, we use related vocabulary to trace major narratives. The following words are frequently used in texts containing

misinformation on the number of Covid-19 cases: *зараженный, врач, заражать, скрывать, смерть, умирать, больница, смертность, статистика, госпиталь, завышать, занижать, лгать, бояться, приказ, закрытый, показатель, информация, медсестра, знакомый, переоборудовать*. Based on the data on the frequencies of these words across the corpus, we conclude that Russian Internet users demonstrated a significant amount of interest in information on the numbers of Covid-19 cases from March 2020 to July 2020. Later, the popularity of the topic faded, and in 2021, it did not appear in the disinformation agenda at all (Fig. 11).

Top fake stories about the numbers of Covid-19 cases and "killer" doctors

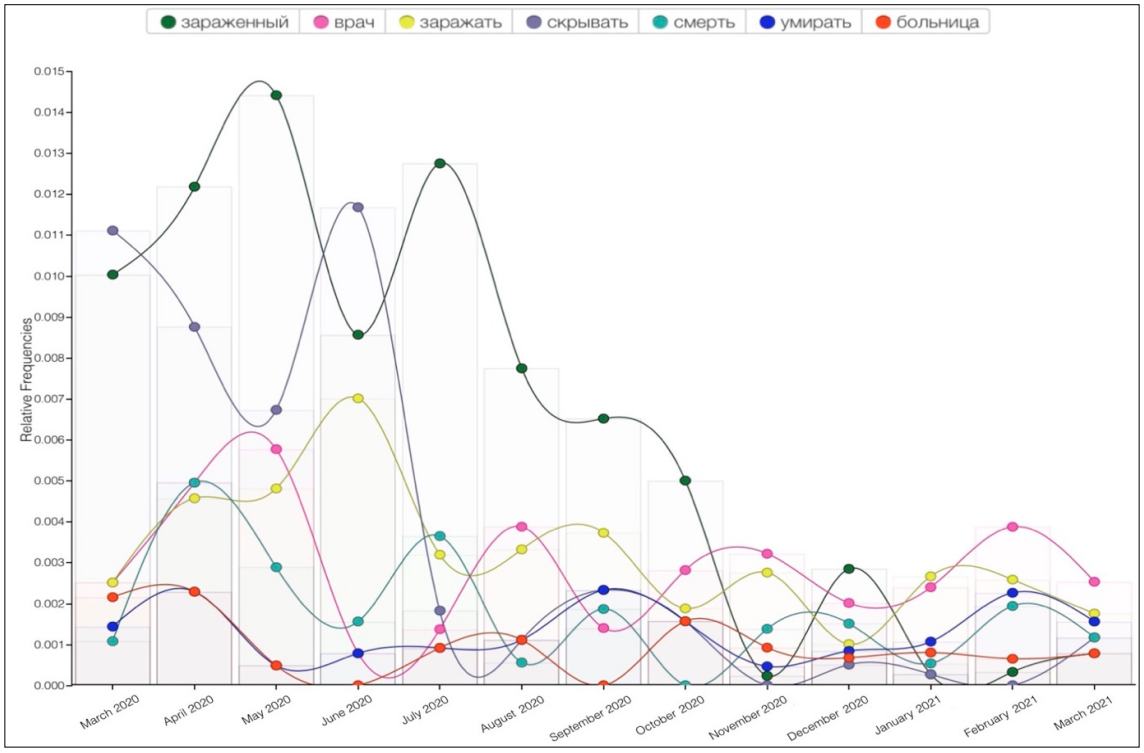
1. There are much more infected people than official statistics say / there are no places in hospitals: «*Девочки, привет. Информация на этот час. Много мертвых, уже больше 20, зараженных больше 1000, ночью их возят скорые. Ситуация очень плохая. Просто этих старых, ничего с ними не делают, просто их отвозят в сторону, а молодых лечат, потому что у них аппаратов для искусственного дыхания нет*». (Episode #12, March 2020)
2. The number of Covid-19 cases is inflated / there are no infected people at all: «*Сегодня из двух непересекающихся источников получила одну и ту же информацию, что родственникам умершего предлагают деньги за то, чтобы они подписали лист подтверждение, что человек умер от Covid-19. Вот отсюда и статистика*». (Episode #57, April 2020)
3. Stories about people who were not infected with Covid-19 but were killed by doctors in hospitals to increase the Covid-19 statistics: «*Для нагона плана статистики смертности, эту самую смертность выполняют при помощи т.н. «лечения». <…> Соседской бабушке 91 год. Бодренькая, живенькая, активная, жить и жить. Но плановая проблема с желудком, необходима операция. Увозят в больницу, делают анализы. <…> Из которой ее уже не выпускают. Труп. Причем, разумеется, никакого вскрытия. <…>*». (Episode #244, October 2020)
4. "Insider" information from the medical staff about doctors purposefully killing Covid-19 patients: «*Фельдшер из Латвии рассказала, как самостоятельно дышащих людей подсаживают на ИВЛ врачи. <…> Они дают самостоятельно дышащим пациентам лошадиную дозу мышечного релаксанта, который полностью расслабляет мышцы. И диафрагма, которая дает нам возможность дышать, расслабляется и перестает «качать» воздух. Таким образом, они останавливают дыхание, чтобы подсадить на ИВЛ*». (Episode #313, December 2020)

The keyword *ВОЗ* (WHO), ranked 27th in the corpus, was found in 19 viral episodes, representing a hybrid of true and false statements. These texts are based on a popular scheme for constructing manipulative narratives, "false headline + true statement (made by a WHO representative) + commentary leading to false conclusions." In the following example, a misleading headline is followed by a true text: *«ВОЗ признала самоизоляцию граждан бесполезной для борьбы с Covid-19. Глава ВОЗ Тедрос Гебреисус признал, что самоизоляция граждан и ограничения на передвижение не помогут ликвидировать пандемию коронавируса. По его словам, такие меры позволят выиграть время с распространением коронавируса, однако для спасения жизней этого недостаточно. Так можно лишь создать «окно возможностей»* (Episode #97, April 2020). Digital narratives with clickbait tools (the use of a headline that does not reflect what the main text is about) proved to be effective in spreading disinformation. According to recent studies (Anspach et al., 2019), more than 70% of social media users only read the headlines of articles before commenting. Other popular headlines of misleading texts representing false interpretations of the statements from the WHO found in the corpus are *«ВОЗ признала, что пандемии не существует» (The WHO has recognized that there is no pandemic), «ВОЗ заявила, что маски бесполезны» (The WHO has announced that masks are ineffective), «ВОЗ признала, что коронавирус никогда не был выделен» (The WHO has admitted that the coronavirus has never been discovered)*.

N-gram extraction from the corpus showed relatively high frequencies of the following digrams: *вторая волна (15), принудительная вакцинация (12), комендантский час (10), система пропусков (9), социальная дистанция (9), цифровой концлагерь (6), тест пцр (5), нулевой пациент (5), массовая вакцинация (4), электронный концлагерь (3)* and trigrams *(режим обязательной самоизоляции (7) тест на антитела (5)*.

Figure 11

Frequencies of the keyword "зараженный" (infected person / covid case) and topic-related vocabulary



Discussion

Diachronic corpora with a narrowed text focus can depict (depending on the unit chosen for analysis) changes in the use of a linguistic feature and certain groups of words over time. In this work, high-frequency words and significant changes in their frequency distribution in a diachronic collection of Russian viral fake Covid-19 stories serve as indicators of a change of topic on the disinformation agenda during the first year of the Covid-19 pandemic.

Recently proposed models evaluating term variation over time (Webber & Stroud, 2013; Mariani et al., 2019) allow measuring the topical change in diachronic collections in articles published in scientific journals (JDDDMP and NLP4NLP). These models are based on the quantitative method of measuring changes in keyword usage. Frequency distribution of keywords points out significant changes in term frequencies across the corpus, respectively, and flagging increases or decreases in the number of texts devoted to the related topic. This methodology is not only effective in tracking the time when a change of a topic takes place but also enables us to identify the dynamic patterns of topical fluctuations, i.e., it shows whether the drifts are sharp or gradual. However, in contrast to the above studies, the focus of the present work on the dynamics of topic change is of particular importance since there is a different factor that affects the keywords frequencies.

Furthermore, when dealing with viral Covid-19 fake stories, the topical changes are largely associated with the public interest drifts over time. Thus, the study demonstrates a correlation between the changing focus in the viral disinformation agenda (and the nature of this change) and some extralinguistic factors that could affect it. The viral disinformation agenda and the nature of this change explicate the type of public interest (rising, falling, or stable) in the topic during the year. The developing situation in the world, media coverage of the topics, official refutations, and materials debunking Covid-19 myths are some examples in this context.

Many works have focused on identifying thematic drifts and emerging topics (cf. Nel et al., 2011; Weismayer & Pezenka, 2017; Pesta et al., 2018). A keyword analysis alone is applied to diachronic corpora to determine the

keywords (the words that occur more frequently in a corpus than in the reference corpus) representing topics. However, in a case of a stylistically heterogeneous corpus containing texts across many genres and categories (as in the case of the corpus of Russian viral Covid-19 fake stories), the technique for measuring topical changes over time can be supplemented by analyzing frequencies of the words found in close proximity to the key terms in a wide variety of contexts (Williams, 1998; Brezina et al, 2015; Murakami et al, 2017). In a sense, our research can be viewed as an application of Brezina's comprehensive methodology for capturing peaks and troughs in diachronic data based on collocation networks (Brezina, 2018). We use the cluster data to extract the collocates of a word of interest and then to trace the occurrences of this topic-related vocabulary across the diachronic corpus (as shown in the section "Network Graph and Term Frequencies"). Counting frequencies of such tokens enabled us to evaluate the authors' attitude of false narratives to the covered topic and compare the lexical environment of keywords in media texts and fake stories (the Results section).

Although it is beyond the scope of this study to find explanations for all the identified drifts in public interest in the topics over time, the obtained data contributes a clearer understanding of the significant topical fluctuations in the Russian viral Covid-19 disinformation agenda during the first year of the pandemic. In addition, this study also depicts the key fake storylines related to Covid-19 that were most popular with Russian Internet users.

Conclusion

The frequency-based analysis showed that globally the most popular terms changed over the months indicating an intense dynamic of the topical change in the Russian Covid-19 disinformation agenda. In the initial months of the pandemic, Russian users were interested in issues related to restrictive measures and quarantine, symptoms of the new virus, the impact of 5G towers on the spread of coronavirus infection; this can be traced in the high frequencies of the corresponding Russian keywords. Already in the summer of 2020, these terms have lost popularity, and a few topic-related vocabulary units showed zero frequencies since that period. Increases in the footprint of the keywords *вакцина* (vaccine) and *вакцинация* (vaccination) indicate a shift in the topical focus of Covid-19 disinformation at the end of 2020. Misleading texts about masks, tests, and folk recipes against coronavirus enjoyed relatively stable user popularity; keywords related to these topics occurred in Russian fake narratives during the first year of the pandemic without showing a sharp rise and drop in the frequencies.

Notably, the words that are not related to the realities of life coronavirus life also appeared to be relatively popular within the corpus. For example, the terms *графен* (graphene), *червь* (worm), *абортивный* (abortive), *химтрейл* (chemtrail) frequently occurred in high proximity to the corresponding keywords (*вакцина* (vaccine), *маска* (маска), *ковид* (ковид) in fake texts, but these words are never found in proximity to the same keywords in the Russian texts outside the Covid-19 disinformation agenda (except for the articles and reports that debunk the ideas presented in coronavirus-related fake narratives). Thus, with the help of misleading viral texts, a new discourse is being created on the Internet, based on concepts that do not correspond to reality. A person who receives information from certain channels that massively spreads fakes has a false picture about several topics related to coronavirus infection.

The negative effects of viral disinformation about Covid-19 indicate the need to continue the fight against the spread of fakes, which is already in place in Russia today at the state level. High-quality argumentation against fakes stimulates the growth of skepticism about the infodemic among Russian users. Theories about Covid-19 and vaccines have been popular for some time, but numerous reports in the media exposing these theories have contributed to the fading of public interest in them. For instance, the term *вакцина* (vaccine) has been the most frequent corpus word since December 2020 while the corresponding topic-related vocabulary, referring to various vaccine-related theories, was losing popularity within one or two months. Public interest in the topics that were completely based on false ideas (for example, 5G, chemtrails) dropped very quickly and did not recover within a year.

As this article is being prepared, the coronavirus pandemic continues, new topics emerge, and so do false arguments against vaccines and new regulations. This research can be further developed to track the topical changes in Covid-19 disinformation. It is also necessary to investigate the factors that determine public interest

in certain topics and misleading texts as it will contribute to developing effective strategies for combating the spread of fakes.

Declaration of Competing Interest

None declared.

References

Anspach, N., M., Jennings, J., T., & Arceneaux, K. (2019). A little bit of knowledge: Facebook's news feed and self-perceptions of knowledge. *Research and Politics*, *6*(1), 1–9. https://doi.org/10.1177/2053168018816189

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

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics*. *Investigating language structure and use*. Cambridge University Press. https://doi.org/10.1017/CBO9780511804489

Biber, D., & Jones, J. K. (2009). Quantitative methods in corpus linguistics. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (vol. 2, p. 1286-1304). De Gruyter Mouton. https://doi.org/10.1515/9783110213881.2.1286

Brezina, V. (2018). *Statistics in Corpus linguistics: A practical guide*. Cambridge University Press. https://doi.org/10.1017/9781316410899.008

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistic, 20*(2), 39–73. https://doi.org/10.1075/ijcl.20.2.01bre

Budge, I, & Pennings, P. (2007) Do they work? Validating computerised word frequency estimates against policy series. *Electoral Studies*, *26*(1), 121–129. https://doi.org/10.1016/j.electstud.2006.04.002

Curzan, A. (2009) Historical corpus linguistics and evidence of language change. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (vol. 2, pp. 1091–1109). De Gruyter Mouton. https://doi.org/10.1515/9783110213881.2.1091

Faust, O. (2018) Documenting and predicting topic changes in Computers in Biology and Medicine: A bibliometric keyword analysis from 1990 to 2017. *Informatics in Medicine Unlocked*, *11*, 15-27. https://doi.org/10.1016/j.imu.2018.03.002

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Islam, M. S., Sarkar, T., Khan, S. H., Kamal, A. M., Hasan, S. M., Kabir, A., Yeasmin, D., Islam, M. A., Chowdhury, K. I. A., Anwar, K. S., Chughtai, A. A., & Seale, H. (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. https://doi.org/10.4269/ajtmh.20-0812

Ivanenko, A. A., & Zhuravlyova, K. A. (2020). CHto zaraznee: Ocenka upotrebitel'nosti leksem koronavirusnoj epohi [What's more contagious: Assessing the prevalence of coronavirus-era lexemes]. In N.V. Kozlovskaya (Ed.), *Novye slova i slovari novyh slov* [New words and dictionaries of new words] (pp. 63-70). RAS. https://doi.org/10.30842/9785604483862

Jurisica, I., &Wigle D. (2005). *Knowledge discovery in proteomics*. Chapman and Hall. https://doi.org/10.1201/9781420035162

Karjus, A., Blythe, R., Kirby, S., Smith, K. (2020). Quantifying the dynamics of topical fluctuations in language. *Language Dynamics and Change*, *10*(1), 86–125. https://doi.org/10.1163/22105832-01001200

Kim, J. D., Ohta, T., & Tsujii, J. (2008). Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, *9*(1), 1–25. https://doi.org/10.1186/1471-2105-9-10

Kim, S. N., Baldwin, T., & Kan, M. Y. (2010). Evaluating N-gram based evaluation metrics for automatic keyphrase extraction. In *Proceedings of the 23rd International Conference on Computational Linguistic*s (pp. 572–580). Coling.

Koplenig, A. (2017). The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets - Reconstructing the composition of the German corpus in times of WWII. Digital Scholarship in the Humanities, *32*(1), 169–188. https://doi.org/10.1093/llc/fqv037

Leech, G. (2005). Adding linguistic annotation. Developing linguistic corpora: A guide to good practice (pp. 17–29). Oxbow Books.

Leech, G., & Roger, F. (1992). Computer corpora - What do they tell us about culture? *ICAME Journal, 16*, 29–50.
MacFarlane, D., Tay, L.Q., Hurlstone, M. J., & Eckera, U. K. H. (2021). Refuting spurious Covid-19 treatment claims reduces demand and misinformation sharing. *Journal of Applied Research in Memory and Cognition, 10*(2), 248–258. https://doi.org/10.1016/j.jarmac.2020.12.005
Mariani, J., Francopoulo, G., Paroubek, P., & Vernier, F. (2019). The NLP4NLP corpus (II): 50 years of research in speech and language processing. *Frontiers in Research Metrics and Analytics*, *3*, 1–30. https://doi.org/10.3389/frma.2018.00036
McCarthy, M., & Carter, R. (2001). Size isn't everything: Spoken English, corpus and the classroom. *TESOL Quarterly, 35*(2), 337–340. https://doi.org/10.2307/3587654
Meurers, W. D. (2005). On the use of electronic corpora for theoretical linguistics. Case studies from the syntax of German. *Lingua, 115*(11), 1619–1639. https://doi.org/10.1016/J.Lingua.2004.07.007
Murakami, A., Thompson, P., Hunston, S., & Vajn, D. (2017). What is this corpus about?': Using topic modelling to explore a specialised corpus. *Corpora, 12*(2), 243-277. https://doi.org/10.3366/cor.2017.0118
Nel, D., van Heerden, G., Chan, A., Ghazisaeedi, M., Halvorson,W., & Steyn, P. (2011). Eleven years of scholarly research in the Journal of Services Marketing. *Journal of Services Marketing*, *25*(1), 4–13. https://doi.org/10.1108/08876041111107014
Ngula, R. (2018) Corpus methods in language studies. In *Perspectives on Conducting and Reporting Research in the Humanities* (pp. 205–223). University of Cape Coast Press.
Pesta, B., Fuerst, J., & Kirkegaard, E.O.W. (2018). Bibliometric keyword analysis across seventeen years (2000–2016) of intelligence articles. *Journal of Intelligence, 6*(4), 1–46. https://doi.org/10.3390/jintelligence6040046
Sampsel, L. J. (2018). Voyant tools. *Music Reference Services Quarterly*, *21*(3), 153-157. https://doi.org/10.1080/10588167.2018.1496754
Sarica, S., Luo, J., & Wood, K. L. (2020). TechNet: Technology semantic network based on patent data. *Expert Systems with Applications*, *142*, 112995. https://doi.org/10.1016/j.eswa.2019.112995
Scott, M. (2001). Comparing corpora and identifying key words, collocations, and frequency distributions through the WordSmith tools suite of computer programs. *Small corpus studies and ELT: Theory and practice* (pp. 47–67). John Benjamins. https://doi.org/10.1075/scl.5.07sco
Sinclair, J. (2005) Corpus and text - basic principles. Developing linguistic corpora: A guide to good practice (pp. 1–16). Oxbow Books.
Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford University Press.
Stefanowitsch, A. (2006). Negative evidence and the raw frequency fallacy. *Corpus Linguistics and Linguistic Theory*, *2*(1), 61–77. https://doi.org/10.1515/CLLT.2006.003
Stefanowitsch, A., & Gries, S. Th. (2009). Corpora and grammar. *Corpus linguistics: An international handbook* (vol. 2, pp. 933–952). De Gruyter Mouton. https://doi.org/10.1515/9783110213881.2.933
Stubbs, M. (2001). Texts, corpora, and problems of interpretation: A response to Widdowson. *Applied Linguistics*, *22*(2), 149–172. https://doi.org/10.1093/applin/22.2.149
Webber, R., & Stroud, D. (2013) How changes in word frequencies reveal changes in the focus of the JDDDMP. *Journal of Direct, Data Digit Marketing Practice*, *14*, 310–320. https://doi.org/10.1057/dddmp.2013.19
Weismayer, C., & Pezenka, I. (2017) Identifying emerging research fields: A longitudinal latent semantic keyword analysis. *Scientometrics, 113*(3), 1757–1785. https://doi.org/10.1007/s11192-017-2555-z
Wilbur, W. J., Rzhetsky, A., Shatkay, H. (2006). New directions in biomedical text annotation: Definitions, guidelines, and corpus construction. *BMC Bioinformatics*, *7*(1), 1–10. https://doi.org/10.1186/1471-2105-7-356
Williams, G. (1998). Collocational networks: Interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics*, *3*(1), 151–71. https://doi.org/10.1075/ijcl.3.1.07wil