ИЗДАТЕЛЬСТВО ГРАМОТА

Филологические науки. Вопросы теории и практики
Philology. Theory & Practice

ISSN 2782-4543 (online)
ISSN 1997-2911 (print)

2024. Том 17. Выпуск 11 | 2024. Volume 17. Issue 11
Материалы журнала доступны на сайте (articles and issues available at): philology-journal.ru

RU

Разработка цифровой модели идентификации фейковых новостей: анализ лингвистических маркеров и контекстуальных особенностей

Ляшенко Д. И.

Аннотация. Цель исследования – установление критериев, позволяющих отграничить фейк от смежных лингвистических явлений на основе анализа частотности словоупотреблений, что послужит основой цифровой модели первичной идентификации фейковых новостей. Исследование набора N-грамм и полнознаменательных лексем в формате KWIC с учетом контекста позволило установить, что для фейковых новостей 2014-2021 гг. преобладающей тематикой является внешняя политика. Научная новизна исследования состоит в том, что совпадение контекстов и отсутствие уникальных лексем, установленное в рамках анализа, дало основания сделать вывод о сходстве нефейковых и фейковых текстов вследствие маскировки последних под типичные тексты медиадискурса, что усложняет процедуру их обработки. В результате исследования доказана гипотеза о выполнении неполнознаменательными словами такой дифференцирующей и идентифицирующей роли. Наличие подобных слов помогает скрыть ложный характер сообщения и блокировать критическое мышление читателя. Текстам фейковых новостей свойственны признак анонимности автора и наличие смыслового компонента «неопределенность», что на вербальном уровне выражается в уменьшении доли личных местоимений и преобладании безличных и неопределенно-личных конструкций.

EN

Development of a digital model for identifying fake news: Analysis of linguistic markers and contextual features

D. I. Lyashenko

Abstract. The purpose of the study is to establish criteria for distinguishing a fake from related linguistic phenomena based on the analysis of the frequency of word usage, which will serve as the basis for a digital model of primary identification of fake news. A study of a set of N-grams and content words in the KWIC format, taking into account the context, allowed us to establish that foreign policy is the predominant topic for fake news in 2014-2021. The scientific novelty of the study lies in the following: the coincidence of contexts and the absence of unique lexemes, established in the framework of the analysis, made it possible to conclude that the similarity of non-fake and fake texts is due to the masking of the latter under typical media discourse texts, which complicates the procedure for their processing. As a result of the study, the hypothesis about the fulfillment of such a differentiating and identifying role by functional words is proved. The presence of such words helps to hide the false nature of the message and block the reader's critical thinking. Fake news texts are characterized by the anonymity of the author and the presence of a semantic component "uncertainty", which at the verbal level is expressed in a decrease in the proportion of personal pronouns and the predominance of impersonal and indefinite personal constructions.

Введение

Фейковые новости, ставшие во втором десятилетии XXI века неотъемлемой частью медиадискурса, представляют собой один из наиболее актуальных объектов исследования современной медиалингвистики и теории речевого воздействия. Реализуя имплицитную интенцию, подобные сообщения позволяют коммуникатору репрезентировать те или иные идеи, удовлетворяющие его интересам и потребностям, скрытно и эффективно воздействовать на когнитивную и эмоциональную сферы адресата. Реализация речевого воздействия посредством привлечения фейковых материалов позволяет увеличить круг потенциальных реципиентов, а также трансформировать их сознание и/или поведение посредством дезинформирования.

Научная статья (original research article) | https://doi.org/10.30853/phil20240587

Актуальность данного вопроса определяется тем, что одной из основных проблем современной теории речевого воздействия является установление не только содержания понятий «фейк» и «фейковые новости», но и критериев, которые позволили бы как выявлять подобные сообщения в текстах разного рода, так и отграничивать их от смежных лингвистических явлений – коммуникативного давления, манипуляции, кликбейта и т. д. Немаловажной задачей в этой связи представляется автоматизация поиска и отбора фейковых текстов среди «нефейковых» при помощи методов компьютерной лингвистики, основанных на качественном и количественном анализе собранных материалов. Это обусловлено большим потоком фейковых новостей и невозможностью их «ручной» идентификации.

Теоретическую базу исследования составляет ряд работ в рамках современной компьютерной лингвистики. Так, М. Хорас, К. Деместикас, А. Гилчик и др. (Choraś, Demestichas, Gielczyk et al., 2021) предлагают использовать автоматизированные методы выявления подделок, основанные на технологиях машинного обучения: анализ на основе текста и обработки естественного языка (NLP), репутационный анализ, сетевой анализ и распознавание манипуляций с изображениями. По свидетельству авторов, именно анализ языковых единиц, не связанных с лингвистическим контекстом (например, наборов слов или N-грамм), поможет сделать выводы о намерениях автора фейкового текста, его настроениях, предполагаемых эмоциях и других поведенческих параметрах, которые могут оказаться ценными при выявлении ложной информации. В основу сетевых моделей обнаружения ложной информации исследователи помещают сетевое встраивание и глубокие нейронные сети. Примечательно то, что один из предложенных методов, основанных на энтропии данных, может быть использован для обнаружения фейковых новостей в случае применения классификатора Maximum Entropy Discrimination (MED), который оценивает репутацию веб-страниц. Вектор репутации, лежащий в основе этого механизма, включает в себя различные факторы, такие как местоположение сервиса, местоположение заблокированного IP-адреса и время регистрации домена. Основываясь на указанных методах, мы можем отметить, что предлагаемые модели используют закономерности в распространении поддельных новостей, которые, как правило, привлекают больше пользователей и создают более плотные сетевые соединения по сравнению с подлинными новостями.

Подобный подход обнаруживаем в работах Г. П. Луна (Luhn, 1957), Х. Уинна и З. Винта (Wynne, Wint, 2019), Н. Хасана, У. Джомаа, Дж. Хориба, М. Хаггаг (Hassan, Gomaa, Khoriba et al., 2020) и других исследователей (Allcott, Gentzkow, 2017; Kong, Tan, Gan et al., 2020; Saquete, Tomás, Moreda et al., 2020), предлагающих на основе количественного анализа «установить встречаемость тех или иных языковых единиц в тексте». Определение набора N-грамм (биграмм, триграмм и т. д.), учет последовательности языковых единиц и их унификация в процессе предобработки данных способствуют установлению уникальных лексем в конкретном корпусе текстов, выявлению особенностей совокупности текстов в результате сопоставительного анализа (Luhn, 1957). Применение подобного инструментария в отношение фейковых новостей позволит выявить их дифференциальные признаки, лингвистические особенности, которые отличают их от типичных достоверных текстов медиадискурса.

В отечественном научном дискурсе также предпринимаются попытки изучения фейкового материала при помощи методов количественного анализа (Кушнерук, 2020; Кушнерук, Курочкина, 2020; Monogarova, Shiryaeva, Tikhonova, 2023; Воронцов, 2021). Так, профессор К. В. Воронцов (2021) указывает на тот факт, что компьютерные технологии AI/NLP, являясь катализаторами эпохи постправды, часто рассматриваются как усиление угрозы распространения фейковых материалов. В то же время исследователь утверждает, что именно технологии NLP, основанные на частотном анализе, позволяют обнаруживать в поддельных текстах проявления обмана (Deception Detection) и пропаганды (Propaganda Detection), выявить приемы воздействия и лингвистические маркеры фейков. Привлечение частотного анализа для установления лингвистических критериев фейка на основе корпусов фейковых текстов представляется перспективным направлением современной компьютерной лингвистики, которое позволило бы ускорить процесс верификации новостных материалов и, соответственно, противодействия дальнейшему распространению подделок.

В рамках данного исследования представляется целесообразным применить цифровые методы и приемы для установления сущности феномена «фейк» и выявления ряда закономерностей, свойственных функционированию фейковых материалов в современном медиадискурсе. Компьютерная обработка требует наличия унифицированных предобработанных данных, которые возможно исследовать при помощи таких инструментов, как AntConc и Национальный корпус русского языка (далее – НКРЯ). В соответствии с этим была проведена лемматизация собранного корпуса фейковых новостей с помощью морфологического анализатора русского языка Mystem (Segalovich, 2003), в результате чего был получен набор лемм – исходных форм слов, служащих материалом нашего исследования. На следующем этапе на основе полученных лемматизированных текстов с помощью корпус-менеджера AntConc (Anthony, 2005) был составлен частотный список для дальнейшего исследования, в том числе и компаративного анализа на базе НКРЯ (Савчук, Архангельский, Бонч-Осмоловская и др., 2024). Методологию исследования составили метод количественного анализа и корпусный метод, использованные для создания корпуса фейковых текстов и последующего количественного анализа составляющих его лексем, сравнительно-сопоставительный метод, позволивший соотнести полученный корпус и НКРЯ с целью выявления сходной тематики, употреблений и контекстов, а также метод системного научного описания для фиксации полученных результатов исследования.

Материалом для исследования послужили фейковые новости, опубликованные в 2014-2021 годах на интернет-страницах российских и украинских русскоязычных СМИ разного уровня и масштаба. В это время

данный лингвистический феномен стал активно развиваться в отечественном медиадискурсе, при этом основной тематикой данных новостей, обусловленной причинами их появления, стала эскалация конфликта России и Украины и связанные с этим политические и социально-экономические явления. В рамках исследования нами были проанализированы тексты из ряда информационных источников, включающих ТАСС, «Комсомольскую правду», РИА Новости, Газета.ru, «Российскую газету», Lenta.ru, «Аргументы и факты», РБК, РЕН ТВ, НТВ, 360tv, ТВ «Звезда», Известия.ру, «Московский Комсомолец», News NSK, Privet-Rostov, NewsGid.net, Star hit, «Царь Град» ТВ, The Village (*), L!fe.ru, «Новости Крыма», ТСН «Украина», «Прессу Украины», Свідок.info, Корреспондент.net, «Днепр вечерний», «Голос русскоязычной Америки», Украина.ру. При этом были отобраны только те новостные материалы, для которых были выпущены официальные опровержения. Сбор фейков осуществлялся в соответствии с двумя критериями – уровнем СМИ (локальный, региональный, международный) и тематикой (общество, политика, экономика, культура, образование и спорт), что позволило классифицировать материал с учетом ряда параметров (время появления, предмет речи, сфера функционирования, форма существования, уровень описываемых событий, цель сообщения). Объем собранного корпуса составил 53 текста (8054 токена). Частотный список в сравнении с Частотным словарем газетного подкорпуса Национального корпуса русского языка (ЧСНКРЯ) представлен в Таблице 1.

Для достижения цели данного исследования необходимо решить следующие задачи:

– создать корпус лемматизированных фейковых медиатекстов для первичной автоматизированной обработки;
– определить основную тематику данных сообщений;
– проанализировать контексты и выявить специфические лингвистические единицы (уникальные лексемы).

Гипотеза заключается в том, что специфика фейковых текстов связана с функционированием неполнознаменательных слов, что определяется наличием таких характеристик фейковых новостей, как анонимности, семы неопределенности и интенсивности. В то же время набор полнознаменательных слов в фейковых и нефейковых текстах в целом совпадает, что обусловлено сходной тематикой подобных текстов и одной сферой функционирования – медиадискурсом.

Практическая значимость исследования состоит в возможности применения его результатов для создания цифровой модели первичной обработки фейковых материалов, что позволит частично автоматизировать процесс распознавания фейков.

Обсуждение и результаты

При помощи методов компьютерного анализа мы установили, что наиболее часто в фейковых новостях встречается предлог «*в*», что обусловлено наличием множества деталей о месте, времени и т. д. описываемых событий, некоторой попытке блокировать критическое мышление реципиента, его способность отличить истину от лжи (частое употребление предлога «*в*» указывает на попытку погружения вглубь, в то время как союз «*и*» ассоциируется с неким «скольжением по поверхности» (Эпштейн, 2003, с. 86-87)). В данном случае имеет место чрезмерная детализация гипотетической ситуации (Баранов, 2018, с. 9), направленная на маскировку ложных сообщений.

Кроме того, предлог «*в*» используется преимущественно для указания на место произошедших событий, описанных в тексте статьи. Отчасти это обусловлено спецификой жанра (1990-2000-е гг.) – публицистический текст. В частотном словаре газетного подкорпуса НКРЯ предлог «*в*» также занимает первое место, что можно считать характеристикой новостного дискурса в целом. Еще одно подтверждение этого факта находим в частотном словаре Л. Н. Засориной, где данный предлог «*в*» также является наиболее частотной леммой, а значительную часть материала словаря (47,4%) составляют публицистические, газетные и журнальные тексты «от произведений Ленина и Горького до 60-х годов» (1977, с. 6).

Примечательно то, что частица «*не*» перемещается на 4-е место, соответственно, доля отрицания и отрицательных конструкций в подобных текстах уменьшается, чтобы потенциальный читатель чаще соглашался с сообщаемой ему информацией. В подобном случае можно говорить о меньшем отторжении информации, противоречащей мнению, позиции предполагаемых реципиентов.

В качестве одной из ключевых характеристик фейковых текстов следует отметить признак анонимизации: автор текста не называет себя, при этом повествование ведется от третьего лица, о чем свидетельствует меньшее употребление личного местоимения «*я*». Данное местоимение присутствует только в цитатах экспертов или очевидцев. В то же время создатели фейковых новостей достаточно редко ссылаются на конкретных экспертов, а наличие ссылки не всегда подкрепляется прямым цитированием.

Примечательно, что первым знаменательным словом в рассматриваемом корпусе фейковых текстов является лексема «*Россия*», занимающая 30-е место в полученной частотной таблице. Данное имя существительное выступает преимущественно в роли обстоятельства места, где происходят описываемые события («*по югу России*», «*по всей России*», «*пришла из России*» и т. д.), либо несогласованного определения («*стремление России*», «*МЧС России*», «*президента России*»), указывающего на принадлежность чего-либо России; особое место отводится метонимии – в данном случае «*Россию*» можно рассматривать как обобщенного исполнителя действия в рамках описываемых в материале событий («*заставляют свидетельствовать против России*», «*агрессия России*», «*стремление России*»). Анализ N-грамм – устойчивых сочетаний из нескольких элементов, включающих лексему «Россия», – показывает, что в данном корпусе представлены преимущественно сочетания, использующиеся

для описания места действия («*посольства в России*», «*по югу Росии*»* (*орфография сохранена), «*не только в России*» и т. д.), либо характеризующие действия страны («*связанной с Россией*», «*символ стремления России*», «*учитывая агрессию в России*»). Следует отметить, что большинство сочетаний, встречающихся в российских источниках, носит нейтральный или положительный характер, в то время как сочетания, обнаруживаемые в украинских СМИ, носят преимущественно отрицательный характер, что связано с конфликтом двух государств.

Таблица 1. *Частотность лемм газетного подкорпуса НКРЯ и фейковых новостей (2014-2021 гг.)*

Порядковый номер леммы по частотности употребления	ЧСНКРЯ (газетный подкорпус)	ФЕЙКИ	Freq (fake news)
1	в	в	488
2	на	и	269
3	и	на	214
4	что	не	153
5	с	что	145
6	по	по	104
7	год	с	104
8	быть	из	96
9	россия	как	61
10	не	о	59
11	это	а	57
12	о	это	50
13	который	за	46
14	он	от	44
15	президент	для	39
16	за	его	37
17	об	к	36
18	из	то	35
19	страна	об	34
20	к	но	33
21	заявить	она	33
22	для	уже	33
23	%	у	30
24	также	их	29
25	сша	после	29
26	этот	был	28
27	как	было	28
28	то	он	28
29	украина	до	26
30	российский	россии	25

Среди глаголов в полученной частотной таблице первые строчки занимают «*есть*» («*быть*») и «*сообщать*». Частое употребление первой лексемы связано с ее использованием в качестве глагола-связки в составном именном сказуемом в подавляющем большинстве случаев. Большая часть контекстов представлена сочетанием глагола «*быть*» с краткой формой страдательного причастия, что объясняется особенностями публицистического стиля, сферой употребления – медиадискурс – и частым отсутствием указания на исполнителя действия («*сайт был взломан*»; «*оригинал был взят из статьи*»; «*Лашрауи также был причастен к парижским терактам*»; «*когда был съеден чудищем из водных глубин*» и т. д.).

Лексема «*есть*» используется преимущественно в качестве глагола-связки, части составного именного сказуемого. Иной контекст находим в сочетании «*то есть*», использующемся для пояснения определенных моментов; указание наличия чего-либо где-либо: «*есть нормальная потребность*»; «*есть Николай (святой)*»; «*есть дом в Майами*».

Форма «*было*» используется преимущественно в безличных предложениях, т. к. исполнитель действия не указан либо неизвестен. N-граммы в первую очередь указывают на наличие у кого-либо какого-либо предмета или объекта: «*у Орбакайте есть*», «*украинского народа есть*», «*у нас есть*», «*есть билеты*», «*есть нормальная биологическая…*». Прочая часть N-грамм отсылает к союзу «*то есть*» как средству связи частей сложного предложения или пояснению того, что было изложено ранее.

В то же время следует отметить частотность употребления глагола «*сообщать*», преимущественно в форме «*сообщает*». В подавляющем большинстве случаев он указывает на источник сообщения данных, при этом в текстах преобладает фактологическая информация, представленная преимущественно посредством имен существительных либо местоимений. Подобную закономерность отчасти можно объяснить особенностью новостных текстов в целом, сферой функционирования – медиадискурсом. Среди N-грамм часто встречается сочетание «*сообщ(или/ает) о/об*», что объясняется жанровой принадлежностью текстов и указанием на объект новости, предмет речи.

Особое внимание среди лемм в таблице частотности привлекает первое по частотности имя прилагательное «*украинский*», которое часто содержит отрицательную коннотацию: «*украинские мошенники*», «*украинских*

шпионов», «*украинский национализм*», «*украинские санкции*» и пр. Примечательно, что подобные характеристики встречаются в текстах русскоязычного медиадискурса, где репрезентировано соответствующее отношение к объектам, связанным с Украиной. В то же время необходимо отметить, что значительная часть контекстов связана с политикой и образованием: «*украинские СМИ*», «*украинские власти*», «*украинские культурологи*», «*украинские патриоты*», «*украинские университеты*», «*украинские фантазии*» и т. д.

Таким образом, на основе лемматизированного корпуса фейков были установлены наиболее частотные лексемы. Следует также отметить, что с помощью AntConc в представленном материале не были выявлены ключевые слова, которые отличали бы фейки, относящиеся к разным временным промежуткам, что может свидетельствовать о приблизительно одинаковом наборе лексики, использующемся как для текстов медиадискурса, так и для фейковых текстов.

Следующий инструмент, который был использован для компьютерной обработки корпуса фейков, – Национальный корпус русского языка, который содержит огромное количество текстов разных стилей и жанров. В соответствии с тем, что анализируемые в рамках настоящего исследования фейки функционируют в сфере медиадискурса, особый интерес представляет газетный подкорпус НКРЯ. Целью данной части исследования является сравнение функционирования отдельных лексем в фейках и в стандартных медиатекстах. Так, для анализа было выбрано два слова (/леммы): «*год*» и «*украинский*», при этом выбор данных лексем объясняется частотностью их употребления в корпусе фейков.

В целом в газетном подкорпусе НКРЯ было найдено 1 576 022 документа, в которых встречается лемма «*год*», 5 717 420 вхождений, что свидетельствует о широком распространении данной словоформы в газетном дискурсе. Частность леммы «*год*» объясняется обязательным указанием даты событий, описываемых в новости. Часть употреблений слова «*год*» связана с указанием каких-либо измерений. Из 12 контекстов 6 связано с измерением проседания крымского моста, 1 – с возрастом лица (Сталлоне), 5 – непосредственно с датой/временем.

Анализ N-gramm позволил определить круг основных употреблений леммы «*год*», среди которых значительная часть отводится указанию на конкретный год описываемых событий, их продолжительность или срок завершения. Более подробные результаты представлены в Таблице 2.

Таблица 2. *Употребления леммы «год»*

N-gramm	Значение/ функция	Число вхождений	Пример употребления
2-граммы	указание на конкретный год произошедших событий	905608	«*“Роскосмос” ожидает подписания соглашения по лунной станции с Китаем в **2022 году***» («Роскосмос» ожидает подписания соглашения по лунной станции с Китаем в 2022 году // Ведомости. 31.12.2021).
	продолжительность какого-либо события / срок его завершения	718892	«*За **два года** в результате прокурорского вмешательства заблокировано свыше 100 подобных сайтов*» (В Приморье демонтировали «китовую тюрьму» // Парламентская газета. 30.12.2021) «*В рамках поручения президента до конца **текущего года** будет обновлена схема размещения атомных и гидроэлектростанций*» (В Минэнерго рассказали о планах построить атомную станцию в Приморье // Парламентская газета. 20.10.2021).
3-граммы	количество денег, выделенное, полученное или потраченное на определенное событие или нужды	17413	«*Мы ожидаем, что Китай приобретёт сельскохозяйственных товаров на 40-50 миллиардов **долларов в год***» (Китай обязался увеличить покупки продукции из США на $200 млрд за два года // Московский комсомолец. 14.12.2019).
5-граммы	сравнение двух или более событий, результатов и т. д. в разные периоды времени	26927	«*по сравнению с прошлым годом*», «*за аналогичный период прошлого года*», «*больше, чем в прошлом году*», «*по сравнению с предыдущим годом*», «*как и в прошлом году*» и т. д.
	срок лишения свободы и отбывания наказания в исправительных учреждениях	23185	«*на срок до пяти лет*», «*к трем годам лишения свободы*», «*грозит до пяти лет лишения*» и т. д.
	количество, объем, единицы измерения, связанные преимущественно с экспортом газа и нефтепродуктов	5725	«*млрд куб. м в год*», «*млн т нефти в год*», «*1 кв. м в год*», «*куб. м газа в год*», «*миллиардов кубометров газа в год*»
	указание на «особые события» (ВОВ, переговоры, военная операция, экстремистская деятельность)	8105	«*в годы Великой Отечественной войны*», «*в годы Второй мировой войны*», «*терактов 11 сентября 2001 года*», «*с 1992 года ведутся переговоры*», «*2014 года в Крыму прошел*» и т. д.

При анализе контекстов в формате KWIC (Key Word in Context, показывающий представление корпуса относительно анализируемого слова) (Савчук, Архангельский, Бонч-Осмоловская и др., 2024) с учетом левого/правого контекста употребления по дате создания в обратном порядке было установлено, что среди первых 56 контекстов 16 из них относятся к «Новому году» как к празднику. Вероятно, это связано с размещением в газетах поздравлений с данным праздником, а также разнообразных статей, связанных с соответствующими традициями и блюдами праздничного стола. Соответственно, в контекстах также фигурирует формулировка «*уходящий год*» как противопоставление новому:

- «*Отпраздновать Новый **год** хочется с размахом – с гостями ←…→*» (Что запретят в новогоднюю ночь // Парламентская газета. 31.12.2021);
- «*Но на Новый **год** можно позволить полбокала вина, считает ←…→*» (Что будет есть Жириновский в новогоднюю ночь // Парламентская газета. 31.12.2021);
- «*Турчак поздравил россиян с Новым **годом** и подвел итоги работы…*» (Турчак поздравил россиян с Новым годом и подвел итоги работы // Парламентская газета. 31.12.2021).

Кроме того, большая часть примеров связана с указанием года произошедшего события, о котором идет речь в новости, что подтверждает выводы, сделанные выше в рамках данного анализа.

- «*Провести в 2022 **году** в Российской Федерации Год культурного…*» (Путин подписал указ о проведении Года культурного наследия народов России // Парламентская газета. 31.12.2021);
- «*В 2020 **году** приставы возбудили более 100 тысяч…*» (Пенсии, вычеты и пособия: новые социальные законы // Парламентская газета. 31.12.2021);
- «*За четыре года до этого, в 1894 **году**…*» (Когда москвичи и питерцы впервые поздравили друг друга с новым годом по телефону // Парламентская газета. 31.12.2021).

Теперь перейдем ко второй выбранной нами лемме – «*украинский*». В целом в корпусе найдено 123 843 документа, 267 279 вхождений, что свидетельствует о среднем распространении данной словоформы в газетном дискурсе. Следует отметить то, что данная словоформа начинает постепенно употребляться с 1983 года, в то время как ее отсутствие в ранних источниках можно объяснить лакуной в газетном подкорпусе. В 1990-х годах заметен небольшой рост числа словоупотреблений, что, вероятно, связано с отделением Украины как самостоятельного государства и упоминанием предметов/явлений, относящихся к ней как украинских.

Особого внимание заслуживает резкий рост употребления словоформы, начавшийся в 2010 году и достигший пика приблизительно в 2016-2017 гг. Предположительно, это связано с созданием напряженности на Украине, Майданом, а затем и эскалацией конфликта в данной стране. В то время большая часть новостей была посвящена именно российско-украинской тематике. Также необходимо отметить последовавшее за этим постепенное уменьшение количества словоупотреблений, что связано, вероятно, с выходом на первый план пандемии COVID-2019 начиная с середины 2019 года. До этого же ослабление позиций можно связать со ставшей привычной информацией о российско-украинских отношениях, конфликте на Донбассе и отсутствии большого числа сенсаций.

Анализ 2-грамм показал, что наиболее частотными устойчивыми сочетаниями являются те, что:

1) связаны с политической сферой (34 599 вхождений): «*украинские власти*», «*украинского президента*», «*украинского лидера*», «*украинские политики*» и т. д.;

2) указывают на принадлежность чего-либо Украине, отношение к ней (22 536 вхождений): «*украинские новости*», «*украинские сми*», «*украинского народа*», «*украинский язык*», «*украинской столице*» и т. д.;

3) связаны с силовыми структурами, армией (19 138 вхождений): «*украинские силовики*», «*украинских военных*», «*украинской армии*», «*украинских националистов*» и т. д.;

4) связаны с международными отношениями (17 173 вхождения): «*украинская делегация*», «*украинская сторона*», «*российско украинской*»* (*орфография сохранена), «*украинской границе*» и т. д.

Анализ 3-грамм показал, что наибольшее распространение получили сочетания, указывающие в первую очередь на украинскую армию и политическую сферу (11 620 вхождений). Вероятно, отчасти это связано с эскалацией конфликта на Украине и обострением ситуации в ДНР и ЛНР.

- «*Степан Бандера – лидер Организации украинских националистов (ОУН, запрещена в России) и **Украинской повстанческой армии** (УПА, запрещена в России)*» («Профессиональный агент Гитлера»: в США рассекречены документы о Степане Бандере // Vesti.ru. 24.01.2020);
- «*Немцы по согласованию с Центральным проводом Организации **украинских националистов (ОУН)*** (**) *начали формирование “Легиона”…*» (Захарова назвала выступление постпреда Украины в ООН позорным // РИА Новости, 03.12.2020).

Интересно отметить тот факт, что среди прочих 3-грамм только один раз встречается сочетание, относящееся к сфере религии, – «*украинская православная церковь*». Это связано со скандалом вокруг ПЦУ (Православной церкви Украины, созданной в 2018-2019 гг. после разрыва «отношений» с Московским патриархатом) и соответствующих новостных материалов, в которых освещается данное событие:

- «***Украинская православная церковь** Московского патриархата (УПЦ МП) посчитала регистрацию ПЦУ незаконной*» (Зеленский призвал священников к диалогу // Коммерсант. 28.07.2019);
- «*Единственный диалог, который патриарх и **Украинская православная церковь** готовы и хотят вести с РПЦ, – это диалог о признании Москвой автокефалии украинской церкви*» (Киев опроверг просьбу патриарха Украины о помиловании // Ведомости. 30.11.2017).

Следует отметить, что в 4-граммах преобладает тема политики и вооруженного конфликта на Украине (9757 вхождений):

- «*27 февраля **украинский парламент утвердил состав** так называемого правительства народного доверия…*» (СМИ: трагедия на Украине вызвана невниманием к русскоязычному народу // РИА Новости. 03.06.2014);
- «*Переговоры “нормандской четверки”, состоявшиеся 12 февраля, прошли на фоне резкого обострения ситуации в **Донбассе**, **спровоцированного украинскими силовиками**…*» (Советник Порошенко: силовики потеряли в Донбассе 1,7 тысячи человек // РИА Новости. 07.06.2015).

При работе с 5-граммами обнаружены интересные сочетания, выполняющие следующие функции:

1) указание на политическую сферу и силовые структуры (5503 вхождения): «*украинский парламент утвердил состав так*», «*с украинским лидером владимиром зеленским*», «*базу данных украинского сайта миротворец*», «*украинские власти назначили выборы в*» и т. д.;

2) указание на источник информации (СМИ, организацию) (3102 вхождения): «*об этом сообщает украинская правда*», «*об этом сообщают украинские новости*», «*со ссылкой на пресс-службу украинского*», «*об этом сообщает украинское издание*», «*со ссылкой на украинские сми*» и т. д.;

3) указание на экономическую сферу и изменения в ней (1684 вхождения): «*индекс украинской биржи повысился на*», «*акции на торгах украинской биржи*», «*среди украинских банков по активам*» и т. д.;

4) указание на религиозную структуру и организации Украины (1243 вхождения): «*украинской православной церкви киевского патриархата*», «*канонической украинской православной церкви упц*», «*и украинская автокефальная православная церковь*» и т. д.;

5) указание на «особые события» (военная операция, экстремистская деятельность и др.) (1167 вхождений): «*проблема урегулирования украинского кризиса обсуждается*», «*украинской повстанческой армии оун упа*» (**) и т. д.;

6) указание на языковую политику в отношении украинского языка (599 вхождений): «*об обеспечении функционирования украинского языка*», «*на любом языке кроме украинского*» и т. д.

В свою очередь, анализ контекстов в формате KWIC (Савчук, Архангельский, Бонч-Осмоловская и др., 2024) с учетом левого/правого контекста употребления по дате создания в обратном порядке показал, что наибольшее распространение получили контексты, связанные с внешней политикой Украины, военной деятельностью и эпохой СССР. Вероятно, это связано с эскалацией конфликта на Украине, усиленным развитием военной отрасли и отсылкам к «советскому прошлому» данной страны. Часть примеров посвящена современной спецоперации, начавшейся в феврале 2022 года:

- «*…союз объединились четыре республики*: *РСФСР*, ***Украинская*** *ССР*, *Белорусская ССР и Закавказская…*» (Когда был создан СССР // Парламентская газета. 30.12.2021);
- «*Любое нарушение* ***украинской*** *границы и любое вторжение будет…*» (Глава МИД Франции заявил о необходимости диалога с Россией // Ведомости. 21.11.2021);
- «*…2018 год он по заданию* ***украинской*** *спецслужбы осуществлял сбор и передачу…*» (Аксёнов поблагодарил Бортникова за обеспечение безопасности Крыма // Парламентская газета. 15.11.2021).

Кроме того, значительная часть примеров связана с темой транзита газа в Европу по территории Украины и деятельности Нафтогаза и Газпрома, что объясняется актуальностью европейского энергетического кризиса и остротой вопроса о поставке нефти и газа:

- «*Ранее* ***украинский*** *“Нафтогаз” сообщил*, *что его крупнейшие…*» (Миллер предупредил о снижении запасов газа в ПХГ Украины до исторического минимума // Ведомости. 29.12.2021);
- «***Украинская*** *компания утверждает*, *что “Газпром” резко…*» (В «Газпроме» прокомментировали обвинения в недостаточных объемах поставок газа // Ведомости. 26.12.2021).

Примечательно, что по запросу «*украинский*» корпус также выдает контексты, в которых затрагиваются «острые» темы, актуальные как для украинских, так и для российских граждан: вопросы языковой политики (в том числе о государственном языке и запрете русского языка), вопросы ограничения телерадиовещания отдельных каналов, суверенности государства и санкционной политики:

- «*Wildberries оценил потери бизнеса от* ***украинских*** *санкций…*» (Wildberries оценил потери бизнеса от украинских санкций // Ведомости. 27.07.2021);
- «*…МОК исправить карту Олимпиады с “****украинским****” Крымом…*» (Российское посольство призвало МОК исправить карту Олимпиады с «украинским» Крымом // Ведомости. 24.07.2021);
- «*…подписал закон “Об обеспечении функционирования* ***украинского*** *языка как государственного”*, *закрепляющий исключительные…*» (На Украине предложили сменить «русскоязычные названия» населенных пунктов // Парламентская газета. 03.11.2021).

Таким образом, проведенный анализ показал значительную распространенность словоформы «*год*» и среднюю распространенность словоформы «*украинский*» в газетном подкорпусе НКРЯ. Следует отметить, что первая из словоформ используется преимущественно для указания на время произошедшего события, описанного в теле статьи, а вторая получила наибольшее распространение в связи с эскалацией конфликта на Украине в 2014-2016 гг. и пережила второй «рост популярности» в связи с проведением специальной военной операции в 2022 году. Употребления, установленные при анализе НКРЯ, совпадают с данными корпуса фейковых текстов, что свидетельствует об умелой маскировке подделок под достоверные новости из официальных источников с целью увеличения срока функционирования фейка.

Заключение

На основании проведенного исследования мы пришли к следующим выводам. Применение цифровых методов при анализе фейковых текстов позволяет эффективно осуществить их первичную автоматизированную обработку для последующего анализа (лингвистической экспертизы). Набор N-грамм в комплексе с анализом полнознаменательных лексем в формате KWIC с учетом левого/правого контекста употребления

по дате создания в обратном порядке показал, что для фейковых новостей 2014-2021 гг. преобладающей тематикой является внешняя политика, в том числе эскалация конфликта на Украине. Кроме того, совпадение смыслового контекста и обнаруженное в ходе анализа отсутствие уникальных лексем как таковых позволяет говорить о сходстве фейковых и нефейковых текстов, что, в свою очередь, свидетельствует о маскировке фейков под типичные тексты медиадискурса, что помогает увеличить эффективность их воздействия и время функционирования. Сходный лексический состав медиатекстов указывает на необходимость обращения к анализу неполнознаменательных слов (модальных слов, предлогов, частиц и т. д.) в качестве одного из специфических критериев различения фейковых и нефейковых материалов. В статье доказано, что фейковый характер материала может быть связан с особенностями функционирования неполнознаменательных слов, влияющих на репрезентацию признаков анонимности (преобладание предлогов и частиц над личными местоимениями), интенсивности и неопределенности (модальные слова). Проведенный анализ корпуса фейковых текстов подчеркивает значимость структурных характеристик языка в выявлении лингвистических параметров фейковых текстов, что позволит разработать более точные модели автоматического распознавания фейков. В рамках дальнейших исследований представляется возможным лингвистическое параметризирование данных языковых показателей с целью установления авторства (определения специфики языковой личности коммуникатора).

Примечания:
(*) – выполняет функции иностранного агента.
(**) – признана в России экстремистской и запрещена.

Источники | References

1. Баранов А. Н. Злоупотребление правом как лингвистический феномен // Язык. Право. Общество: сборник статей V международной научно-практической конференции (г. Пенза, 22-25 мая 2018 г.). Пенза: Пензенский государственный университет. 2018.
2. Воронцов К. В. Фейковые новости и другие виды потенциально опасного дискурса: типология, подходы, датасеты, соревнования // Международная независимая открытая конференция по ИИ “OpenTalks.AI” (г. Москва, 3-5 февраля 2021 г.). М., 2021.
3. Засорина Л. Н. Частотный словарь русского языка. М.: Русский язык, 1977.
4. Кушнерук С. Л. Дискурсивный мир информационно-психологической войны: репрезентационная структура по данным корпуса // Политическая лингвистика. 2020. № 5 (83).
5. Кушнерук С. Л., Курочкина М. А. Информационно-психологическая война в зарубежной медиакоммуникации: взгляд дискурсолога // Вестник Челябинского государственного университета. 2020. № 7 (441).
6. Савчук С. О., Архангельский Т. А., Бонч-Осмоловская А. А., Донина О. В., Кузнецова Ю. Н., Ляшевская О. Н., Орехов Б. В., Подрядчикова М. В. Национальный корпус русского языка 2.0: новые возможности и перспективы развития // Вопросы языкознания. 2024. № 2.
7. Эпштейн М. Н. Предлог «В» как философема. Частотный словарь и основной вопрос философии // Вопросы философии: научно-теоретический журнал. 2003. № 6.
8. Allcott H., Gentzkow M. Social media and fake news in the 2016 election // Journal of Economic Perspectives. 2017. Vol. 31. No. 2.
9. Anthony L. AntConc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom // Proceedings of the International Professional Communication Conference (Limerick, 10-13 Jily 2005). Limerick, 2005.
10. Choraś M., Demestichas K., Gielczyk A., Herrero A., Ksieniewicz P., Remoundou K., Urda D., Wozniak M. Advanced Machine Learning techniques for fake news (online disinformation) detection: A systematic mapping study // Applied Soft Computing. 2021. Vol. 101.
11. Hassan N., Gomaa W., Khoriba G., Haggag M. Credibility detection in twitter using word n-gram analysis and supervised machine learning techniques // International Journal of Intelligent Engineering and Systems. 2020. Vol. 13.
12. Kong S., Tan L., Gan K., Samsudin N. Fake news detection using deep learning // 2020 IEEE 10th symposium on computer applications & industrial electronics (ISCAIE) (Malaysia, 18-19 April 2020). Piscataway, 2020.
13. Luhn H. P. A statistical approach to mechanized encoding and searching of literary information // IBM Journal of Research and Development. 1957. Vol. 1. No. 4.
14. Monogarova A., Shiryaeva T., Tikhonova E. The words that make fake stories go viral: A corpus-based approach to analyzing Russian Covid-19 disinformation // Russian Journal of Linguistics. 2023. Vol. 27. No. 3.
15. Saquete E., Tomás D., Moreda P., Martínez-Barco P., Palomar M. Fighting post-truth using natural language processing: A review and open challenges // Expert Systems With Applications. 2020. Vol. 141.
16. Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine // Proceedings of the International conference of Machine Learning: Models, Technologies and Applications (MLMTA’03) (23-26 June 2003). Las Vegas, 2003. Vol. 2003.
17. Wynne H., Wint Z. Content based fake news detection using n-gram models // Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services (Munich, 2-4 December 2019). Munich, 2019.

Информация об авторах | Author information

RU

Ляшенко Дарья Игоревна[1]
[1] Южный федеральный университет, г. Ростов-на-Дону

EN

Darya Igorevna Lyashenko[1]
[1] Southern Federal University, Rostov-on-Don

[1] *dlyashenko@sfedu.ru*

Информация о статье | About this article

Дата поступления рукописи (received): 14.10.2024; опубликовано online (published online): 25.11.2024.

Ключевые слова (keywords): фейковые новости; смысловой компонент «неопределенность»; корпусная лингвистика; речевое воздействие; неполнознаменательные лексемы; fake news; semantic component “uncertainty”; corpus linguistics; speech impact; functional lexemes.