МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»

МЕТОДИЧНІ ВКАЗІВКИ

до виконання лабораторних робіт

з курсу «Корпусна лінгвістика»
для студентів спеціальності
«Прикладна та комп'ютерна лінгвістика»

Частина 2

Харків 2021

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»

МЕТОДИЧНІ ВКАЗІВКИ

до виконання лабораторних робіт
з курсу «Корпусна лінгвістика»
для студентів спеціальності
«Прикладна та комп'ютерна лінгвістика»

Частина 2

Затверджено
редакційно-видавничою
радою університету,
протокол № 2 від 29.06.2021 р.

Харків
НТУ «ХПІ»
2021

Методичні вказівки до виконання лабораторних робіт з курсу «Корпусна лінгвістика» для студентів спеціальності «Прикладна та комп'ютерна лінгвістика». Частина 2 / уклад. Н. Ф. Хайрова, С. В. Петрасова, О. О. Оробінська. – Харків : НТУ «ХПІ», 2021. – 51 с.

Укладачі: Н. Ф. Хайрова
С. В. Петрасова
О. О. Оробінська

Рецензент Н. В. Шаронова

Кафедра інтелектуальних комп'ютерних систем

ВСТУП

Поняття корпусної лінгвістики виникло в 80-х роках минулого століття. В даний час під **корпусною лінгвістикою** розуміється розділ мовознавства, що займається розробкою, створенням і використанням текстових корпусів. Водночас корпусні лінгвістика є швидше не складовою частиною загальної лінгвістики, а являє собою методологію або способи використання конкретних ресурсів, що представляють великі обсяги текстових даних. Отже, корпус та спеціальні програмні засоби роботи з цим корпусом є спеціалізованим інструментом лінгвістичних досліджень. Таке спеціалізоване програмне забезпечення, що використовується для дослідження великих обсягів текстових даних, які зібрані у корпусі, називається **concordancer.**

Традиційно в корпусній лінгвістиці можна виділити два напрями вивчення: створення корпусів і дослідження мовних закономірностей за допомогою корпусних методів на базі створених корпусів. Проте, нерідко розробники корпусів проводять одночасно і власні лінгвістичні дослідження. Таким чином корпусна лінгвістика передбачає одночасне використання лінгвістичних знань та знань комп'ютерних технологій, що і зумовлює використання даної дисципліни в навчальному плані студентів спеціальності «Прикладна та комп'ютерна лінгвістика».

Дані методичні рекомендації спрямовані на отримання навичок роботи з усіма складовими корпусних завдань, а саме: створення власних корпусів, опанування різних видів корпусної розмітки, використання Інтернету для корпусних досліджень та вживання різних статистичних методів при роботі з корпусами.

В другу частину **Методичних вказівок** до виконання лабораторних робіт з курсу «Корпусна лінгвістика» включено чотири лабораторні роботи, виконання яких дозволяє отримати навички роботи з системою семантичної розмітки USAS, аналізу корпусу в online середовищі CQPweb та обробки кластерів корпусів конкордансером AntConc.

Лабораторна робота 4

СЕМАНТИЧНА РОЗМІТКА

4.1. Автоматична семантична розмітка системи USAS

Система ***USAS (UCREL (University Centre for Computer Corpus Research on Language) Semantic Analysis System)*** призначена для автоматичного семантичного аналізу тексту. Система розроблена на базі лексикону сучасної англійської мови (Tom McArthur's Longman Lexicon of Contemporary English) з використанням підходів комп'ютерної лінгвістики і NLP. Система семантичної класифікації пропонує найбільш підходящу тезаурусну класифікацію смислу слова. Базова множина семантичних полів включає 21 домен (Додаток Г), які в свою чергу таксономічно включають 232 категорії (Додаток Д). Семантичний тег є більш загальним, ніж категорія тезауруса і визначає смисл слова не так вузько і точно, як це зроблено в тлумачному словнику.

Наприклад, слово "bank" буде розмічено наступним чином (рис. 4.1).

```
0000001 002  -----   -----
0000003 010  NN1     bank                    I1/H1 I1.1/I2.1c W3/M4 A9+/H1 O2 M6
```

Рисунок 4.1 – Семантична розмітка

Мітка I1 позначає категорію грошей і комерцію для фінансового домену, тег W3 позначає географічний термін, H1 – архітектура / тип будови. При цьому більш докладного тлумачення семантична розмітка не робить. Приблизна точність роботи системи 91%.

Починаючи з 2013 року такі тестовані системи автоматичної семантичної класифікації розроблені також для голландської, китайської, італійської, португальської та іспанської мов. При цьому здійснюється автоматичний переклад лексики мови, можлива багатозначність якої призводить до помилок, які необхідно перевіряти "вручну".

Для російської мови розроблено (перекладено) множину семантичних тегів (Додаток Е). Крім того існує візуальне подання ієрархічної множини семантичних тегів (рис. 4.2).

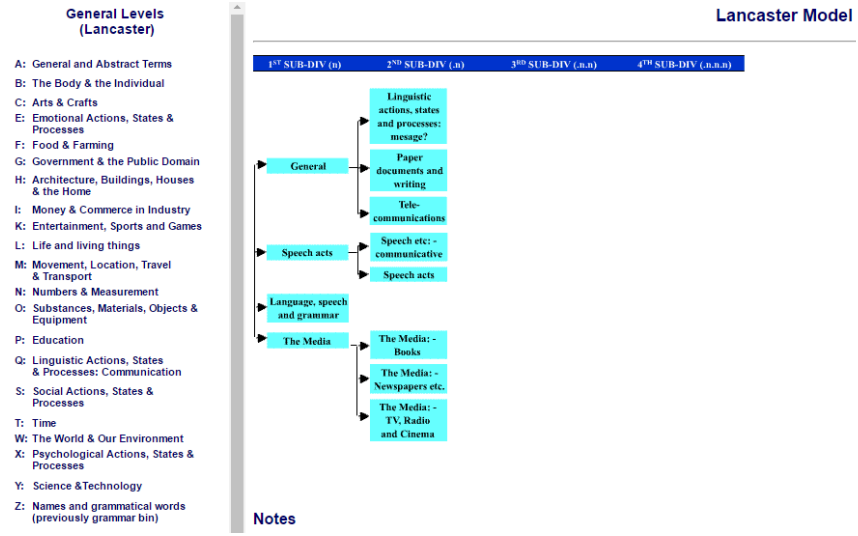


Рисунок 4.2 – Ієрархія семантичних тегів

Результат семантичного тегування системи USAS може бути видано в трьох різних форматах: горизонтальному, вертикальному і в XML-форматі. У вертикальному форматі (рис. 4.3) кожне слово розміщується в одному рядку зі своїм POS і семантичним тегами.

```
0000001 002  -----    -----
0000003 010  NN1      work                  I3.1 A1.1.1 Q4 K2 C1
0000003 020  VBZ      is                    Z5 A3+
0000003 030  VVN      expected              X2.6+ B1
0000003 040  TO       to                    Z5
0000003 050  VVI      start                 T2+ E5- E3-
0000003 060  IF       for                   Z5
0000003 070  AT       the                   Z5
0000003 080  NN1      road                  M3 X4.2
0000003 090  MD       next                  T1.1.3[i2.2.1 N4 T1.3[i1.2.1 P1/S2mf[i1.2.1
0000003 100  NNT1     year                  T1.1.3[i2.2.2 T1.3 P1c T1.3[i1.2.2 P1/S2mf[i1.2.2
```

Рисунок 4.3 – Вертикальний формат семантичної розмітки

Наприклад, слово "work" розмічено як I3.1 – *Work and employment: Generally*, слово "start" розмічено як час T2+ *Time: Beginning and ending*, слово "road" розмічено як M3 – *Vehicles and transport on land*. Крім того при вертикальній розмітці показані альтернативні теги, які є менш ймовірними в даному контексті. Такі альтернативні варіанти виводяться далі правіше по рядку. Наприклад, слово "work" може мати семантичні значення: A1.1.1 – *General actions, making etc.*, Q4 – *The Media*, K2 – *Music*

and related activities, C1 – *Arts and crafts*. Інший приклад, слово "is" може бути розмічено як A3+ – *Being*, коли воно є основним дієсловом в реченні, але в даному випадку – це допоміжне дієслово (*Z5 – Grammatical bin*).

Вирази, що складаються з декількох слів, показані з додатковим позначенням після квадратної дужки, що відкривається. Наприклад, вираз "next year" позначено як T1.1.3[i2.3.1 і T1.1.3[i2.3.2, що позначає *Time: General: Future*. При цьому одиночне слово "year" було б розмічено як T1.3 – *Time: Period*.

Антонімічні концептуальні класифікатори використовують перед тегом знак +/–. Наприклад, A5.1+ (good) і A5.1– (bad) належать до однієї і тієї ж категорії A5.1, але вони чітко розмічені. Вищий і найвищий ступені отримують подвійні і потрійні +/– маркери відповідно.

Закритий клас слів, таких як прийменники або імена власні позначаються тегом, що починається на "Z", який позначає граматичну категорію *bin*. Така категорія не може бути підрахована при остаточному статистичному аналізі (за винятком власних назв), але може бути легко знайдена при необхідності.

Для отримання більш короткої форми тегування слід вибрати горизонтальний формат (рис. 4.4).

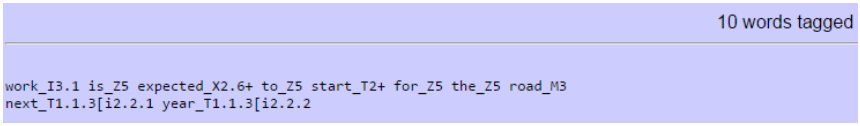


Рисунок 4.4 – Горизонтальний формат семантичної розмітки

У горизонтальному форматі відображається тільки найбільш ймовірний семантичний тег, який записується в кінці слова, після знака підкреслення. Таке позначення легко може бути прочитано і оброблено конкордансерами. Крім того можна вибрати XML-стиль, що додає в елемент *w* після POS-розмітки додатковий атрибут *sem*, значенням якого є семантична мітка (рис. 4.5).

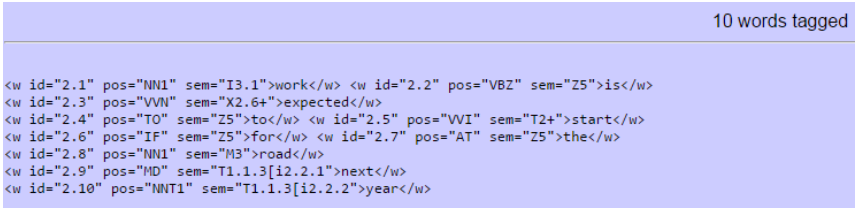


Рисунок 4.5 – XML-формат семантичної розмітки

Завдання до лабораторної роботи 4

1. На сайті http://ucrel.lancs.ac.uk/usas/ ознайомитися з описом системи автоматичного семантичного тегування.

2. Провести семантичну розмітку фрагмента англійського тексту (горизонтальна, вертикальна і XML-розмітка).

3. Описати значення тегів.

4. Використовуючи додаток AntConc в семантично розміченому тексті, визначити кількість слів, що відносяться до одного з 21 *major discourse fields*.

5. Використовуючи додаток AntConc в семантично розміченому тексті, визначити колокації слів, що відносяться до одного з 21 *major discourse fields*.

6. Відсортувати результат видачі за семантичним значенням наступного слова.

Лабораторна робота 5

ПРАКТИЧНИЙ АНАЛІЗ КОРПУСУ В ONLINE СЕРЕДОВИЩІ CQPWEB

Середовище CQPweb (Corpus Query Processor) надає безліч потужних засобів обробки і аналізу корпусів. CQPweb дозволяє отримати доступ до великої кількості корпусів, включаючи корпуси British English 06 і American English 06, які розроблені в 2006 році на основі Brown і LOB корпусів. Використання середовища CQPweb не вимагає розробки власних програмних засобів для аналізу корпусів і здійснює online обробку даних на сервері Lancaster University, і потім результат запиту відображається в браузері комп'ютера клієнта.

Після реєстрації система CQPweb дозволяє отримати доступ до великої кількості корпусів. Однак до одного з найважливіших корпусів, British National Corpus або BNC, можна підключитися тільки через власний вхід. BNCweb являє собою відкриту систему доступу до British National Corpus, що дозволяє будь-якому охочому отримати доступ до BNC.

Крім того можна отримати доступ до BNCweb через сервер Lancaster University з обмеженими можливостями використання.

Потрібно зайти на сайт Lancaster BNCweb Server Usernames: http://bncweb.lancs.ac.uk/bncwebSignup/user/login.php (рис. 5.1).

Рисунок 5.1 – Сайт Lancaster BNCweb Server Usernames

Для реєстрації необхідно перейти за посиланням *Register for an Account* і заповнити поля реєстрації, вказавши в полі *Institution* університет.

Після підтвердження реєстрації можна перейти за посиланням *Login to BNCweb*.

Створити обліковий запис в системі CQPweb можна, зайшовши на сторінку: https://cqpweb.lancs.ac.uk, і перейти за посиланням *Create account*.

Після реєстрації стає доступним вхід на головну сторінку CQPweb-сервера: https://cqpweb.lancs.ac.uk/, яка являє собою базовий Веб-інструмент обробки запитів до корпусів. Основна ідея даного сервісу полягає в тому, що можна інсталювати корпус на сервер і скористатися інструментарієм сервісу для обробки корпусу, доступним з будь-якого комп'ютера клієнта.

У головному вікні CQPweb представлено список корпусів, які інстальовані на сервер (рис. 5.2).

Рисунок 5.2 – Головне вікно CQPweb

Клацання по корпусу (наприклад, American English 2006 Corpus) дозволяє відкрити його для пошуку та інших видів аналізу. Для аналізу будь-якого з корпусів головне меню CQPweb надає однакові макети, але розділяє їх індивідуальними кольорами.

У лівій частині сторінки, що дозволяє аналізувати той чи інший завантажений на сервер корпус засобами CQPweb, розташоване меню, розділене на групи: *Corpus queries, Saved query data, Corpus info* і *About CQPweb*. Кожна група об'єднує дії, що відповідають її назві. Наприклад,

група *About CQPweb* включає пункт *CQPweb main menu*, вибір якого здійснює перехід на головну сторінку CQPweb для вибору, для аналізу і обробки іншого корпусу.

Використання пункту меню *Corpus documentation* групи *Corpus info* дозволяє отримати інформацію про конкретний корпус, запропоновану розробниками цього корпусу.

У вікні аналізу корпусу (рис. 5.3) можна здійснити більшість стандартних процесів обробки корпусів. Наприклад, якщо ввести в поле пошуку *Standard Query* слово "somewhere" і натиснути кнопку *Start Query*, досить швидко буде отримано результат пошуку.

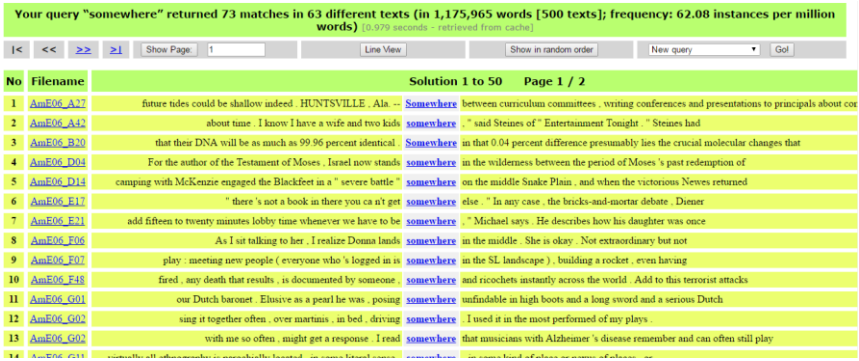
Рисунок 5.3 – Вікно аналізу корпусу

У верхній частині браузера відображається кількість запропонованих варіантів (73 matches in 63 different texts), розмір оброблюваного корпусу (in 1,175,965 words [500texs]) і нормування частоти зустрічальності на 1 мільйон слів (62, 08 instances per million words). Нижче відображається конкорданс, що показує результат пошуку. На сторінці відображено 50 збігів слова "somewhere", з лівостороннім і правостороннім контекстом.

Для повернення у вікно стандартного пошуку і створення нового запиту досить у верхній правій частині вікна (рис. 5.4) вибрати *New query* і натиснути *Go!*

За умовчанням конкордансер (рис. 5.4) відображає шукані терміни в центрі вікна. Таке відображення називається *KWIC (Keyword In Context*

display). Кількість відображуваних слів перед і після заданого слова залежить від кожного конкретного корпусу і описується самим конкретним корпусом. Натискання кнопки *Line View* у верхній частині вікна змінює зображення, вирівнюючи контекст по лівому краю і прибираючи центрування навколо шуканого терміну.



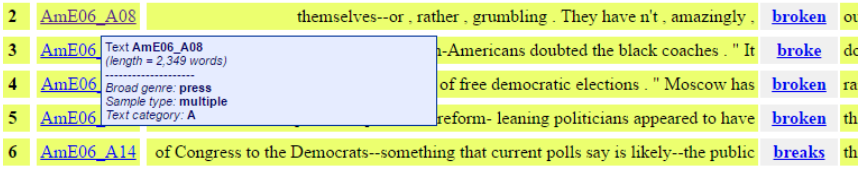
Your query "somewhere" returned 73 matches in 63 different texts (in 1,175,965 words [500 texts]; frequency: 62.08 instances per million words) [0.979 seconds - retrieved from cache]

I< << >> >1 Show Page: 1 Line View Show in random order New query Go!

No	Filename	Solution 1 to 50 Page 1 / 2		
1	AmE06_A27	future tides could be shallow indeed . HUNTSVILLE , Ala . --	Somewhere	between curriculum committees , writing conferences and presentations to principals about cor
2	AmE06_A42	about time . I know I have a wife and two kids	somewhere	. " said Steines of " Entertainment Tonight . " Steines had
3	AmE06_B20	that their DNA will be as much as 99.96 percent identical .	Somewhere	in that 0.04 percent difference presumably lies the crucial molecular changes that
4	AmE06_D04	For the author of the Testament of Moses , Israel now stands	somewhere	in the wilderness between the period of Moses 's past redemption of
5	AmE06_D14	camping with McKenzie engaged the Blackfeet in a " severe battle "	somewhere	on the middle Snake Plain , and when the victorious Newes returned
6	AmE06_E17	" there 's not a book in there you ca n't get	somewhere	else . " In any case , the bricks-and-mortar debate , Diener
7	AmE06_E21	add fifteen to twenty minutes lobby time whenever we have to be	somewhere	. " Michael says . He describes how his daughter was once
8	AmE06_F06	As I sit talking to her , I realize Donna lands	somewhere	in the middle . She is okay . Not extraordinary but not
9	AmE06_F07	play : meeting new people (everyone who 's logged in is	somewhere	in the SL landscape) , building a rocket , even having
10	AmE06_F48	fired , any death that results , is documented by someone ,	somewhere	and ricochets instantly across the world . Add to this terrorist attacks
11	AmE06_G01	our Dutch baronet . Elusive as a pearl he was , posing	somewhere	unfindable in high boots and a long sword and a serious Dutch
12	AmE06_G02	sing it together often , over martinis , in bed , driving	somewhere	. I used it in the most performed of my plays .
13	AmE06_G02	with me so often , might get a response . I read	somewhere	that musicians with Alzheimer 's disease remember and can often still play

Рисунок 5.4 – Результати запиту

У лівій частині вікна (рис. 5.4) для кожного знайденого екземпляра шуканого терміна відображається файл (в більшості випадків корпусу складаються з багатьох тисяч текстових файлів), в якому цей термін знайдено. Кожне ім’я такого файла забезпечено міткою, яка відображається при підведенні миші до імені файла (рис. 5.5).



2	AmE06_A08	themselves--or , rather , grumbling . They have n't , amazingly ,	broken	ou
3	AmE06	-Americans doubted the black coaches . " It	broke	do
4	AmE06	of free democratic elections . " Moscow has	broken	ra
5	AmE06	reform- leaning politicians appeared to have	broken	th
6	AmE06_A14	of Congress to the Democrats--something that current polls say is likely--the public	breaks	th

Text **AmE06_A08**
(length = 2,349 words)

Broad genre: **press**
Sample type: **multiple**
Text category: **A**

Рисунок 5.5 – Відображення імені файла

Якщо клацнути мишею по імені файла можна отримати інформацію про метадані файла (рис. 5.6), яка використовується і надається конкретним корпусом.

Metadata for text *AmE06_A08*	
Text identification code	AmE06_A08
Text category	A
Broad genre	press
Title	The Wall Street Journal: 'Grace Under Pressure; Difficult times call for less-contentious politics' (no page number); 'Volkswagen Restructuring Drive Starts to Sputter' (p. A3)
Author	Peggy Noonan; Stephen Power
Publication date	December 1, 2006; December 30, 2006
Sampled from	All; Beginning
Sample type	multiple
Source URL	Factiva
No. words in text	2,349

Рисунок 5.6 – Метадані файла

Повернувшись до основної сторінки конкордансера в спливаючій мітці, подібній опису файла, можна отримати значення анотації всієї контекстної лінії шуканого слова (найчастіше це POS-розмітка), до якого підведено курсор миші (рис. 5.7).

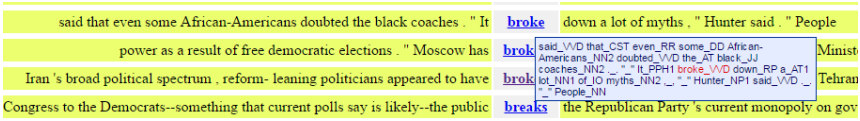


Рисунок 5.7 – Відображення розмітки

Клік миші по будь-якому шуканому слову конкордансу відкриває вікно розширеного контексту цього слова (рис. 5.8).

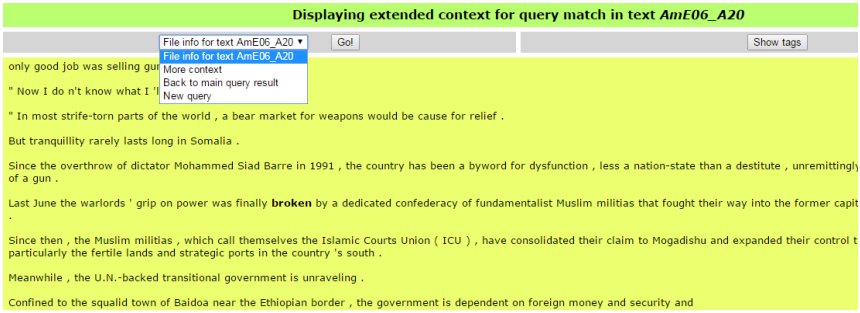


Рисунок 5.8 – Вікно розширеного контексту

Вкладка *Standard Query* дозволяє здійснити деякі опціональні налаштування пошуку. В полі *Query mode* можна вибрати необхідність врахування регістру символів (*Simple query (case-sensitive)*)

Використання пунктів списку *Restriction* дозволяє обмежити видачу результатів пошуку певними жанрами або типами підкорпусів аналізованого корпусу. При цьому список, що випадає, пункту *Restriction* різний для різних корпусів. Наприклад, для British English 2006 corpus цей список надає вибір з кількох жанрів: Fiction, General prose (non-fiction), Learned (academic), Press і Last restriction, що дозволяють ідентифікувати основні секції даного корпусу. Вибір пункту Last Restrictions дозволяє організовувати пошук в тих самих обмеженнях (підсекціях), в яких здійснювався попередніх пошук.

Завдання до лабораторної роботи 5

1. Створити свій обліковий запис на сайтах BNCweb і CQPweb.

2. На сайті CQPweb порівняти результати видачі конкордансерів заданих термінів і фраз відповідних варіантів для корпусів British English 2006 і American English 2006.

3. Визначити метадані одного з файлів, відповідного заданій тематиці підкорпусів.

4. Визначити і описати тип розмітки даного файла (POS-tagging, синтаксична або семантична).

5. Організувати пошук заданого терміна в обмеженому підкорпусі і визначити обсяг підкорпусу.

6. Використовуючи Simple Query Language Syntax організувати пошук лексеми:

– за заданим терміном (всіх форм словозміни цього слова);

– з використанням POS-розмітки в запиті;

– з використанням регулярного виразу в запиті;

– з використанням XML-розмітки в запиті.

Лабораторна робота 6

РОЗШИРЕНІ МОЖЛИВОСТІ СЕРЕДОВИЩА CQPWEB

6.1. Функція Restricted Query середовища CQPweb

Вхідною сторінкою будь-якого корпусу середовища CQPweb є сторінка *Standard query*. Для того щоб потрапити на сторінку розширеного пошуку необхідно в групі меню *Corpus queries* вибрати пункт *Restricted query* (рис. 6.1).

Рисунок 6.1 – Сторінка розширеного пошуку

Верхня частина вікна розширеного пошуку практично збігається з вікном стандартного пошуку. Розширений пошук пропонує обмеження запиту за декількома напрямами.

Всі тексти в корпусах, що завантажуються на CQPweb, класифіковані за однією або кількома засадами класифікації. Наприклад, корпус Brown Family (C8 tags) має п’ять різних схем класифікації (рис. 6.2), кожна з яких містить кілька категорій:

– чотири жанрові категорії;

– п’ять категорій, що визначає корпус приналежності тексту (Brown Family являє собою колекцію декількох корпусів);

– п’ятнадцять текстових категорій;

– три категорії періоду часу;

– дві категорії регіону мови (British-English, American-English).

Рисунок 6.2 – Схеми класифікації

Ті ж класифікаційні схеми можна побачити в метаданих файла (рис. 6.3).

Metadata for text *BE06_C06*	
Text identification code	BE06_C06
Corpus	BE2006 Corpus
Text category	C. Press: Reviews (theatre, books, music, dance)
Broad Genre	Press
Time period	Texts from 2006
Regional variety	British English
Source text	Tony Blair Rock Star Friends & Crocodiles Elizabeth David: A Life in Recipes Eleventh Hour; 'Get Carter!' : Sunday Telegraph
No. words in text	2,185

Рисунок 6.3 – Класифікаційні схеми в метаданих файла

Вказавши у вікні обмеженого пошуку (рис. 6.2) умови пошуку і натиснувши кнопку *Start query*, буде отримана сторінка конкордансера з результатом пошуку, аналогічна стандартному пошуку, на верхніх рядках якої буде описано результат обмеженого запиту (рис. 6.4).

Рисунок 6.4 – Результат обмеженого запиту

Якщо в якійсь схемі класифікації будуть обрані одна або кілька категорій і не будуть обрані категорії в інших класифікаційних схемах (наприклад, вказано тільки часовий період, а інші види класифікацій не вибрані), то ті групи, в яких жодна з категорій не обрана, припускають вибір будь-якої категорії або відображають результат пошуку по всіх категоріях.

6.2. Історія запитів

Пункт меню *Query history* доступний в групі *Saved query data* (рис. 6.5) на сторінці як стандартного, так і розширеного пошуку.

Рисунок 6.5 – Меню Query history

Виклик цього пункту відображає всі запити до поточного корпусу, здійснені користувачем за допомогою системи CQPweb. Кожен здійснений запит описано за допомогою п'яти стовпців: номер по порядку (No), текст запиту, введеного у вікно пошуку (Query), обмеження запиту (Restriction), кількість виданих результатів пошуку (Hits) і дата та час запиту (Date). Інформація в трьох стовпцях представлена посиланнями:

– клацання по посиланню стовпчика *Hits* повертає в конкордансер, при цьому результат видачі отримано з кешу (рис. 6.6);

– клацання миші по посиланню в полі *Query* здійснює перехід у вікно стандартного запиту, в якому заповнені необхідні поля;

– клацання миші по посиланню в полі *Restriction* здійснює перехід у вікно розширеного запиту із заповненими полями і вибраними обмеженнями.

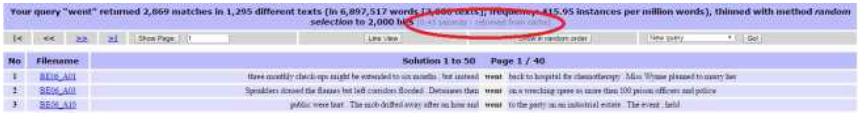


Рисунок 6.6 – Відображення запитів до поточного корпусу

6.3. Розподіл частоти (Distribution option)

Система CQPweb дозволяє побачити розподіл частоти шуканого слова в різних секціях корпусу. Секції являють собою групи текстів, які визначені різними типами метаданих і повністю відповідають категоріям розширеного пошуку. Для переходу у вікно розподілу частот необхідно у

верхній правій частині вікна конкордансера вибрати пункт *Distribution* і натиснути кнопку *Go!* (рис. 6.7).

Рисунок 6.7 – Перехід до вікна розподілу частот

Отримані таблиці розподілу частот шуканого слова будуть відповідати наявним класифікаційним категоріям аналізованого корпусу.

Наприклад, для корпуса *Brown Family (extended)* виділяється кілька таблиць розподілу, відповідних базовій класифікації: *Broad Genre, Corpus, Text category, Time period* і *Regional variety* (рис. 6.8).

У другому стовпці таблиці *Words in category* показано наскільки дана категорія велика, тобто скільки вона містить слів. Наприклад, Total: 6897517 – це загальний розмір корпусу.

У третьому стовпці *Hits in category* показано скільки разів шукане слово знайдено в кожній категорії. Наприклад, слово "sometime" всього в корпусі знайдено 43 рази, з них в категорії Fiction – 18, в категорії Press – 10 і так далі.

У четвертому стовпці *Dispersion* показано в скількох файлах даної категорії зустрічається це слово. Наприклад, запис "18 out of 756" позначає, що з 756 файлів, що відносяться до категорії *Fiction* шукане слово зустрічається в 18 файлах.

В останньому стовпці *Frequency* показана відносна частота шуканого слова в даній категорії, тобто частота цього слова на мільйон слів даної категорії. В даному прикладі, частота 10.04 = 18 (скільки разів пошукові слова знайдено в даній категорії) / 1791997 (загальна кількість слів в даній категорії) * 1 000 000 (нормування частоти).

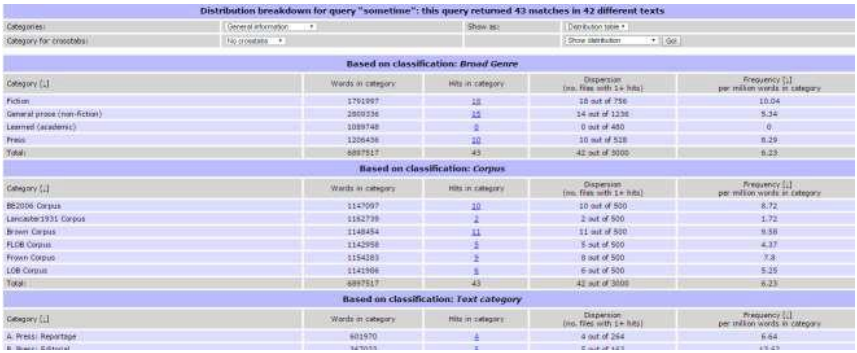


Рисунок 6.8 – Таблиці розподілу

Розподіл можна відобразити у вигляді гістограми (рис. 6.9), для цього у верхньому правому куті вікна розподілу в групі *Show as:* потрібно вибрати пункт *Bar chart* і натиснути кнопку *Go!*

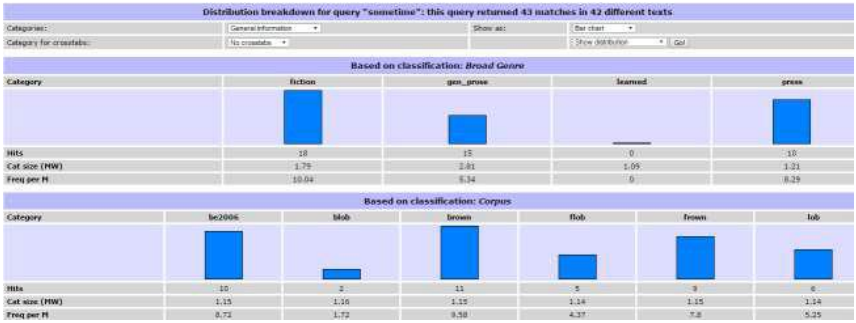


Рисунок 6.9 – Розподіл у вигляді гістограми

6.4. Функція розрідженого запиту (Thin Query function)

Функція розрідженого запиту може бути використана для спрощення обробки результатів запиту слова, що досить часто зустрічається. Наприклад, запитавши в стандартному запиті *(Standard Query)* корпусу BNC Sampler (2 304 310 слів) лемму {have}, отримаємо результат збігів 31 202. Неможливо проаналізувати таку кількість результатів видачі. У тому випадку якщо до результату видачі застосовується не автоматичний статистичний аналіз, а аналізуються результати "вручну", CQPWeb має функцію зменшення кількості прикладів результату видачі запиту *(Thin Query function).*

Опція *Thin Query function* знаходиться в згорнутому списку верхньої частини вікна результатів видачі конкордансера поруч із кнопкою *Go!* (рис. 6.10).

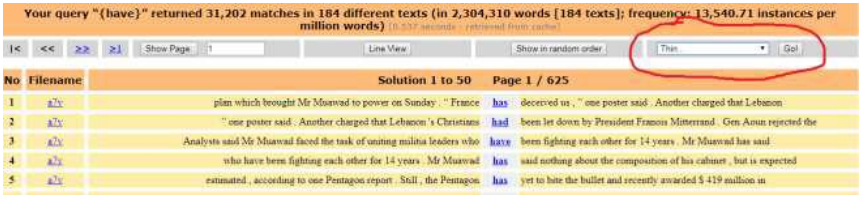


Рисунок 6.10 – Опція Thin Query function

В полі *Number of instances or percentage* вікна, що відкривається, необхідно ввести число прикладів, які необхідно отримати (рис. 6.11).

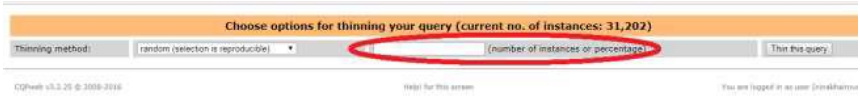


Рисунок 6.11 – Поле Number of instances or percentage

Наприклад, якщо ввести число 500, то результат видачі буде зменшений до 500 збігів (по 50 на 10 сторінок) (рис. 6.12). Цей результат є розрідженим, так як в ньому випадковим чином залишено 500 збігів (з наявних 31 202).

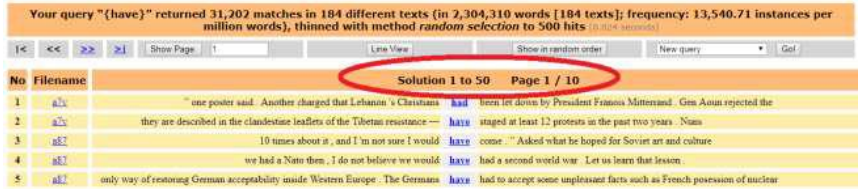


Рисунок 6.12 – Результат видачі

При налаштуванні методу розрядження запиту *(Thinning method)* можна вибрати два варіанти: *random (selection is reproducible)* і *random (selection is not reproducible)*. При цьому результат розрядженого запиту завжди виходить випадковим чином, тобто 500 збігів вибирається з 31 202 випадково. Але якщо обрано метод *random (selection is reproducible)*, то

при декількох однакових запитах буде відтворено один і той же результат тонкого або розрідженого запиту. Тобто в нашому прикладі будуть отримані ті ж 500 результатів збігу леми. Якщо буде обрано протилежний метод розрідженого запиту *random (selection is not reproducible)* і той самий запит буде поставлено знову, то буде отримано іншу множину довільно обраних збігів леми {have} зі словами корпусу.

Різниця в цих двох варіантах полягає в загальному підході комп'ютера до обробки випадковості *(computers treat randomness)*, якому необхідно задати джерело – число, з якого починає обчислюватися випадковість. Тоді для вибірки, що відтворюється, використовується одне і те ж число, обчислення випадковості, а для невідтворюваності вибірки дане число реально генерується генератором випадкових чисел комп'ютера, в залежності від поточного моменту часу.

Завдання до лабораторної роботи 6

1. Описати класифікаційні схеми корпусу, заданого варіанта *Лабораторної роботи 6.*

2. Організувати пошук і відображення конкордансу для шуканої лексеми на заданому корпусі, при заданих обмеженнях пошуку.

3. Описати результати пошуку (розмір підкорпусів, в яких здійснювався пошук; кількість запропонованих варіантів; кількість файлів, в яких знайдено збіг; нормовану частоту).

4. Відобразити метадані будь-якого файла корпусу, в якому знайдена задана лексема.

5. Відобразити історію своїх запитів в даному корпусі.

6. Описати частотний розподіл заданої лексеми в табличному вигляді і у вигляді гістограми.

7. Здійснити пошук заданої лексеми в повному корпусі, потім випадковим чином вибрати 20 результатів видачі для аналізу.

Таблиця 6.1 – Варіанти завдання

Номер варіанта	Корпус для обробки	Обмеження	Шукана лексема
1	American English 2006	Press, Beginning	national
2	British English 2006	Press, Text category “A” та “B”	nation
3	American English 2006	Fiction, Text category “M”	stand
4	British English 2006	General prose (non-fiction), Text category “F” та “E”	stand
5	British National Corpus (XML edition)	Academic prose, 1985-1993, Written books and periodicals	broad
6	British National Corpus (XML edition)	Written books and periodicals, 1960-1974, Fic-tion and verse	follow

Лабораторна робота 7

РОБОТА З КЛАСТЕРАМИ І N-ГРАМАМИ

7.1. Визначення кластерів в корпусах конкордансером AntConc

Кластери являють собою повторювані шаблони двох і більше слів в корпусі. Інструментарій кластера дозволяє визначати шаблони з двох, трьох або більше слів навколо шуканого терміна.

Конкордансер *AntConc* має вбудовану вкладку *Clusters / N-Grams*, яка дозволяє шукати слова або шаблони в корпусі і групувати або кластеризувати результат праворуч або ліворуч від шуканого терміна, після чого сортувати результат, наприклад, за частотою.

Для роботи з кластерами потрібно перейти на вкладку *Clusters / N-Grams* конкордансера, ввести шуканий термін в поле *Search Term* і натиснути кнопку *Start*. Якщо не змінювати налаштування, встановлених за умовчанням, то результат видачі буде аналогічний, показаному на рис. 7.1. В даному прикладі розмір кластера використовувався за умовчанням *Cluster Size (min-2 words, max-2 words).*

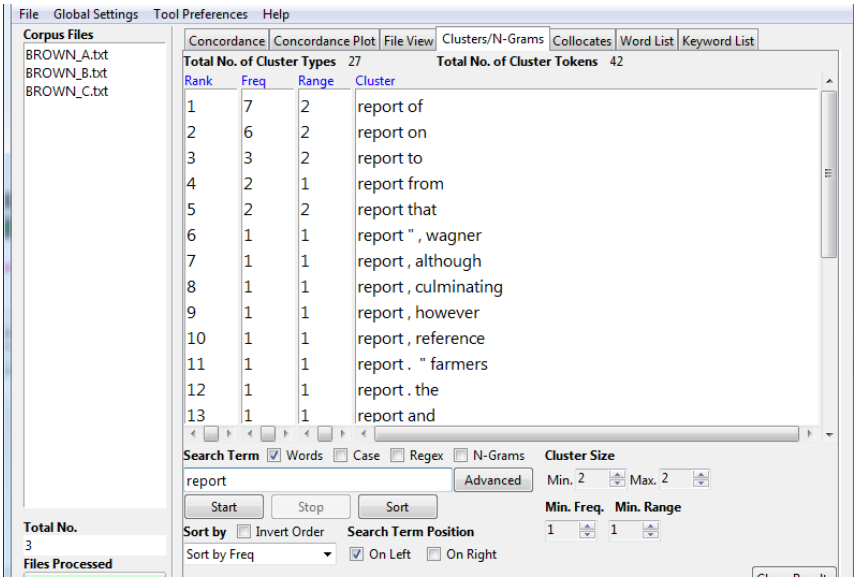


Рисунок 7.1 – Вкладка Clusters / N-Grams

У разі необхідності можна налаштувати пошук і видачу результатів кластеризації в нижній частині вікна вкладки. В полі *Min.Freq.* вказується частота появи кластера, в полі *Min.Range* вказується кількість файлів корпусу, в яких присутній знайдений кластер.

У прикладі, показаному на рисунку 7.1, з двухсловних кластерів найчастіше (7 разів) зустрічається послідовність слів "report of" в двох файлах (завантажений підкорпус складається всього з трьох файлів).

Якщо клацнути мишею по будь-якому словосполученню в стовпці *Cluster*, можна перейти на вкладку *Concordance* і подивитися всі появи даного словосполучення в текстах (рис. 7.2).

Рисунок 7.2 – Вкладка Concordance

Отримані результати кластерів (рис. 7.1) можна впорядкувати за частотою – *Sort by Freq*, за кількістю файлів, в яких присутній даний кластер – *Sort by Rang*, за алфавітом – *Sort by Word*, за алфавітом закінчення кластера – *Sort by Word End* (рис. 7.3).

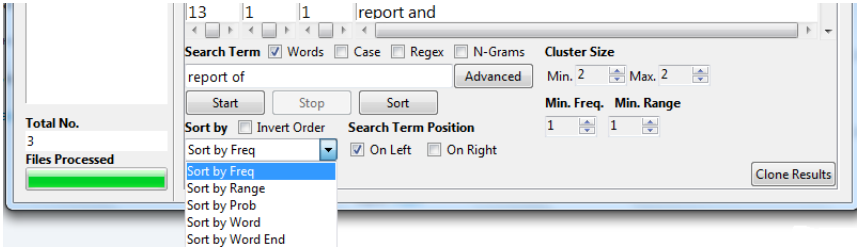


Рисунок 7.3 – Опція сортування

Крім того можна впорядкувати видані в результаті пошуку кластери за розміром ймовірності переходу між першим словом і другим словом (або всіма іншими словами) – *Sort by Prob*. Для цього потрібно на вкладці *Clusters / N-Grams* пункту меню *Tool Preferences* включити опцію *Transitional probability between first and other words*. Після чого в основному вікні *Clusters / N-Grams* з'явиться додаткова колонка *Prob* (рис. 7.4). У цій колонці буде відображатися ймовірність переходу між першим і другим словом, якщо в кластері два слова, або між шуканим словом і іншими словами кластера, якщо в кластері більше одного слова.

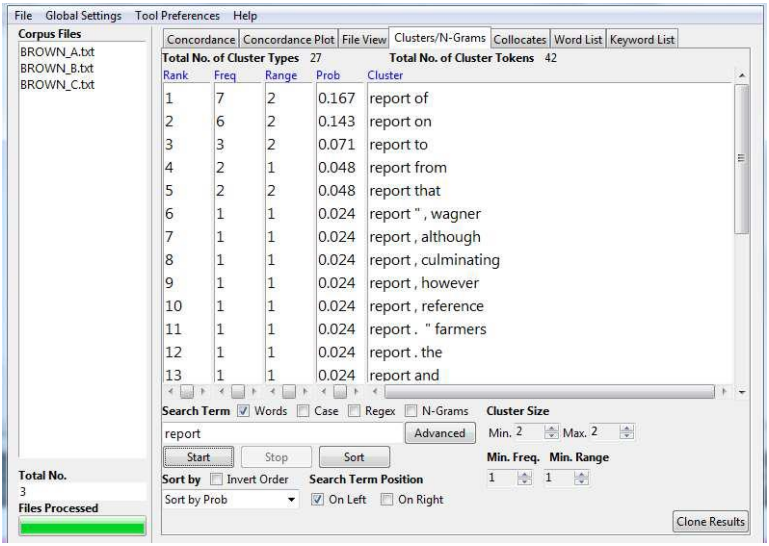


Рисунок 7.4 – Колонка Prob на вкладці Clusters / N-Grams

Наприклад, показана на рисунку 7.4 ймовірність переходу 0.167 для кластера "report of" показує, з якою ймовірністю в даному корпусі друге слово буде з'являтися після першого слова. Якби ці слова "report" і "of" завжди з'являлися в корпусі разом, то значення даної ймовірності переходу дорівнювало б одиниці.

Продовжуючи приклад, якщо в опціях *Tool Preferences Clusters / N-Grams Preferences* вимкнено опцію *Treat all data as lowercase* (вкладка *Other Options*), то в результаті ймовірність переходу в кластері "Report of" буде дорівняти 1. Це означає, що слова Report з великої літери і of з маленької завжди в корпусі з'являються разом.

7.2. Визначення n-грам в корпусах конкордансером AntConc

Механізм пошуку n-грам дозволяє визначати в корпусі повторювані послідовності ланцюжків слів без необхідності пошуку термінів.

Наприклад, 2-грамами для речення "This is a pan" будуть "This is", "is a" і "a pen". Пошук n-грам дозволяє знаходити найбільш загальні кластери (ланцюжки слів) в корпусі без необхідності пошуку термінів. Для n-грамного пошуку потрібно в нижній частині вікна вкладки *Clusters / N-Grams* (рис. 7.5) включити опцію *N-Gram* групи *Search Term*. Даний інструмент дозволяє сканувати весь корпус в пошуку послідовностей з n слів.

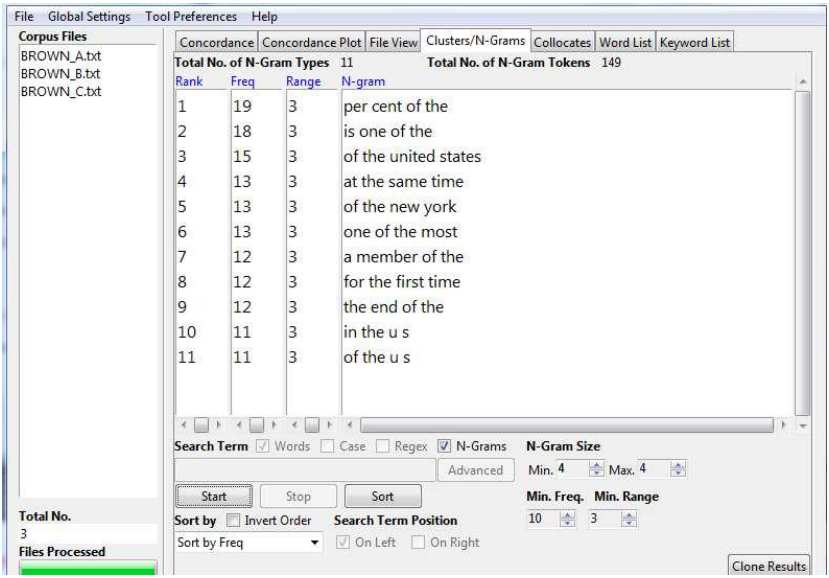


Рисунок 7.5 – N-грамний пошук

В налаштуваннях нижньої частини вікна (рис. 7.5) можна задати розмір шуканих в корпусі n-грам (*Min, Max*), мінімальну частоту їх появи в корпусі (*Min. Freq.*) і кількість файлів, в яких n-грами повинні бути присутніми (*Min. Range*).

Аналогічно функції, показаної на рис. 7.2, клацанням по знайденій фразі можна перейти на вкладку *Concordance* і побачити використання цієї фрази в контексті. Знайдені n-грами можна також сортувати за частотою – *Sort by Freq*, за кількістю файлів, в яких присутній даний кластер –

Sort by Rang, за алфавітом – *Sort by Word*, за алфавітом закінчення кластера – *Sort by Word End* і по ймовірності переходу між словами ланцюжка – *Sort by Prob*.

Аналогічним чином можна здійснювати пошук кластерів в розмічених корпусах. При цьому при виборі розміру кластера теги так само вважаються окремою одиницею (рис. 7.6).

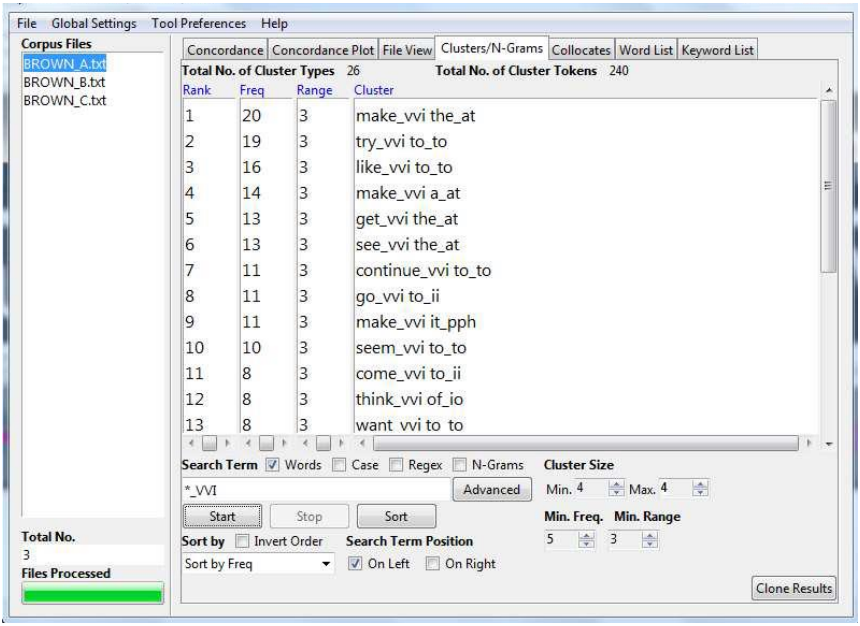


Рисунок 7.6 – Пошук кластерів в розмічених корпусах

Завдання до лабораторної роботи 7

1. Відкрити підкорпус свого варіанта в конкордансері AntConc.

2. Використовуючи налаштування, задані за умовчанням, визначити кластери з заданим словом.

3. Відсортувати отриманий список за частотою появи в файлах.

4. Відсортувати отриманий список за ймовірністю переходу від першого слова кластера до наступних. Пояснити отримані ймовірності переходу.

5. Провести пошук шаблонів з заданим словом, з урахуванням регістру букв.

6. Знайти шаблони ланцюжків з заданим словом довжиною не менше трьох слів, що зустрічаються не менше трьох разів, не менше ніж у трьох файлах, що враховують слова праворуч від заданого.

7. Переглянути появи даних словосполучень в текстах корпусу.

8. Визначити n-грами, що складаються з 4 слів, які зустрічаються не менш ніж 4 рази, не менше ніж у 4 файлах. Проаналізувати отримані n-грами.

9. Використовуючи розмічені корпуси заданого варіанта, визначити частини мови, що найбільш часто стоять ліворуч і праворуч від заданого терміна.

Таблиця 7.1 – Варіанти завдання

Номер варіанта	Корпус для обробки	Термін
1	Текстові файли BROWN_A, BROWN_B, BROWN_C, BROWN_D, BROWN_E корпусу Brown	year
2	Текстові файли LOB_A, LOB_B, LOB_C, LOB_D, LOB_E корпусу LOB	Britain
3	Текстові файли BROWN_F, BROWN_G, BROWN_H, BROWN_J, BROWN_K корпусу Brown	American
4	Текстові файли LOB_F, LOB_G, LOB_H, LOB_J, LOB_K корпусу LOB	war
5	Текстові файли BROWN_L, BROWN_M, BROWN_N, BROWN_P, BROWN_R корпусу Brown	people
6	Текстові файли LOB_L, LOB_M, LOB_N, LOB_P, LOB_R корпусу LOB	world

ДОДАТКИ

Додаток Г
СИСТЕМА БАЗОВИХ КАТЕГОРІЙ USAS

A general and abstract terms	**B** the body and the individual	**C** arts and crafts	**E** emotion
F food and farming	**G** government and public	**H** architecture, housing and the home	**I** money and commerce in industry
K entertainment, sports and games	**L** life and living things	**M** movement, location, travel and transport	**N** numbers and measurement
O substances, materials, objects and equipment	**P** education	**Q** language and communication	**S** social actions, states and processes
T Time	**W** world and environment	**X** psychological actions, states and processes	**Y** science and technology
Z names and grammar			

Додаток Д

ПОВНА СИСТЕМА СЕМАНТИЧНИХ КАТЕГОРІЙ USAS

A1 General and abstract terms
A1.1.1 General actions, making etc.
A1.1.2 Damaging and destroying
A1.2 Suitability
A1.3 Caution
A1.4 Chance, luck
A1.5 Use
A1.5.1 Using
A1.5.2 Usefulness
A1.6 Physical/mental
A1.7 Constraint
A1.8 Inclusion/Exclusion
A1.9 Avoiding
A2 Affect
A2.1 Affect:- Modify, change
A2.2 Affect:- Cause/Connected
A3 Being
A4 Classification
A4.1 Generally kinds, groups, examples
A4.2 Particular/general; detail
A5 Evaluation
A5.1 Evaluation:- Good/bad
A5.2 Evaluation:- True/false
A5.3 Evaluation:- Accuracy
A5.4 Evaluation:- Authenticity
A6 Comparing
A6.1 Comparing:- Similar/different
A6.2 Comparing:- Usual/unusual
A6.3 Comparing:- Variety
A7 Definite (+ modals)
A8 Seem
A9 Getting and giving; possession
A10 Open/closed; Hiding/Hidden; Finding; Showing
A11 Importance

A11.1 Importance: Important
A11.2 Importance: Noticeability
A12 Easy/difficult
A13 Degree
A13.1 Degree: Non-specific
A13.2 Degree: Maximizers
A13.3 Degree: Boosters
A13.4 Degree: Approximators
A13.5 Degree: Compromisers
A13.6 Degree: Diminishers
A13.7 Degree: Minimizers
A14 Exclusivizers/particularizers
A15 Safety/Danger
B1 Anatomy and physiology
B2 Health and disease
B3 medicines and medical treatment
B4 Cleaning and personal care
B5 Clothes and personal belongings
C1 Arts and crafts
E1 Emotional actions, states and processes General
E2 Liking
E3 Calm/Violent/Angry
E4 Happy/sad
E4.1 Happy/sad: Happy
E4.2 Happy/sad: Contentment
E5 Fear/bravery/shock
E6 Worry, concern, confident
F1 Food
F2 Drinks
F3 Cigarettes and drugs
F4 Farming & Horticulture
G1 Government, Politics and elections
G1.1 Government etc.
G1.2 Politics
G2 Crime, law and order
G2.1 Crime, law and order: Law and order
G2.2 General ethics

G3 Warfare, defence and the army; weapons
H1 Architecture and kinds of houses and buildings
H2 Parts of buildings
H3 Areas around or near houses
H4 Residence
H5 Furniture and household fittings
I1 Money generally
I1.1 Money: Affluence
I1.2 Money: Debts
I1.3 Money: Price
I2 Business
I2.1 Business: Generally
I2.2 Business: Selling
I3 Work and employment
I3.1 Work and employment: Generally
I3.2 Work and employmeny: Professionalism
I4 Industry
K1 Entertainment generally
K2 Music and related activities
K3 Recorded sound etc.
K4 Drama, the theatre and showbusiness
K5 Sports and games generally
K5.1 Sports
K5.2 Games
K6 Childrens games and toys
L1 Life and living things
L2 Living creatures generally
L3 Plants
M1 Moving, coming and going
M2 Putting, taking, pulling, pushing, transporting &c.
M3 Vehicles and transport on land
M4 Shipping, swimming etc.
M5 Aircraft and flying
M6 Location and direction
M7 Places
M8 Remaining/stationary
N1 Numbers

N2 Mathematics
N3 Measurement
N3.1 Measurement: General
N3.2 Measurement: Size
N3.3 Measurement: Distance
N3.4 Measurement: Volume
N3.5 Measurement: Weight
N3.6 Measurement: Area
N3.7 Measurement: Length & height
N3.8 Measurement: Speed
N4 Linear order
N5 Quantities
N5.1 Entirety; maximum
N5.2 Exceeding; waste
N6 Frequency etc.
O1 Substances and materials generally
O1.1 Substances and materials generally: Solid
O1.2 Substances and materials generally: Liquid
O1.3 Substances and materials generally: Gas
O2 Objects generally
O3 Electricity and electrical equipment
O4 Physical attributes
O4.1 General appearance and physical properties
O4.2 Judgement of appearance (pretty etc.)
O4.3 Colour and colour patterns
O4.4 Shape
O4.5 Texture
O4.6 Temperature
P1 Education in general
Q1 Linguistic actions, states and processes; communication
Q1.1 Linguistic actions, states and processes; communication
Q1.2 Paper documents and writing
Q1.3 Telecommunications
Q2 Speech acts
Q2.1 Speech etc:- Communicative
Q2.2 Speech acts
Q3 Language, speech and grammar

Q4 The Media
Q4.1 The Media:- Books
Q4.2 The Media:- Newspapers etc.
Q4.3 The Media:- TV, Radio and Cinema
S1 Social actions, states and processes
S1.1 Social actions, states and processes
S1.1.1 Social actions, states and processes
S1.1.2 Reciprocity
S1.1.3 Participation
S1.1.4 Deserve etc.
S1.2 Personality traits
S1.2.1 Approachability and Friendliness
S1.2.2 Avarice
S1.2.3 Egoism
S1.2.4 Politeness
S1.2.5 Toughness; strong/weak
S1.2.6 Sensible
S2 People
S2.1 People:- Female
S2.2 People:- Male
S3 Relationship
S3.1 Relationship: General
S3.2 Relationship: Intimate/sexual
S4 Kin
S5 Groups and affiliation
S6 Obligation and necessity
S7 Power relationship
S7.1 Power, organizing
S7.2 Respect
S7.3 Competition
S7.4 Permission
S8 Helping/hindering
S9 Religion and the supernatural
T1 Time
T1.1 Time: General
T1.1.1 Time: General: Past
T1.1.2 Time: General: Present; simultaneous

T1.1.3 Time: General: Future
T1.2 Time: Momentary
T1.3 Time: Period
T2 Time: Beginning and ending
T3 Time: Old, new and young; age
T4 Time: Early/late
W1 The universe
W2 Light
W3 Geographical terms
W4 Weather
W5 Green issues
X1 Psychological actions, states and processes
X2 Mental actions and processes
X2.1 Thought, belief
X2.2 Knowledge
X2.3 Learn
X2.4 Investigate, examine, test, search
X2.5 Understand
X2.6 Expect
X3 Sensory
X3.1 Sensory:- Taste
X3.2 Sensory:- Sound
X3.3 Sensory:- Touch
X3.4 Sensory:- Sight
X3.5 Sensory:- Smell
X4 Mental object
X4.1 Mental object:- Conceptual object
X4.2 Mental object:- Means, method
X5 Attention
X5.1 Attention
X5.2 Interest/boredom/excited/energetic
X6 Deciding
X7 Wanting; planning; choosing
X8 Trying
X9 Ability
X9.1 Ability:- Ability, intelligence
X9.2 Ability:- Success and failure

Y1 Science and technology in general
Y2 Information technology and computing
Z0 Unmatched proper noun
Z1 Personal names
Z2 Geographical names
Z3 Other proper names
Z4 Discourse Bin
Z5 Grammatical bin
Z6 Negative
Z7 If
Z8 Pronouns etc.
Z9 Trash can
Z99 Unmatched

Додаток Е

СИСТЕМА СЕМАНТИЧНИХ КАТЕГОРІЙ USAS РОСІЙСЬКОЮ МОВОЮ

A Общие Понятия
A1 Общие понятия
A1.1.1 Обычные действия / Изготовление ч.-л.
A1.1.1- Бездействие
A1.1.2 Повреждение и разрушение
A1.1.2- Налаживание и починка
A1.2 Пригодность
A1.2+ Пригодное
A1.2- Непригодное
A1.3 Осторожность
A1.3+ Осторожно
A1.3- Неосторожно
A1.4 Шансы, везение
A1.4+ Удача
A1.4- Неудача
A1.5 Употребление, применение, использование
A1.5.1 Использование
A1.5.1+ В использовании
A1.5.1- Неиспользуемое
A1.5.2 Польза
A1.5.2+ Полезное
A1.5.2- Бесполезное
A1.6 Материальное/Нематериальное
A1.7+ Ограничение
A1.7- Без ограничений
A1.8+ Включение
A1.8- Исключение
A1.9 Уклонение
A1.9- Неизбежное
A2 Влияние
A2.1 Изменение
A2.1+ С изменениями

A2.1- Без изменений
A2.2 Причина и следствие / Связь
A2.2+ Причина/Следствие/Связь
A2.2- Несвязанность
A3 Бытие, существование
A3+ Существующее
A3- Несуществующее
A4 Классификация
A4.1 Виды, группы, образцы
A4.1- Не отнесящиеся к определённой категории
A4.2 Общее и частное
A4.2+ Подробно
A4.2- В общем
A5 Оценка
A5.1 Оценка: Хорошо/Плохо
A5.1+ Оценка: Хорошо
A5.1- Оценка: Плохо
A5.2 Оценка: Правда/Неправда
A5.2+ Оценка: Правда
A5.2- Оценка: Неправда
A5.3 Оценка: Верность, точность
A5.3+ Оценка: Верно, точно
A5.3- Оценка: Неверно, неточно
A5.4 Оценка: Подлинность
A5.4+ Оценка: Подлинно
A5.4- Оценка: Неподлинно
A6 Сравнение
A6.1 Сравнение: Сходство/Различие
A6.1+ Сравнение: Сходство
A6.1- Сравнение: Различие
A6.2 Сравнение: Обычное/Необычное
A6.2+ Сравнение: Обычное
A6.2- Сравнение: Необычное
A6.3 Сравнение: Разнообразие
A6.3+ Сравнение: Много разнообразия
A6.3- Сравнение: Мало разнообразия
A7 Вероятность

A7+ Вероятно
A7- Невероятно
A8 Кажущееся
A9 Давать/Получать; владение, обладание
A9+ Получение и обладание
A9- Дача
A10 Открыто/Закрыто; Прятать; Находить; Показывать
A10+ Открыто; Находить; Показывать
A10- Закрыто; Спрятано
A11 Важность/Заметность
A11.1 Важность
A11.1+ Важно
A11.1- Неважно
A11.2 Заметность
A11.2+ Заметно
A11.2- Незаметно
A12 Лёгкость/Трудность
A12+ Легко
A12- Трудно
A13 Степень
A13.1 Степень: В общем
A13.2 Степень: Увеличение до максимума
A13.3 Степень: Усиление
A13.4 Степень: Приблизительно
A13.5 Степень: Неопределенно
A13.6 Степень: Преуменьшение
A13.7 Степень: Уменьшение до минимума
A14 Исключительно / В частности
A15 Опасность/Безопасность
A15+ Безопасно
A15- Опасно
B Тело Человека
B1 Анатомия и физиология
B2 Здоровье и болезнь
B2+ Здоров
B2- Нездоров
B3 Лекарства и лечение

B3- Без медицинского лечения
B4 Уборка и личная гигиена
B4+ Чисто
B4- Нечисто
B5 Одежда и личные принадлежности
B5- Без одежды
C1 Искусства и Ремёсла
C1 Искусства и ремёсла
E Эмоциональные Действия, Состояния и Процессы
E1 Эмоции в целом
E1+ Эмоциональность
E1- Без эмоций
E2 Расположение, предпочтение
E2+ Нравится
E2- Не нравится
E3 Спокойствие/Насилие/Гнев
E3+ Спокойствие
E3- Насилие/Гнев
E4 Счастье и удовлетворённость
E4.1 Счастье/Печаль
E4.1+ Счастлив
E4.1- Печален
E4.2 Удовлетворённость
E4.2+ Удовлетворён
E4.2- Неудовлетворён
E5 Храбрость и страх
E5+ Храбрость
E5- Страх/шок
E6 Беспокойство и уверенность
E6+ Уверен
E6- Беспокоится
F Пища и Сельское Хозяйство
F1 Пища
F1+ Изобилие еды
F1- Нехватка еды
F2 Напитки и алкоголь
F2+ Пьянство

F2- Непьющий
F3 Курение и наркотики
F3+ Курение и наркомания
F3- Некурящие / незлоупотребляющие наркотиками
F4 Сельское хозяйство, садоводство и огородничество
F4- Невозделанная земля
G Сфера Деятельности Государства
G1 Правительство и политика
G1.1 Государство и правительство
G1.1- Неправительственные, негосударственные органы
G1.2 Политика
G1.2- Вне политики
G2 Преступность и правопорядок
G2.1 Правопорядок
G2.1+ Законность
G2.1- Преступность
G2.2 Мораль и нравственность
G2.2+ Этично
G2.2- Неэтично
G3 Военные действия, оборона и армия; оружие
G3- Против войны
H Архитектура, Здания и Сооружения, Дом
H1 Архитектура, здания и сооружения
H2 Части зданий и сооружений
H3 Территории, близлежащие к жилым домам
H4 Местожительство
H4- Непроживание
H5 Мебель и домашнее оборудование
H5- Без мебели
I Деньги, Торговля и Промышленность
I1 Деньги
I1.1 Деньги и плата
I1.1+ Деньги: В достатке
I1.1- Деньги: Недостаток
I1.2 Деньги: Долги
I1.2+ Трата и потеря денег
I1.2- Без долгов

I1.3 Деньги: Цена
I1.3+ Дорого
I1.3- Дешево
I2 Бизнес, коммерческая деятельность
I2.1 Бизнес: В общем смысле
I2.1- Не связанное с коммерческой деятельностью
I2.2 Бизнес: Продажа
I3 Работа и трудоустройство
I3.1 Работа и трудоустройство: В общем
I3.1- Незанятость
I3.2 Работа и трудоустройство: Профессионализм
I3.2+ Профессионально
I3.2- Непрофессионально
I4 Промышленность
I4- С неразвитой промышленностью
K Развлечения, Спорт и Игры
K1 Развлекательные мероприятия
K2 Музыка
K3 Звукозапись
K4 Театр и шоу-бизнес
K5 Спорт и спортивные игры
K5.1 Спорт
K5.2 Спортивные игры
K6 Детские игры и грушки
L Жизнь и Всё Живое
L1 Жизнь и всё живое
L1+ Живое
L1- Неживое
L2 Живые существа: животные, птицы, и т. д.
L2- Без живых существ
L3 Растения
L3- Без растительности
M Движение и Передвижение, Местоположение
M1 Движение, передвижение
M2 Класть, тянуть, толкать, перемещать
M3 Наземный транспорт
M4 Водный транспорт

M4- Неплавающие
M5 Воздушный транспорт
M6 Местоположение and направление
M7 Принадлежность к местности
M8 Без движения
N Числа и Системы мер и Измерений
N1 Числа
N2 Математика
N3 Системы мер и измерений
N3.1 Системы мер и измерений: В общем
N3.2 Меры и измерения: Размер
N3.2+ Размер: Большой
N3.2- Размер: Маленький
N3.3 Меры и измерения: Расстояние
N3.3+ Расстояние: Далеко
N3.3- Расстояние: Близко
N3.4 Меры и измерения: Объём
N3.4+ Объем: Расширен
N3.4- Объем: Сжат
N3.5 Меры и измерения: Вес
N3.5+ Вес: Тяжёлый
N3.5- Вес: Лёгкий
N3.6 Меры и измерения: Пространство
N3.6+ Просторно
N3.7 Меры и измерения: Длина, широта и высота
N3.7+ Длинный, широкий и высокий
N3.7- Короткий и узкий
N3.8 Меры и измерения: Скорость
N3.8+ Скорость: Быстро
N3.8- Скорость: Медленно
N4 Линейный порядок
N4- Нелинейность
N5 Количество
N5+ Количество: много
N5- Количество: мало
N5.1 Полнота, целостность; максимум
N5.1+ Полностью; максимально

N5.1- Часть
N5.2 Превышение
N5.2+ С превышением; расточительство
N6 Частотность
N6+ Часто
N6- Нечасто
O Вещества и материалы, Предметы и Оборудование
O1 Вещества и материалы в целом
O1.1 Вещества и материалы: Твёрдые
O1.2 Вещества и материалы: Жидкие
O1.2- Сухие
O1.3 Вещества и материалы: Газообразные
O1.3- Негазообразные
O2 Предметы
O3 Электричество и электрическое оборудование
O4 Физические свойства
O4.1 Внешний вид и физические свойства в общем
O4.2 Оценка внешнего вида
O4.2+ Оценка внешнего вида: Красиво
O4.2- Оценка внешнего вида: Некрасиво
O4.3 Цвет и цветовые комбинации
O4.4 Форма
O4.5 Текстура
O4.6 Температура
O4.6+ Температура: Горячо/огонь
O4.6- Температура: Холодно
P Образование
P1 Образование
P1- Без образования
Q Лингвистические процессы; Коммуникация
Q1 Коммуникация
Q1.1 Коммуникация в общем
Q1.2 Печатные документы и письмо
Q1.2- Неписаное
Q1.3 Телекоммуникации
Q2 Речь
Q2.1 Речь: Коммуникативность

Q2.1+ Речь: Разговорчивость
Q2.1- Речь: Некоммуникативность
Q2.2 Речевые акты
Q2.2- Отсутствие речевых актов
Q3 Язык, речь и грамматика
Q3- Неречевое поведение
Q4 Средства массовой информации
Q4.1 Книги
Q4.2 Газеты и т. д.
Q4.3 Телевидение, Радио и Кино
S Общественные Действия, Состояния и Процессы
S1 Общественные Действия, Состояния и Процессы
S1.1 Общественные Действия, Состояния и Процессы
S1.1.1 Общественные Действия, Состояния и Процессы
S1.1.2 Взаимность
S1.1.2+ Взаимно, обоюдно
S1.1.2- Односторонне
S1.1.3 Участие
S1.1.3+ Участвовать
S1.1.3- Не участвовать
S1.1.4 Заслуживать ч.-л.
S1.1.4+ Достойный
S1.1.4- Недостойный
S1.2 Черты характера
S1.2.1 Доступность и дружелюбность
S1.2.1+ Неофициальность/Дружелюбность
S1.2.1- Официальность/Недружелюбие
S1.2.2 Жадность
S1.2.2+ Жадный
S1.2.2- Щедрый
S1.2.3 Эгоистичность
S1.2.3+ Себялюбивый
S1.2.3- Бескорыстный
S1.2.4 Вежливость
S1.2.4+ Вежлив
S1.2.4- Невежлив
S1.2.5 Жёсткость; сила/слабость

S1.2.5+ Жёсткий/сильный
S1.2.5- Слабый
S1.2.6 Здравый смысл
S1.2.6+ Здравомыслящий
S1.2.6- Абсурд
S2 Люди
S2- Безлюдно
S2.1 Люди: Женщины
S2.1- Неженственна
S2.2 Люди: Мужчины
S3 Взаимоотношения
S3.1 Личные взаимоотношения: В общем
S3.1- Отсутствие личных взаимоотношений
S3.2 Взаимоотношения: Близость и секс
S3.2+ Взаимоотношения: Сексуальные
S3.2- Взаимоотношения: Несексуальные
S4 Родство
S4- Отсутствие родства
S5 Группы людей и объединения
S5+ Принадлежность к группе
S5- Личная независимость
S6 Обязанность и необходимость
S6+ Большие обязанности / сильная необходимость
S6- Отсутствие обязанностей или необходимости
S7 Отношения власти и подчинения
S7.1 Организация власти
S7.1+ У власти
S7.1- Без власти
S7.2 Уважение
S7.2+ С уважением
S7.2- Без уважения
S7.3 Соперничество
S7.3+ Соперничать
S7.3- Без соперничества
S7.4 Разрешение
S7.4+ Разрешено
S7.4- Не разрешено

S8 Помощь/Помехи
S8+ Помогать
S8- Мешать
S9 Религия и всё сверхъестественное
S9- Нерелигиозное
T Время
T1 Время
T1.1 Время: В общем
T1.1.1 Время: Прошлое
T1.1.2 Время: Настоящее; одновременность
T1.1.2- Время: Асинхронно
T1.1.3 Время: Будущее
T1.2 Время: Мгновенность
T1.3 Время: Период
T1.3+ Промежуток времени: Долго
T1.3- Промежуток времени: Коротко
T2 Время: Начало и конец
T2+ Время: Начало
T2- Время: Конец
T3 Время: Давно/недавно; возраст
T3+ Время: Давно; Взрослые
T3- Время: Недавно; Юные
T4 Время: Рано/поздно
T4+ Время: Рано
T4- Время: Поздно
W Мир и Окружающая Среда
W1 Вселенная
W2 Свет
W2- Темнота
W3 Географические термины
W4 Погода
W5 Экология
X Психологические Действия, Состояния и Процессы
X1 Психологические Действия, Состояния и Процессы
X2 Умственная деятельность
X2.1 Соображения, мнения
X2.1- Не думая

X2.2 Знание
X2.2+ Знать
X2.2- Не знать
X2.3 Учить, узнавать
X2.3+ Познание
X2.4 Исследовать, изучать, проверять, искать
X2.4+ Перепроверять
X2.4- Не исследовано
X2.5 Понимать
X2.5+ Понимание
X2.5- Непонимание
X2.6 Ожидать
X2.6+ Ожидаемое
X2.6- Неожиданное
X3 Органы чувств
X3.1 Органы чувств: Вкус
X3.1+ Вкусно
X3.1- Невкусно
X3.2 Органы чувств: Звук
X3.2+ Звук: Громко
X3.2- Звук: Тихо
X3.3 Органы чувств: Осязание
X3.4 Органы чувств: Зрение
X3.4+ Видно
X3.4- Не видно
X3.5 Органы чувств: Запах
X3.5- Без запаха
X4 Мысленный объект
X4.1 Мысленный объект: Концепт
X4.1- Беспредметно
X4.2 Мысленный объект: Способ, метод
X5 Внимание
X5.1 Внимание
X5.1+ Внимательно
X5.1- Невнимательно
X5.2 Интерес/скука; возбуждён/энергичен
X5.2+ Заинтересован/возбуждён/энергичен

X5.2- Неинтересно/скучно/вяло
X6 Принимать решение
X6+ Решено
X6- Нерешено
X7 Хотеть; планировать; выбирать
X7+ Желательно
X7- Нежелательно
X8 Пытаться
X8+ Очень стараться
X8- Не пытаться
X9 Способности
X9.1 Способности и интеллект
X9.1+ Способный/умный
X9.1- Неспособность/неумность
X9.2 Успех и неуспех
X9.2+ Успех
X9.2- Неуспех
Y Наука и Техника
Y1 Наука и техника
Y1- Антинаучно
Y2 Информационные технологии и вычислительная техника
Y2- Низкие технологии
Z Имена Собственные и Служебные Слова
Z0 Ненайденные имена собственные
Z1 Личные имена
Z2 Географические названия
Z3 Другие имена собственные
Z4 Область дискурса
Z5 Служебные слова
Z6 Отрицание
Z7 Если
Z7- Безусловное
Z8 Местоимения
Z9 Мусорная корзина
Z99 Ненайденные слова и выражения
.

СПИСОК ЛІТЕРАТУРИ

1. Жуковська В. В. Вступ до корпусної лінгвістики : навч. посіб. / В. В. Жуковська. – Вид-во ЖДУ ім. І.Франка, Житомир, 2013. – 142 с.

2. Копотев М. Введение в корпусную лингвистику : учеб. пособ. / М. Копотев. – Корпусная лингвистика: Animedia Company, 2014. – 230 с.

3. Захаров В. Корпусная лингвистика / В. Захаров, С. Богданова. – Litres, 2020. – 148 с.

4. McEnery T. Corpus-based Language Studies: An Advanced Resource Book / T. McEnery, R. Xiao, Yu. Tono. – London: Routledge, 2006. – 408 p.

5. Tognini-Bonelli E. Theoretical overview of the evolution of corpus linguistics / E. Tognini-Bonelli // Routledge Handbook of Corpus Linguistics. – Abingdon: Routledge, 2010. – P. 14–27.

6. Software: AntConc — a freeware corpus analysis toolkit for concordancing and text analysis [Електронний ресурс]. – Режим доступу : http://www.laurenceanthony.net/software.html

7. CLAWS part-of-speech tagger for English [Електронний ресурс]. – Режим доступу : http://ucrel.lancs.ac.uk/claws/

8. Free CLAWS web tagger [Електронний ресурс]. – Режим доступу : http://ucrel-api.lancaster.ac.uk/claws/free.html

ЗМІСТ

Навчальне видання

МЕТОДИЧНІ ВКАЗІВКИ
до виконання лабораторних робіт з курсу
«Корпусна лінгвістика»
для студентів спеціальності
«Прикладна та комп'ютерна лінгвістика»
Частина 2

Укладачі: ХАЙРОВА Ніна Феліксівна
ПЕТРАСОВА Світлана Валентинівна
ОРОБІНСЬКА Олена Олександрівна

Відповідальний за випуск *проф. Н. В. Шаронова*

Роботу до видання рекомендував *проф. М. І. Безменов*

В авторській редакції

План 2021 р., поз. 215

Підп. до друку 29.09.2021. Формат 60×84 1/16. Папір офсетний.
Riso-друк. Гарнітура Times New Roman. Ум. друк. арк. 3,2.
Наклад 100 прим. Зам. №6/10/21. Ціна договірна.

Видавець Видавничий центр НТУ «ХПІ».
Свідоцтво про державну реєстрацію ДК № 5478 від 21.08.2017 р.
61002, Харків, вул. Кирпичова, 2

Виготовлювач: ФОП Панов А.М.
Свідоцтво серії ДК № 4847 від 06.02.2015 р.
м. Харків, вул. Жон Мироносиць, 10, оф. 6,
тел.: +38(057)714-06-74, +38(050)976-32-87
vdele.in.ua copy@vlavke.com