УДК 81'373:578.27
DOI https://doi.org/10.32841/2409-1154.2019.43.4.36

Махачашвілі Р. К.,
доктор філологічних наук, доцент,
завідувач кафедри романської філології
та порівняльно-типологічного мовознавства
Інституту філології
Київського університету імені Бориса Грінченка

Білик К. М.,
викладач кафедри романської філології
та порівняльно-типологічного мовознавства
Інституту філології
Київського університету імені Бориса Грінченка

КОРПУСНЕ ДОСЛІДЖЕННЯ ТЕКСТІВ РУБРИКИ «НАДЗВИЧАЙНІ НОВИНИ» У ФРАНЦУЗЬКІЙ, АНГЛІЙСЬКІЙ ТА УКРАЇНСЬКІЙ МОВАХ

Анотація. Стаття присвячена дослідженню корпусу текстів рубрики «Надзвичайні новини» у французькій, англійській та українській мовах. Для укладання тематичного глосарію надзвичайних новин тематики «Вірусні захворювання» ми використали інструмент корпусної лінгвістики «Voyant Tools». Визначено ключові лексеми, встановлено синонімічний ряд і проаналізовано частоту використання слів і словосполучень вищезазначеної тематики. Метою статті є ознайомлення з базовими поняттями корпусної лінгвістики, опис можливостей роботи з корпусними базами даних у лінгвістичних дослідженнях, аналіз інструменту корпусної лінгвістики «Voyant Tools», укладання інноваційного глосарія текстів рубрики «Надзвичайні новини» тематики «Вірусні захворювання» у французькій, англійській, українській мовах. Корпусна лінгвістика є надзвичайно важливою галуззю сучасної лінгвістичної науки. Представлена робота містить практичний аналіз корпусу текстів новинного спрямування. У центрі уваги корпусного дослідження опиняється мовна особа, тобто її мовна діяльність, масова комунікація, проблема її опису. Саме тому досліджуваний інструмент корпусної лінгвістики приділяє увагу конкордансу, відносній та абсолютній частотності слів, ключовим словам, сполучуваності слів, багатокомпонентним групам, контекстуальному вживанню слів.

Реалізація поставленої мети передбачає виконання таких завдань:

– дослідити новинні тексти рубрики «Надзвичайні новини» у французькій, англійській та українській мовах;

– відібрати характерні для тематики «Вірусні захворювання» лексичні одиниці;

– проаналізувати семантичну специфіку відібраного лексичного матеріалу.

У ході дослідження новинних повідомлень було встановлено, що найчастіше вживаними лексемами є: вірус, зараження, поширення, люди, Китай, адже саме ці слова дають читачеві схематичну картину того, про що йтиметься у газетній шпальті.

Ключові слова: корпусна лінгвістика, надзвичайні новини, лексеми, текст, частота, коронавірус, «Voyant Tools».

Постановка проблеми. Останнім часом у потоці інформаційно-комунікаційних технологій набуває особливого значення автоматичне опрацювання інформації. Завдяки інформаційно-комунікаційним технологіям, які мають універсальні технічні можливості для аналізу, збереження та відбору матеріалу мови, розвивається нова галузь мовознавчих досліджень – корпусна лінгвістика, основним завданням якої є аналіз великої кількості емпіричного матеріалу та створення об'єктивних висновків щодо лінгвістичного корпусу тексту.

Корпусна лінгвістика передбачає застосування загальних принципів побудови, обробки лінгвістичних корпусів текстів, які здійснюються за допомогою інформаційно-комунікаційних технологій, розробку методик збору мовних явищ. Згідно з дослідженням В.А. Широкова корпус характеризується як «обмежений за об'ємом набір електронних текстів, зібраних із метою максимально точно представити варіант мови, яка досліджується» [4, с. 24]. В.П. Захаров зазначає, що корпус текстів – «значний за обсягом, представлений в електронному вигляді, уніфікований, структурований, розмічений, філологічно компетентний масив мовних даних, створений для вирішення конкретних лінгвістичних завдань» [8, с. 3].

О.М. Демська-Кульчицька визначає: «Корпус текстів – машиночитане стандартно організоване зібрання репрезентативних для певної мови, діалекту або іншої підмножини мови писемних або усних текстів, призначених для лінгвістичного аналізу й опису, відібраних і впорядкованих згідно з експліцитними екстра- та інтралінгвальними критеріями» [7, с. 20].

Спираючись на дослідження вищезгаданих науковців, виокремимо характерні ознаки корпусу текстів, а саме:

– емпіричний підхід до синтезу мовних даних;

– створення великого за обсягом корпусу тексту як основи для дослідження лінгвістичного аналізу;

– залучення комп'ютерних технологій для розбору матеріалу;

– аналіз частоти використання слів і словосполучень певної тематики.

Виклад основного матеріалу. Контентом корпусів текстів слугували 12 повідомлень рубрики «Надзвичайні

новини» французького видання «Le Parisien», англійського «The Guardian» та українського «Газета по-українськи» тематики «Вірусні захворювання».

Зважаючи на дослідження вищезазначених науковців, ми дійшли висновку, що аспекти корпусу текстів були досліджені з погляду на загальні характеристики. Крім того, у працях не зверталася увага на застосування програми для дослідження корпусу текстів «Voyant tools». Більш того, окремі фахівці акцентують на вивченні корпусу текстів лише українською мовою, тому **метою** нашого дослідження є використання інструменту корпусної лінгвістики відкритого доступу «Voyant Tools» (https://voyant-tools.org), що є веб-платформою, яка виконує аналіз тексту.

Програма дозволяє здійснити науковий розбір тексту, виділити у корпусі частоту вживання лексичних одиниць, створити діаграми, схеми тощо. Також «Voyant tools» можна використовувати для аналізу онлайн-текстів або текстів, завантажених користувачами.

Здійснюючи аналіз вибірки новин англійського видання «The Guardian» тематики «Вірусні захворювання» за допомогою веб-програми «Voyant Tools», ми зафіксували, що цей корпус налічує 2 909 слів і 955 словосполучень. Із них виокремлено 44 ключові лексичні одиниці, які стосуються тематики «Коронавірус», визначено частоту вживання та використання слів у реченні. Серед найбільш вживаних слів: *virus (29); coronavirus (27); people (27); health (20); cases (18); china (17); wuhan (16); confirmed (14); symptoms (12); new (11); spread (11); chinese (10); ago (9); infected (9); said (9); tested (9); uk (9); government (8); hospital (8); patients (8); far (7); level (7); public (7); 1d (6); city (6); flu (6); highest (6); hubei (6); medical (6); negative (6); number (6); province (6); reported (6); risk (6); situation (6); weeks (6); centre (5); cough (5); died (5); doctor (5); emergency (5); hong (5); hospitals (5); news (5); provinces (5); world (5); year (5); affected (4); announced (4); appears (4); breathing (4); coronaviruses (4); countries (4); crisis (4); days (4); human (4); infection (4); kong (4); organization (4)*

За сегментною діаграмою простежується топ 5 ключових слів тематики «Коронавірус» (рис. 1).

Демонструючи результат роботи програми, котра дозволяє зробити аналіз лексем новин французького видання «Le Parisien», виводимо показники, за якими статистика корпусу тексту налічує 4 293 слів та 1 254 словосполучень, із яких можна виокремити частоту використання слів від найбільш частого використання слова у тексті до найменшого, а саме: *cas (31); France (29); personnes (28); coronavirus (27); chine (25); virus (25); Wuhan (20); janvier (19); parisien (18); chinois (17); 2020 (16); santé (15); morts (12); symptômes (12); afp (11); Paris (11); patients (11); samedi (11); ville (10); bordeaux (9); jours (9); matin (9); vendredi (9); 2019 (8); contact (8); l'épidémie (8); ministère (8); savoir (8); annoncé (7); autorités (7); charge (7); dimanche (7); faire (7); faut (7); français (7); l'instant (7); modifié (7); mortalité (7); premier (7); taux (7); actualités (6); adresse (6); collectée (6); commerciales (6); d'autres (6); hôpital (6); l'actu (6); l'actualité (6); l'essentiel (6); l'hôpital (6); m'inscris (6); mail (6); maladie (6); newsletter (6); nouveau (6).*

Як бачимо, найбільш вживані лексеми-номінанти у новинах французькою мовою є: *cas (випадок); france (Франція); personnes (люди); coronavirus (коронавірус); chine (Китай); virus (вірус)* (рис. 2).

Здійснено також аналіз корпусу текстів українських новин, згідно з яким програма «Voyant Tools» продемонструвала такі ключові слова: *у (96); в (93); з (73); на (65); і (45); що (40); коронавірус (38); також (37); Китаю (35); не (32); до (30); Китаї (29); та (27); про (26); коронавірусу (24); від (23); здоров'я (23); охорони (22); за (21); коментувати (18); коронавірусом (18); людей (18); це (16); із (16); через (15); який (15); а (14); для (14); закрили (14); зараження (14); зафіксували (14); осіб (14); повідомляє (14); фото (14); читайте (14); ВООЗ (13); він (13); вірусу (13); захворювання (13); нового (13); січня (13); які (13); його (12); карантин (12); кількість (12); лікарні (12); поки (12); китайського (11); влада (10); й (10); місто (10); місті (9); року (9); ухань (9); Франції (9); час (9); ще (9); як (9).*

Крім того, можна застосувати для візуалізації термінів, що найчастіше зустрічаються в корпусі текстів, «Хмару слів «Cirrus (Циррус)»». (рис. 3).

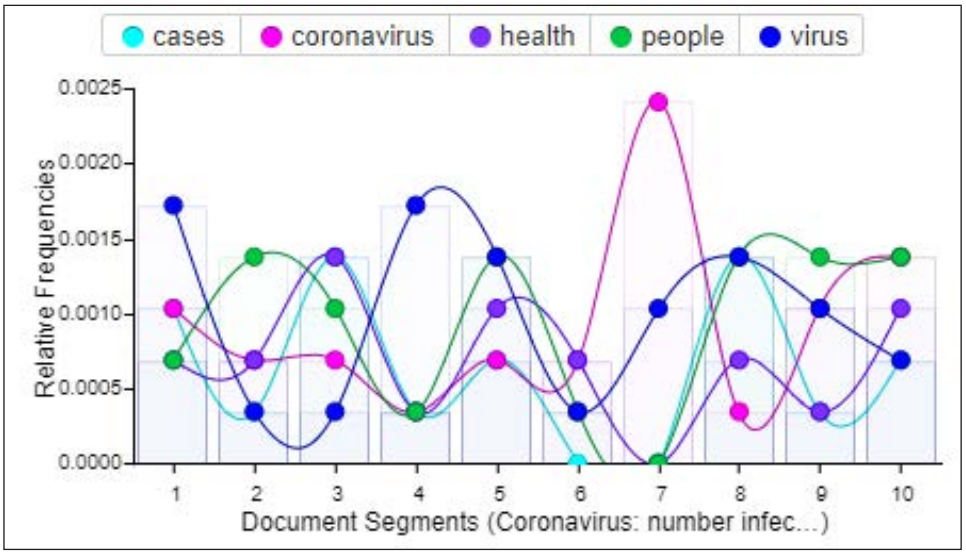

Рис. 1. Діаграма частоти вживання слів тематики «Коронавірус» в англійській мові

Результати показали, що найбільше застосовані в повідомленнях українською мовою прийменники *у, в, на*, сполучник *і*. Другу позицію займають такі слова: *коронавірус, Китай, зараження*. На третій позиції зазначені слова, які французькими й англійськими ЗМІ були виокремлені як топ-5 найуживаніших – *Ухань, Франція, вірус, захворювання*.

Таким чином, використовуючи інструмент корпусної лінгвістики «Voyant Tools», можемо виокремити ключові лексеми корпусу текстів, провести системні, широкі за кількістю охопленого емпіричного мовного матеріалу дослідження мови.

Отже, завдяки розвитку сучасних інформаційних технологій з'явилася нова сучасна лінгвістична дисципліна – корпусна лінгвістика, для якої характерний аналіз інтерпретаційних процесів, корпусу текстів, лексичний розбір. Однією з переваг дисципліни є оперування великим обсягом тексту, виклад аналізу корпусу тексту в електронному варіанті, що дозволяє економити час.

У перспективі подальшого розгляду і вивчення цієї теми доцільно дослідити та порівняти лексеми-номінанти концептів СМЕРТЬ, ВІРУС і встановити спільне та відмінне в лінгвістичному аналізі у трьох мовах.

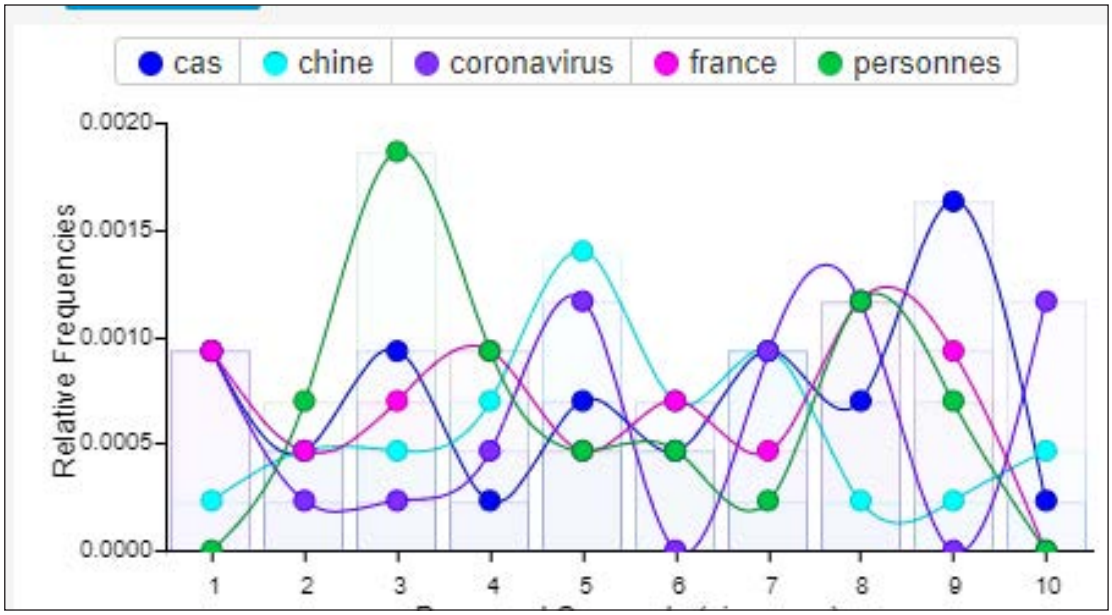

Рис. 2. Діаграма частоти вживання слів тематики «Коронавірус» у французькій мові

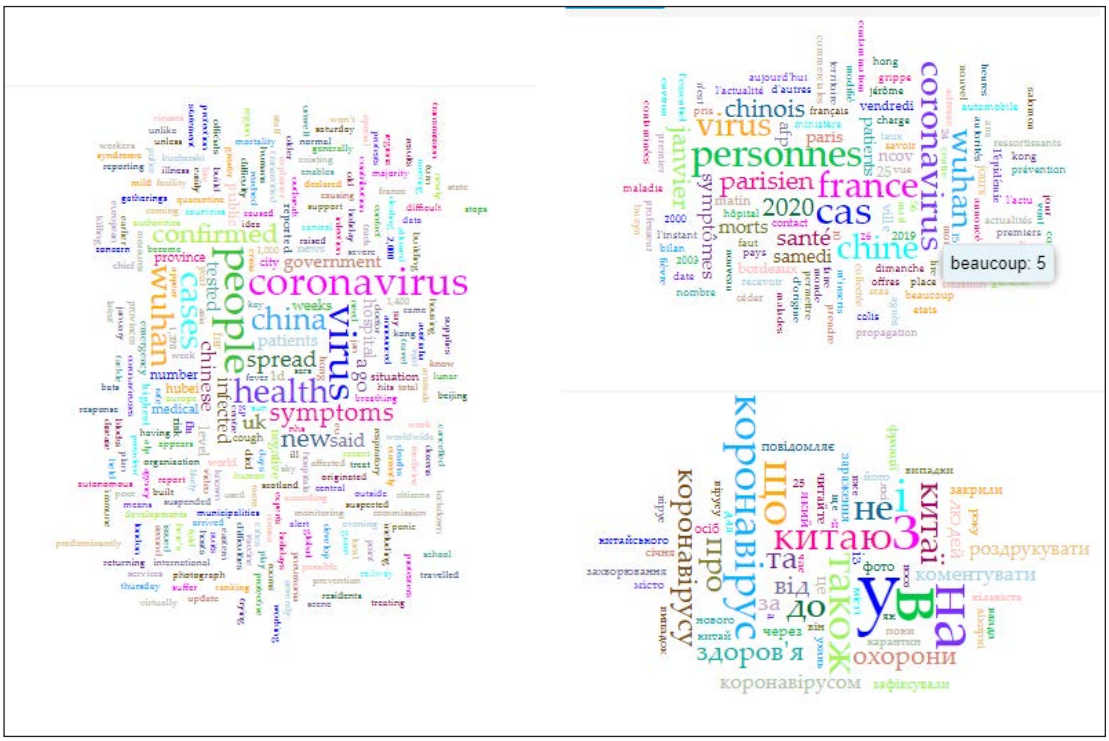

Рис. 3. «Хмара слів «Cirrus (Циррус)»»

Література:

1. Демська О.М. Текстовий корпус: ідея іншої форми. Київ : ВПЦ НаУКМА, 2011. 282 с.
2. Демська-Кульчицька О. Основи Національного корпусу української мови. Київ : Наук. видання ІУМ НАН України, 2005. 219 с.
3. Захаров В.П. Корпусная лингвистика : учебно-методическое пособие. Санкт-Петербург, 2005. 48 с.
4. Корпусна лінгвістика / В.А. Широков, О.В. Букагов, Т.О. Грязнухіна та ін. Київ : Довіра, 2005. 471 с.
5. Махачашвілі Р.К. Открытая вербальная е-среда: исследовательские принципы и ИКТ инструменты. *Open Educational E-environment of modern university*. 2016. № 2. С. 27–33.
6. Le Parisien. URL: http://www.leparisien.fr/faits-divers/4/.
7. The Guardian. URL: https://www.theguardian.com/world.
8. Газета по-українськи. URL: https://gazeta.ua/news/np.
9. MacEnery T., Hardie A. Corpus Linguistics: Method, Theory and Practice. Cambridge University Press, 2012. 294 p.
10. Voyant Tools. URL: https://voyant-tools.org.

Makhachashvili R., Bilyk K. Corpus linguistic study of heading "breaking news" in French, English and Ukrainian languages

Summary. The article is devoted to the research of the corpus of texts of the "Extraordinary News" section in French, English and Ukrainian. When compiling a thematic innovative glossary of extraordinary viral diseases topics, we used the corpus linguistics tool "Voyant Tools". Key tokens have been identified, a synonym has been established and the frequency of use of words and phrases of the above subject has been analyzed.

The purpose of the article is to get acquainted with the basic concepts of corpus linguistics, describe the possibilities of working with corpus databases in linguistic research, analysis of corpus linguistics tools "Voyant Tools", compilation of an innovative glossary of texts "Emergency News" topics in "Viral News" in French, English, Ukrainian. Corpus linguistics is an extremely important branch of modern linguistic science. The presented work contains a practical analysis of the body of news texts. The focus of the corpus research is the linguistic person, that is, his linguistic activity, mass communication, the problem of his description. That is why the tools of corpus linguistics under study pay attention to concordance, relative and absolute word frequency, keywords, word compatibility, multicomponent groups, contextual use of words.

The realization of this goal implies the following tasks:

– to research news articles of the "Extraordinary News" section in French, English and Ukrainian.

– select vocabulary units that are specific to the topic of "Viral diseases";

– analyze the semantic specificity of the lexical material selected.

In a study of news reports, it was found that the most commonly used tokens are: virus, infection, spread, humans, China, because these words give a schematic picture to the reader of what the language in the newspaper covers. In view of further consideration and study of this topic, it is advisable to research and compare the token-nominees of the concepts of DEATH, VIRUS and to establish common and different in linguistic analysis in three languages.

Key words: corpus linguistics, breaking news, lexical items, text, frequency, coronavirus, Voyant Tools.