Корпусна лінгвістика

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. Достоєвського.
- 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 англійскої мови (Lancaster-Oslo-Bergen Corpus, LOB).
- British National Corpus.
- International Corpus of English.
- Bank of English.
- Cobuild Corpus.
- Мангеймський корпус Німецької мови.
- Чеський національный корпус.
- Національный корпус російської мови.
- Корпуси китайскої, турецкої, эстонскої, албанскої 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.

Лінгвістичний корпус текстів:	***Електронна бібліотека***:
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 матеріала на основі критеріїв репрезентативності, лінгвістичної та історико-культурної значимості	відбір тексті, 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 функціонурованння 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 сполучуваності, 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	•Письменні •Мовні •Змішані
Мова текстів	•Українська •Англійська і т.д.
«Паралельність»	•Одномовні •Двомовні •Багатомовні
«Літературність»,специфічність	•Літературні •Діалектні •Розмовні •Термінологічні •Змішані
Жанр	•Літературні •Фольклорні •Драматичні •Публіцистичні

Класифікація корпусів

Ознак	Типи корпусів
Доступність	•Вільно доступні •Комерційні •Закриті
Призначення	•Дослідницькі •Ілюстративні
Динамічність	•Динамічні (моніторні) •Статичні
Розмітка	•Розмічені •Нерозмічені
Характер розмітки	•Морфологічні •Синтаксичні •Семантичні •Просодичні.
Об`єм текстів	•Повнотекстові •«Фрагментнотекстові»

Користувачі

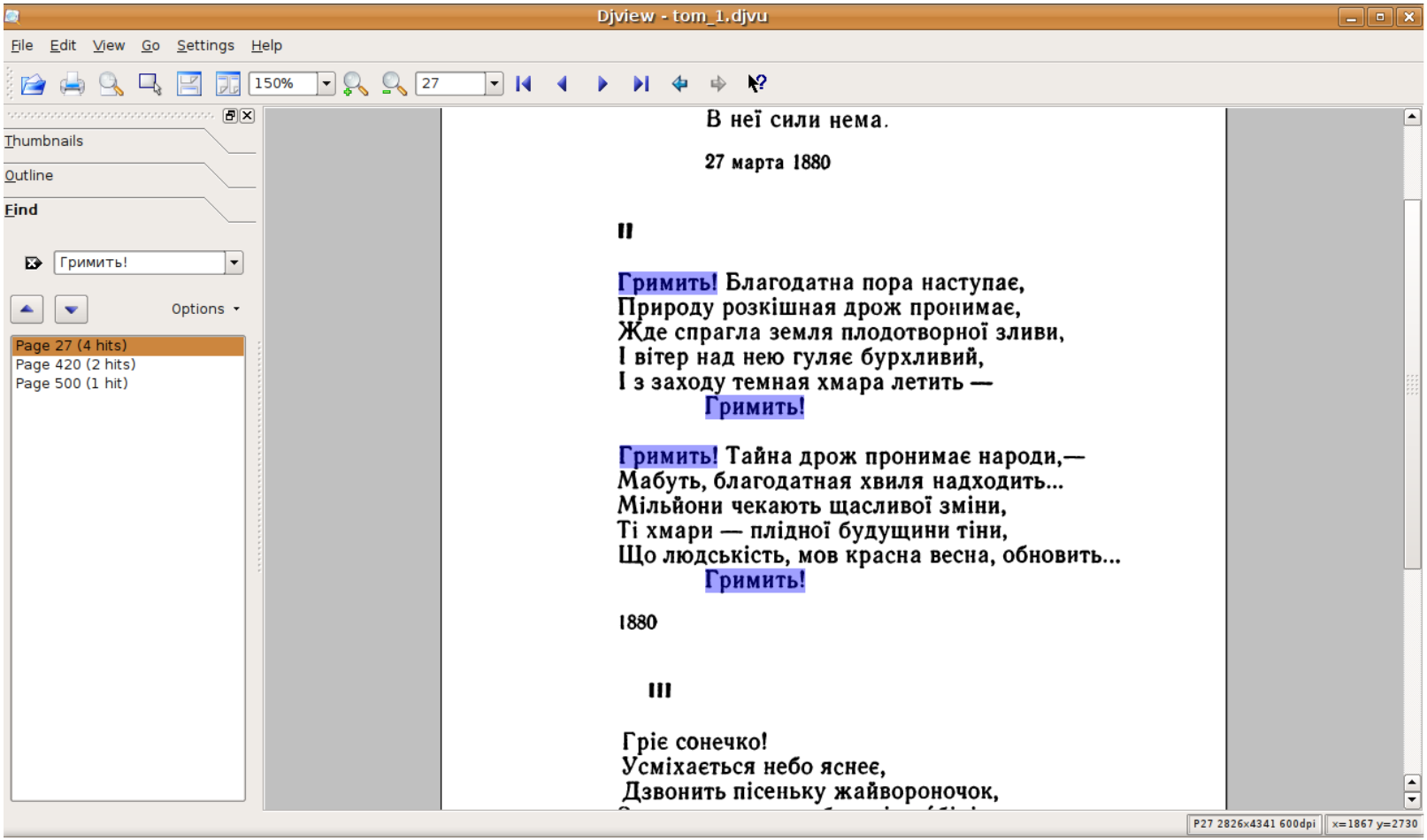
- Прикладні лінгвісти;
- лексикографи;
- лінгвісти-теоретики;
- викладачі;
- комп`ютерні лінгвісти;
- інші спеціалісти з мови (літературознавці, редактори);
- Корпуси як інструмент для розробки і налаштування різних автоматизованих систем (машинний переклад, розпізнавання мови, інформаційний пошук).

Корпус мови Івана Франка

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

- В процесі створення електронної бібліотеки відсканована **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



Поділ текстів на речення

- *В основу методу покладено наступні правила:*

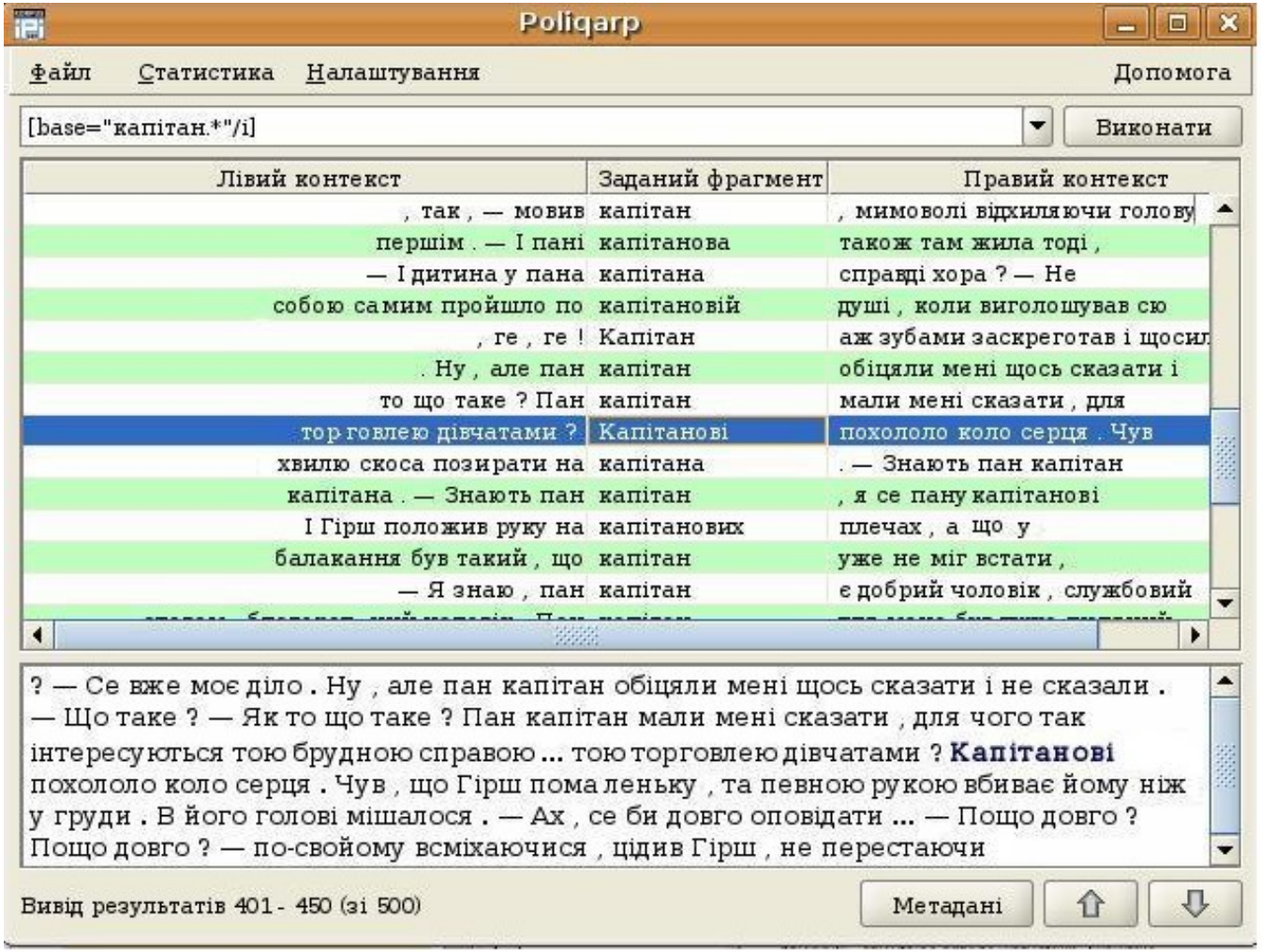
- 1. Потенційний знак кінця, після якого виступає знак пунктуації, не є знаком кінця.
- 2. Потенційний знак кінця, після якого виступає мала літера, не є знаком кінця.
- 3. Крапка, оточена з обох боків цифрами не є знаком кінця.
- 4. Крапка, перед якою є скорочення, що пишеться без крапки є кінцем речення.
- 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 POLIQARP

```
<?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



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

Дякую за увагу