

# *Корпусна лінгвістика*

A decorative horizontal bar consisting of a thick teal line, a thin white line, and another thick teal line, spanning the width of the slide.

□ **Корпусна лінгвістика** – розділ мовознавства, що займається створенням, обробкою та використанням корпусів.

Корпусна лінгвістика сьогодні має дві лінії розвитку – **лінгвістичний аналіз тексту і інформаційний аналіз тексту.**

- ***Корпус текстів*** - це вид корпусу даних, одиницями якого є тексти або їх достатньо значні фрагменти, що включають, наприклад, якісь повні фрагменти макроструктури текстів даної проблемної області.

## Цінність корпусу вбачається в наступному:

- одного разу зроблений корпус може багато разів використовуватися;
- корпус показує мовні дані в їх реальному оточенні;
- корпус характеризується показністю, або збалансованим складом текстів, що дозволяє використовувати його для тестування пошукових машин, машинних морфологій, систем перекладу, а також використовувати його в різних лінгвістичних дослідженнях;
- корпус має важливе значення для викладання мови.

## *Електронні бібліотеки та їх різноманіття:*

- Корпус латинських текстів “Персей”.
- Корпус текстів Ф. М. Достоевського.
- Електронна енциклопедія “Брокгауз і Ефрон”.
- Фундаментальна електронна бібліотека.
- Російська віртуальна бібліотека.
- Бібліотека М. Мошкова.
- Електронна бібліотека Хімічного фак-ту МГУ.

# Лінгвістичні корпуси

- Brown Corpus.
- Ланкастерський корпус англійської мови (Lancaster-Oslo-Bergen Corpus, LOB).
- British National Corpus.
- International Corpus of English.
- Bank of English.
- Cobuild Corpus.
- Мангеймський корпус Німецької мови.
- Чеський національний корпус.
- Національний корпус російської мови.
- Корпуси китайської, турецької, естонської, албанської і багатьох інших мов

# Корпус і електронна бібліотека

Тексти в корпусах розглядаються перш за все як зразки текстів.

Тексти в електронних бібліотеках, виходячи з призначення, краще за все називати творами з усіма характерними для них рисами.

<u>Лінгвістичний корпус текстів:</u>	<u>Електронна бібліотека:</u>
Зразки текстів	повні тексти
лінгвістична розмітка	бібліографічні та історико-культурні елементи даних (якщо є)
лінгвостатистика	Відсутність статистики
Репрезентативність мовного матеріалу «умовна»	повнота текстів електронної бібліотеки
Відбір мовного матеріалу на основі критеріїв репрезентативності, лінгвістичної та історико-культурної значимості	відбір текстів, визначений матеріалом бібліотеки

# Корпус

Власне корпус  
(накопичені дані)

+

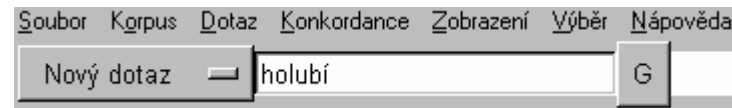
корпусний менеджер  
(спеціалізована пошукова система)



# Конкорданс

□ **Конкорданс** є спеціалізованою лінгвістичною прикладною програмою, за допомогою якої здійснюється автоматична вибірка заданих мовних одиниць з електронних текстів.

**Чеський національний корпус: Пошук використання слова *holubí* (голубиний).**



jej . Hedvábným slunečníkem **holubí** barvy procházelo slunce a  
 podobné zábrany rovněž slabé a **holubí** samec dokáže v těsné kleci  
 řeklo by se , že vybírala **holubí** vejce , to se o ní vědělo  
 odpověděl , " červené jako **holubí** nožky , červenější než ohromné  
 uniformu tvořila tunika v barvě **holubí** šedi , která se oblékala  
 znám jitra nadšená jak hejna **holubí** , a časem viděl jsem , co

***Призначення корпусу*** – показати функціонування лінгвістичних одиниць в їх природньому контекстному оточенні.

На основі корпусу можна одержати дані:

- ✓ про частоту словоформ, лексем, граматичних категорій,
- ✓ про зміну частот
- ✓ про зміни контекстів в різні періоди часу
- ✓ про поведінку мовних одиниць різних авторів
- ✓ про особливості їх сполучованості, управління
- ✓ і т.д.

# Типи корпусів

- Корпуси, що відносяться до однієї мови;
- корпуси, що відносяться до підмови(жанр, стиль, мова певного автора).

Існує багато інших типів корпусів.

Можна виділити такі **класифікації корпусів.**

# Класифікація корпусів

Ознака	Типи корпусів
Тип даних	<ul style="list-style-type: none"><li>• Письменні</li><li>• Мовні</li><li>• Змішані</li></ul>
Мова текстів	<ul style="list-style-type: none"><li>• Українська</li><li>• Англійська і т.д.</li></ul>
«Паралельність»	<ul style="list-style-type: none"><li>• Одномовні</li><li>• Двомовні</li><li>• Багатомовні</li></ul>
«Літературність», специфічність	<ul style="list-style-type: none"><li>• Літературні</li><li>• Діалектні</li><li>• Розмовні</li><li>• Термінологічні</li><li>• Змішані</li></ul>
Жанр	<ul style="list-style-type: none"><li>• Літературні</li><li>• Фольклорні</li><li>• Драматичні</li><li>• Публіцистичні</li></ul>

# Класифікація корпусів

Ознак	Типи корпусів
Доступність	<ul style="list-style-type: none"><li>• Вільно доступні</li><li>• Комерційні</li><li>• Закриті</li></ul>
Призначення	<ul style="list-style-type: none"><li>• Дослідницькі</li><li>• Ілюстративні</li></ul>
Динамічність	<ul style="list-style-type: none"><li>• Динамічні (моніторні)</li><li>• Статичні</li></ul>
Розмітка	<ul style="list-style-type: none"><li>• Розмічені</li><li>• Нерозмічені</li></ul>
Характер розмітки	<ul style="list-style-type: none"><li>• Морфологічні</li><li>• Синтаксичні</li><li>• Семантичні</li><li>• Просодичні.</li></ul>
Об`єм текстів	<ul style="list-style-type: none"><li>• Повнотекстові</li><li>• «Фрагментнотекстові»</li></ul>

# Користувачі

- Прикладні лінгвісти;
- лексикографи;
- лінгвісти-теоретики;
- викладачі;
- комп'ютерні лінгвісти;
- інші спеціалісти з мови (літературознавці, редактори);
- Корпуси як інструмент для розробки і налаштування різних автоматизованих систем (машинний переклад, розпізнавання мови, інформаційний пошук).

# Корпус мови Івана Франка

(Тимчишин О.В.)  
Національний університет  
«Львівська політехніка»  
(Комп'ютерні науки))

- В процесі створення електронної бібліотеки відскановано **50 томів (більше 25,000 сторінок)**, з допомогою інструментів бібліотеки **DjVuLibre** створено електронні книги у форматі **DjVu**, який повністю зберігає поліграфічне оформлення (як у факсимільному виданні); так званий «прихований текстовий шар» у файлах забезпечує можливість повнотекстового пошуку у текстах книг.



# Попередня підготовка текстів

Djview - tom\_1.djvu

File Edit View Go Settings Help

150% 27

В неї сили нема.  
27 марта 1880

II

**Гримить!** Благодатна пора наступає,  
Природу розкішная дрож пронимає,  
Жде спрагла земля плодотворної зливи,  
І вітер над нею гуляє бурхливий,  
І з заходу темная хмара летить —  
**Гримить!**

**Гримить!** Тайна дрож пронимає народи,—  
Мабуть, благодатная хвиля надходить...  
Мільйони чекають щасливої зміни,  
Ті хмари — плідної будучини тіни,  
Що людськість, мов красна весна, обновить...  
**Гримить!**

1880

III

Гріє сонечко!  
Усміхається небо яснее,  
Дзвонять пісеньку жайворончок,

Page 27 (4 hits)  
Page 420 (2 hits)  
Page 500 (1 hit)

P27 2826x4341 600dpi x=1867 y=2730

# Поділ текстів на речення

- *В основу методу покладено наступні правила:*
  - 1. Потенційний знак кінця, після якого виступає знак пунктуації, не є знаком кінця.
  - 2. Потенційний знак кінця, після якого виступає мала літера, не є знаком кінця.
  - 3. Крапка, оточена з обох боків цифрами не є знаком кінця.
  - 4. Крапка, перед якою є скорочення, що пишеться без крапки є кінцем речення.
  - 5. Крапка після скорочення, що вимагає доповнення, не є кінцем речення.
  - 6. Ініціал (велика літера з крапкою) не є кінцем речення. Якщо ініціал є скороченням від імені (так є в більшості випадків), то слово, що йде після нього, з великою імовірністю є прізвищем, отже тут немає межі речення.

# • Локалізація корпусного менеджера РОЛІQАРР

```

• <?xml version="1.0" encoding="UTF-8"?>
• <cesAna>
• <chunkList>
• <chunk type = "p">
• <tok>
• <orth>ОПОВІДАННЯ</orth>
• <lex disamb="1">
• <base>ОПОВІДАННЯ</base>
• <ctag>dummy</ctag>
• </lex>
• </tok>
• <tok>
• <orth>ДЛЯ</orth>
• <lex disamb="1">
• <base>ДЛЯ</base>
• <ctag>dummy</ctag>
• </lex>
• </tok>
• <tok>
• <orth>ДОМАШНЬОГО</orth>
• <lex disamb="1">
• <base>ДОМАШНЬОГО</base>
• <ctag>dummy</ctag>
• </lex>
• </tok>
• <tok>
• <orth>ОГНИЩА</orth>
• <lex disamb="1">
• <base>ОГНИЩА</base>
• <ctag>dummy</ctag>
• </lex>
• </tok>

```

• Рисунок 2. Приклад фрагменту тексту у форматі XCES

# Компіляція корпусу

Роліагр

файл Статистика Налаштування Допомога

[base="капітан.\*"/i] Виконати

Лівий контекст	Заданий фрагмент	Правий контекст
, так , — мовив	капітан	, мимоволі відхиляючи голову
першим . — І пані	капітанова	також там жила тоді ,
— І дитина у пана	капітана	справді хора ? — Не
собою самим пройшло по	капітановій	душі , коли виголошував сю
, ге , ге !	Капітан	аж зубами заскреготав і щосил
. Ну , але пан	капітан	обіцяли мені щось сказати і
то що таке ? Пан	капітан	мали мені сказати , для
торговлею дівчатами ?	Капітанові	похололо коло серця . Чув
хвилю скоса позирати на	капітана	. — Знають пан капітан
капітана . — Знають пан	капітан	, я се пану капітанові
І Гірш положив руку на	капітанових	плечах , а що у
балакання був такий , що	капітан	уже не міг встати ,
— Я знаю , пан	капітан	є добрий чоловік , службовий

? — Се вже моє діло . Ну , але пан капітан обіцяли мені щось сказати і не сказали .  
 — Що таке ? — Як то що таке ? Пан капітан мали мені сказати , для чого так  
 інтересуються тою брудною справою ... тою торгівлею дівчатами ? **Капітанові**  
 похололо коло серця . Чув , що Гірш помаленьку , та певною рукою вбиває йому ніж  
 у груди . В його голові мішалось . — Ах , се би довго оповідати ... — Пощо довго ?  
 Пощо довго ? — по-своєму всміхаючися , щдив Гірш , не перестаючи

Вивід результатів 401 - 450 (зі 500) Метадані ↑ ↓

# Корпусна лінгвістика один із нових продуктивних шляхів розвитку мовознавства...

Дякую за увагу