DOI: https://doi.org/10.32839/2304-5809/2019-12-76-65
УДК 811.161.2'322.2.2.

Яйченя Ю.П., Кульчицький І.М.
Національний університет «Львівська політехніка»

ЛЕМАТИЗАЦІЯ ТВОРУ Р. ІВАНИЧУКА «ЧЕРЛЕНЕ ВИНО»: СТАТИСТИЧНИЙ АСПЕКТ

Анотація. У статті представлений метод обробки корпусу текстів — лематизація в аспекті статистичних досліджень. Зазначено, що лематизацію широко використовують в алгоритмах пошукових систем, вона дозволяє знайти більшу кількість результатів, а не тільки результати за запитом введеної словоформи. Проведено статистичний аналіз тексту Р. Іваничука «Черлене вино» за ознаками, як-от: розподіл за частинами мови словоформ тексту, розподіл за частинами мови лем словника тексту, розподіл значень частотності словоформ тексту, розподіл значень частотності лем словника тексту. У дослідженні додано загальні коефіцієнти слів в тексті, а також загальні коефіцієнти тексту. Об'єктом аналізу було обрано твір Р. Іваничука «Черлене вино». Предметом аналізу є лематизації твору Р. Іваничука «Черлене вино» і проведення статистичного аналізу результатів.

Ключові слова: корпусна лінгвістика, статистика, лематизація, словоформа, лема.

Yaychenya Yuriy, Kulchytskyi Ihor
Lviv Polytechnic National University

LEMATIZATION OF R. IVANICHUK'S WORK "RED WINE" IN STATISTICAL ASPECTS

Summary. The article presents a method for processing the corpus of texts by the method of lemmatization in the aspect of statistical research. Lemmatization is widely used in search engine algorithms. So, it allows you to find a larger number of results, and not just the results for the query word only in the form in which it was entered. According to the results of lemmatization of the work of R. Ivanychuk "Red wine", a statistical analysis of it was carried out according to the following criteria: distribution of speech parts of the text word form, distribution of the speech parts by the text dictionary part, distribution of the frequency values of the word text forms, distribution of the frequency values by the text dictionary part and added common text word odds, as well as general text odds. It is worth noting that the calculation of the lemmatization result was carried out according to the number of word usage, word forms and word lemma, accrual of legomena for word forms and word lemma, the number of word forms and lem taken ten or more times, the number of characters of the extended alphabet in the text, as well as the number of sentences in the text. The object of analysis was selected by R. Ivanychuk's work "Red wine". The subject of analysis is the implementation of the lemmatization of the work of R. Ivanichuk "Red wine", as well as the statistical analysis of the results of this lemmatization. The first task of morphological analysis (lemmatization) is to provide a definition of the normal form of the word from which the word form was formed, and a list of parameters that are part of this word form. The second in this case is to search for the desired word form in the dictionary and copy the morphological information corresponding to the found word form into the program.Lemmatization is widely used in search engine algorithms. So, it allows you to find a larger number of results, and not just the results for the query word only in the form in which it was entered. Lemmatization is also used when checking the uniqueness of text, web development, programming and compiling a semantic core. The practical application of statistical data was investigated on the basis of the data of lemmatization.

Keywords: corpus linguistics, statistics, lemmatization, word form, lemma.

Постановка проблеми. Лематизацію широко використовують в алгоритмах пошукових систем. Так, вона дозволяє знайти більшу кількість результатів, а не тільки результати за запитом слова тільки в тій формі, в якій воно було введено. Її використовують при перевірці унікальності тексту, веб-розробках, програмуванні та складанні семантичного ядра. У статті пропонуємо розглянути та застосувати основні підходи для аналізу та обробки текстової інформації.

Аналіз досліджень і публікацій. Для сьогодення характерна величезна кількість студій пов'язаних зі статистикою, базою знань, пошуковими системами тощо. Обробка тексту – є одним з найнеобхідніших завдань для класифікації документів [1; 5, с. 284]. Попередня обробка тексту використовується насамперед для автоматичної анотації документів. Про це детально можна дізнатись із наукових праць Л.Н. Шавердної, А.Ф. Осики, В.П. Леонова, які дають деталізований огляд наявних спроб автоматизації анотування текстів та показують в загальному характеристику проблеми у дослідженнях

Метою пропонованої статті є опис процесу лематизації тексту в статистичному аспекті. Об'єктом аналізу є твір Р. Іваничука «Черлене вино». **Предметом аналізу** є здійснення лематизації твору та аналіз результатів дослідження.

Виклад основного матеріалу. При створенні корпусу використовують ряд процедур і програм, таких як: токенізація, лематизація, стемінг та синтаксичний аналіз [1, с. 38–41]. Токенізація – це розбиття потоку символів природної мови на окремі значущі одиниці (токени, словоформи). Лематизація – це процес формування початкової форми слова, виходячи з його інших словоформ. Стеммінг полягає в пошуку основи слова.

Завдання морфологічного аналізу полягає в забезпеченні визначення початкової форми слова. При розгляді поняття використовують його початкову форму. Наприклад, візьмемо слово: *столів* – [*стілець*]. Для нього є різні форми: *стільцем, стільця, стільця* і так далі. Кожна форма відповідає ряду параметрів і характеристик, наприклад, *рід*, *число* чи *відмінок*, які характеризують дану

словоформу. Крім того, кожне слово відповідає певній частині мови [7, с. 101–108].

У конкретному місці слово із заданою частиною мови в певній формі в ході машинної обробки даних необхідно формалізувати. Варто відзначити також, що подібна різноманітність створює проблеми при проведенні аналізу тексту. Проблема полягає в обробці всіх словоформ замість обробки єдиного слова. Уникнути проблемної ситуації дозволяють етапи морфологічного аналізу і синтезу.

Синтаксичний аналіз – це процес аналізу синтаксичної структури тексту або частини тексту, заснований на порівнянні лінійної послідовності токенів (слів, токенів) мови з його формальної граматики.

У нашому дослідженні для створення корпусу тексту твору Р.Іваничука «Черлене вино» використовуємо процедуру лематизації. Для полегшення роботи з словами, визначенням частини мови та її граматичних категорій використовуємо два он-лайн словники: «Граматичний словник української літературної мови. Флексія» (опублікований в 2011 році і доступний на Лінгвістичному порталі (http://www.mova.info/grmasl.aspx)) і також «Словники України» (опублікований на порталі http://lcorp.ulif.org.ua/dictua/). Ці он-лайн ресурси допоможуть під час визначення частини мови, опису за граматичними категоріями.

В процесі лематизації тексту «Черлене вино» було обраховано наступні характеристики тексту, що представлені у таблиці 1.

Далі було підраховано показники, як-от: як розподіл за частинами мови словоформ тексту (див. таблицю 2), розподіл за частинами мови лем словника тексту (див. Таблицю 3), розподіл значень частотності словоформ тексту (див. таблицю 4), розподіл значень частотності лем словника тексту (див. таблицю 5), загальні коефіцієнти слів тексту (див. таблицю 6), загальні коефіцієнти тексту (див. таблицю 7). Всі дані внесені у таблиці.

Висновки. Статистичні методи у мовознавстві допомагають правильно організувати лінгвістичні спостереження, отримати об'єктивні дані, незалежні від суб'єктивного сприйняття дослідника, забезпечити надійність, точність, достовірність висновків. Головним завданням статистичної лінгвістики є застосування математичних методів для розкриття закономірностей функціонування одиниць мови у мовленні, а також у встановленні закономірностей будови тексту.

У статистичному дослідженні ідіостилю автора були використані кількісні методи, тобто підрахунок частоти вживання словоформ, лем, певних частин мови, а також і статистичні методи, які використовують різні формули для виявлення правил

Таблиця 1

Загальні характеристики тексту Р. Іваничука «Черлене вино»

Кількість слововживань:	68776
Кількість словоформ:	19551
Кількість лем:	11383
Нарах legomena для словоформ:	12958
Кількість словоформ, що вживаються десять і більше разів:	782
Нарах legomena для лем:	5866
Кількість лем, що вживаються десять і більше разів:	943
Кількість символів розширеного алфавіту у тексті:	365082
Кількість речень у тексті:	4274

Таблиця 2

Розподіл за частинами мови словоформ тексту

Частина мови	Кількість	Частота	Кількість вживань	Частота вживань
	2	0,000102297	2	2,91E-05
вигук	40	0,002045931	107	0,001555775
дієприслівник	343	0,01754386	489	0,007110038
дієслово	5754	0,294307197	12124	0,176282424
дієприслівник	1	5,11E-05	1	1,45E-05
займенник	381	0,019487494	9662	0,140485053
іменник	8108	0,414710245	20441	0,297211236
не визначено	9	0,000460335	9	0,00013086
невідмінюване	13	0,000664928	207	0,003009771
не передбачено	4	0,000204593	10	0,0001454
прийменник	67	0,003426935	7403	0,107639293
прийменникова сполука	1	5,11E-05	1	1,45E-05
прикметник	3819	0,195335277	5461	0,079402699
прислівник	765	0,039128433	3562	0,051791323
прислівникова сполука	1	5,11E-05	1	1,45E-05
сполучник	63	0,003222342	6129	0,089115389
частка	82	0,004194159	2869	0,041715133
числівник	59	0,003017748	259	0,003765849
Разом	19551	1	68776	1

Таблиця 3

Розподіл за частинами мови лем словника тексту

Частина мови	Кількість	Частота	Кількість вживань	Частота вживань
	2	0,000175701	2	2,91E-05
вигук	39	0,003426162	107	0,001555775
дієприслівник	343	0,030132654	489	0,007110038
дієслово	3170	0,278485461	12124	0,176282424
займенник	127	0,011156988	9662	0,140485053
іменник	4540	0,398840376	20441	0,297211236
не визначено	9	0,000790653	9	0,00013086
невідмінюване	13	0,001142054	207	0,003009771
не передбачено	2	0,000175701	10	0,0001454
прийменник	66	0,00579812	7403	0,107639293
прийменникова сполука	1	8,79E-05	1	1,45E-05
прикметник	2097	0,184222086	5461	0,079402699
прислівник	760	0,06676623	3562	0,051791323
прислівникова сполука	1	8,79E-05	1	1,45E-05
сполучник	61	0,005358868	6129	0,089115389
частка	82	0,007203725	2869	0,041715133
числівник	30	0,002635509	259	0,003765849
Разом	11383	1	68776	1

Таблиця 4

Розподіл значень частотності словоформ тексту

Значення	Кількість	Частота	Значення	Кількість	Частота
1	12958	0,6627794	27	3	0,000153445
2	2958	0,15129661	28	7	0,000358038
3	1156	0,05912741	29	3	0,000153445
4	635	0,03247916	30	4	0,000204593
5	361	0,01846453	31	3	0,000153445
6	267	0,01365659	32	3	0,000153445
7	180	0,00920669	33	5	0,000255741
8	149	0,00762109	34	1	5,11E-05
9	105	0,00537057	35	2	0,000102297
10	90	0,00460335	36	1	5,11E-05
11	67	0,00342693	37	2	0,000102297
12	49	0,00250627	38	4	0,000204593
13	55	0,00281316	39	5	0,000255741
14	36	0,00184134	40	6	0,00030689
15	37	0,00189249	41	1	5,11E-05
16	34	0,00173904	42	3	0,000153445
17	23	0,00117641	43	4	0,000204593
18	22	0,00112526	44	2	0,000102297
19	16	0,00081837	45	2	0,000102297
20	19	0,00097182	46	5	0,000255741
21	19	0,00097182	47	1	5,11E-05
22	9	0,00046033	48	1	5,11E-05
23	8	0,00040919	49	2	0,000102297
24	14	0,00071608	50	2	0,000102297
24	1	5,11E-05	1920	1	5,11E-05
25	1	5,11E-05	**Разом**	**19343**	**1**
26	1	5,11E-05			

Таблиця 5

Розподіл значень частотності лем словника тексту

Значення	Кількість	Частота	Значення	Кількість	Частота
1	5866	0,515329878	27	10	0,000878503
2	2033	0,178599666	28	9	0,000790653
3	976	0,085741896	29	13	0,001142054
4	547	0,048054116	30	8	0,000702802
5	343	0,030132654	31	6	0,000527102
6	237	0,020820522	32	7	0,000614952
7	197	0,01730651	33	10	0,000878503
8	137	0,012035492	34	10	0,000878503
9	104	0,009136432	35	7	0,000614952
10	98	0,00860933	36	2	0,000175701
11	78	0,006852324	37	1	8,79E-05
12	58	0,005095318	38	6	0,000527102
13	57	0,005007467	39	9	0,000790653
14	41	0,003601862	40	3	0,000263551
15	46	0,004041114	41	3	0,000263551
16	34	0,00298691	42	7	0,000614952
17	29	0,002547659	43	7	0,000614952
18	29	0,002547659	44	5	0,000439252
19	23	0,002020557	45	2	0,000175701
20	26	0,002284108	46	2	0,000175701
21	17	0,001493455	47	3	0,000263551
22	8	0,000702802	48	4	0,000351401
23	15	0,001317755	49	5	0,000439252
24	24	0,002108407	50	7	0,000614952
25	22	0,001932707	**Разом**	**11258**	**1**
26	15	0,001317755			

Таблиця 6

Загальні коефіцієнти слів тексту

Багатство словника:	0,17
Середня повторюваність слова у тексті:	6,04
Коефіцієнт винятковості для словоформ:	0,19
Коефіцієнт винятковості для лем:	0,52
Коефіцієнт концентрації словника для словоформ:	0,01
Коефіцієнт концентрації словника для лем:	0,08
Автоматичний коефіцієнт читабельності:	11,62

Таблиця 7

Загальні коефіцієнти тексту

Коефіцієнт лексичної щільності:	0,2
Коефіцієнт іменних означень:	3,74
Коефіцієнт дієслівних означень:	0,28
Коефіцієнт номінативності:	1,62
Коефіцієнт дієслівності:	0,18
Коефіцієнт логічної зв'язності:	3,17
Коефіцієнт емболії мовлення:	0,04

розподілу мовних одиниць у мовленні, для виміру зв'язків між мовними елементами, для встановлення тенденцій у розвитку та функціонуванні мови та для встановлення залежності між якісними та кількісними характеристиками мови.

Отже, маючи в розпорядженні числові показники, можна створити інструменти для ефективного аналізу текстів за допомогою лематизації.

Безумовно, створення статистичних (частотних) словників полягає в їхньому подальшому практичному застосуванні, а також може слугувати основою в різних сферах лінгвістичного аналізу. Оскільки українська мова має складну морфологію і велику кількість словоформ, було досить складно створити модель, яка б ефективно узагальнювала введені словоформи.

Список літератури:

1. Бук С. Квантитативна параметризація текстів Івана Франка: спроба проекту. *Іван Франко: Студії та матеріали*. Львів, 2010. URL: http://arxiv.org/abs/1005.5466 (дата звернення: 01.12.2019).
2. Бук С., Ровенчак А. Частотний словник повісті І. Франка «Перехресні стежки». *Стежками Франкового тексту (комунікативні, стилістичні та лексичні виміру роману «Перехресні стежки»)*. Львів : Видав. центр ЛНУ імені Івана Франка, 2007. С. 138–369.
3. Гандзій О.А. Публіцистика Романа Іваничука: проблематика і поетика : автореф. дис. … канд. філол. наук : спец. 10.01.01 «Українська література». Івано-Франківськ, 2011. С. 20.
4. Загнітко А., Данилюк І. Корпус текстів граматичної службовості. *Прикладна лінгвістика та лінгвістичні технології. Довіра*. Київ, 2013. С. 102–112.
5. Демська О.М. Текстовий корпус: ідея іншої форми. Нац. ун-т «Києво-Могилянська академія». Київ : Вид. дім «Києво-Могилянська академія», 2011. С. 284.
6. Крупа М. Лінгвістичний аналіз художнього тексту. Тернопіль, 2005. С. 17–43.
7. Кульчицький І.М. Технічні аспекти функціонування текстів у електронному інформаційному просторі. *Український інформаційний простір. Число 2*. Київський національний університет культури і мистецтв. Київ, 2014. С. 101–108.
8. Перебийніс В.І. Статистичні методи для лінгвістів : навч. посіб. Вінниця : Нова книга, 2001. 268 с.
9. Перебийніс В.С., Муравицька М.П., Дарчук Н.П. Частотні словники та їх використання. Київ : Наукова Думка, 1985. 204 с.
10. Проект ВЕСУМ. URL: https://github:https://github.com/brown-uk/dict_uk (дата звернення: 01.12.2019).
11. Роман Іваничук : бібліографічний покажчик / укладач Л. Панів. Львів : Вид. центр ЛНУ ім. Івана Франка, 2011. 405 с.
12. Тищенко В. Частота частин мови в різних функціональних стилях сучасної української мови. *Питання структурної лексикології*. Київ : Наукова думка, 1970. С. 215–224.

References:

1. Buk, S. (2010). Kvantytatyvna parametryzatsiia tekstiv Ivana Franka: sproba proektu. *Ivan Franko: Studii ta materialy*. Lviv. URL: http://arxiv.org/abs/1005.5466 (accessed 01.12.2019).
2. Buk, S., & Rovenchak, A. (2007). Chastotnyi slovnyk povisti I. Franka "Perekhresni stezhky". *Stezhkamy Frankovoho tekstu (komunikatyvni, stylistychni ta leksychni vymiru romanu "Perekhresni stezhky")*. Lviv: Vydav. tsentr LNU imeni Ivana Franka, s. 138–369.
3. Handzii, O.A. (2011) Publitsystyka Romana Ivanychuka: problematyka i poetyka: avtoref. dys. … kand. filol. nauk: spets. 10.01.01 «Ukrainska literatura». Ivano-Frankivsk, s. 20.
4. Zahnitko, A., & Danyliuk, I. (2013). Korpus tekstiv hramatychnoi sluzhbovosti. *Prykladna linhvistyka ta linhvistychni tekhnolohii*. Kyiv: Dovira, s. 102–112.
5. Demska, O.M. (2011). Tekstovyi korpus: ideia inshoi formy. Nats. un-t «Kyievo-Mohylianska akademiia». Kyiv: Vyd. dim «Kyievo-Mohylianska akademiia», s. 284.
6. Krupa, M. (2005). Linhvistychnyyi analiz khudozhnoho tekstu. Ternopil, s. 17–43.
7. Kulchytskyi, I.M. (2014) Tekhnichni aspekty funktsionuvannia tekstiv u elektronnomu informatsiinomu prostori. *Ukrainskyi informatsiinyi prostir. Chyslo 2. Kyivskyi natsionalnyi universytet kultury i mystetstv*. Kyiv, s. 101–108.
8. Perebyinis, V.I. (2001). Statystychni metody dlia linhvistiv : navch. posib. Vinnytsia: Nova knyha, 268 s.
9. Perebyinis, V.S. Muravytska, M.P., & Darchuk, N.P. (1985). Chastotni slovnyky ta yikh vykorystannia. Kyiv: Naukova Dumka, 204 s.
10. Proekt VESUM. URL: https://github.com/brown-uk/dict_uk (accessed 01.12.2019).
11. Paniv, L. (2011). Roman Ivanychuk: bibliohrafichnyi pokazhchyk. Lviv: Vyd. tsentr LNU im. Ivana Franka, 405 s.
12. Tyshchenko, V. (1970). Chastota chastyn movy v riznykh funktsionalnykh styliakh suchasnoi ukrainskoi movy. *Pytannia strukturnoi leksykolohii*. Kyiv: Naukova dumka, s. 215–224.