Використання корпусних технологій у дослідницькій та навчальній діяльності

М.В. Надутенко

www.ulif.org.ua
http://lcorp.ulif.org.ua/dictua/

Результати та надбання Українського мовно-інформаційного фонду НАН України на поприщі лінгвістичних технологій

Результати та надбання Українського мовно-інформаційного фонду НАН України

- Відкрито лексикографічний ефект в інформаційних системах.
- Побудовано інформаційно-структурну теорію лексикографічних систем.
- Побудовано теорію інтегрованих, індексованих та віртуалізованих Л-систем.
- Розроблено концепцію віртуальних систем професійної взаємодії (спільно з Інститутом кібернетики НАНУ), теорію та технологію віртуальних лексикографічних лабораторій.

Результати та надбання Українського мовно-інформаційного фонду НАН України

- Розроблено парадигматичну словозмінну Л-систему української мови.
- Виділено близько 3000 парадигматичних класів української лексики.
- Створено Граматичну лексикографічну базу даних обсягом понад 570 тисяч одиниць та відповідну Віртуальну лексикографічну лабораторію (електронний Граматичний словник української мови).
- Розроблено алгоритми побудови фонематичної транскрипції для української лексики та побудовано базу транскрибованої словозмінної парадигми на масиві електронного Граматичного словника української мови.

Результати та надбання Українського мовно-інформаційного фонду НАН України

- Розроблено парадигматичну словозмінну Л-систему російської мови.
- Створено граматичну лексикографічну базу даних обсягом понад 180 тисяч одиниць та відповідну Віртуальну лексикографічну лабораторію.

Результати та надбання Українського мовно-інформаційного фонду НАН України

- Побудовано теорію іменної словозміни турецької мови.
- Кваліфіковано іменну словозмінну парадигму турецької мови обсягом 1808 граматичних значень.
- Досліджено фонетичні процеси та розроблено алгоритми побудови повної словозмінної парадигми турецького іменника.
- Створено ВЛЛ «Граматичний словник турецької мови» та «Тлумачний словник турецької мови».

Результати та надбання Українського мовно-інформаційного фонду НАН України

- Німецька мова. Розроблено парадигматичну словозмінну Л-систему.
- Іспанська мова. Розроблено парадигматичну словозмінну Л-систему.
- Створено теорію семантичних станів мовних одиниць, яка надає концептуальний апарат для опису лінгвістичних неоднозначностей різних типів та інтеграції граматичного та лексикографічного типів опису мовної системи.

Результати та надбання Українського мовно-інформаційного фонду НАН України

- Розроблено теорію концептографічних систем. Реалізовано системну концептографію Святого Письма (Євангелія).
- Розроблено систему лінгвістичних алгоритмів автоматичного аналізу українських текстів, яка включає: передморфологічний, морфологічний, синтаксичний й частково семантичний аналіз.
- Побудовано теорію лексикографічної системи українського дієслова та іменника.
- Розроблено лінгвістичну теорію еквівалентів слова. На основі цієї теорії створено цілу низку словників еквівалентів слова.

Результати та надбання Українського мовно-інформаційного фонду НАН України

- Розроблено теорію лексикографічної системи фундаментального тлумачного Словника української мови.
- Розроблено теорію лексикографічних середовищ, яка стала теоретичним підґрунтям для побудови інтегрованих лексикографічних систем.
- За допомогою цієї теорії було розроблено **Інтегровану лексикографічну систему «Словники України»** - найбільший і найдовершеніший український електронний словник.

Результати та надбання Українського мовно-інформаційного фонду НАН України

- Розроблено теорію лінгвістичного корпусу української мови. Сформовано УНЛК обсягом близько 180 млн. слововживань.
- Розроблено інтегровану систему двомовної слов'янської лексикографії «МОНДІЛЕКС». Створено 30 двомовних ВЛЛ (мови болгарська, польська, російська, словацька, словенська, українська).
- Розроблено теорію лінгвістичних онтологій.
- Створено Український лінгвістичний портал.

Результати та надбання Українського мовно-інформаційного фонду НАН України

- Розроблено теорію та технологію віртуалізації лексикографічних систем. Здійснено реалізацію УНЛК у хмаринному середовищі та на суперкомп’ютері СКІТ-3.
- Сформульвано та обґрунтовано концепцію системи Національних лінгвістичних ресурсів та Національної лінгвістичної інфраструктури.

Електронні ресурси Національної словникової бази (Віртуальні лексикографічні лабораторії, лінгвістичні корпуси, словники, платформи «LEX» та «ULISS»)

УКРАЇНСЬКИЙ
МОВНО-ІНФОРМАЦІЙНИЙ
ФОНД НАН УКРАЇНИ

РЕСУРСИ на сайті Українського мовно-інформаційного фонду НАН України

- «Словники України on-line»
- Тлумачний словник української мови у 20-ти томах (попередня версія)
- Тлумачний словник української мови у 20-ти томах *
- Система лінгвістичної взаємодії «ВЛЛ»
- «Український національний лінгвістичний корпус»

- Система лінгвістичної взаємодії «ВЛЛ» (нова версія)
- Словник української мови за редакцією Б.Д. Грінченка *
- Словарь русского словоизменения
- Віртуальна лексикографічна система "MONDILEX"
- Віртуальна лексикографічна система "Turkic"
- Український біографічний архів

- Сайт "Українсько-російсько-англійський словник зі зварювання" (веб версія)
- Російсько-українсько-англійський словник зі зварювання

- Російсько-українсько-англійський словник з механіки
- Російсько-українсько-англійський словник з механіки (версія 2)

- Російсько-українсько-англійський словник з радіоелектроніки

- Віртуальна термінографічна лабораторія з фізики
- Віртуальна термінографічна лабораторія з біології
- Віртуальна лексикографічна лабораторія «Українсько-кримськотатарський словник»

- "Словники України" (Версія з диску 4.0 (мережева))

- ВЛЛ «Словник української мови»
- ВЛЛ «Етимологічний словник української мови»
- УКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ ЛІНГВІСТИЧНИЙ КОРПУС
- ВЛЛ «Словарь русского языка»
- ВЛЛ «Словник турецької мови»
- ВЛЛ «Граматичний словник української мови»
  ВЛЛ «Граматичний словник російської мови»
- ВЛЛ «Граматичний словник турецької мови»
- ВЛЛ «Граматичний словник німецької мови»
- ВЛЛ «Граматичний словник іспанської мови»
- ВЛЛ «Словник синонімів української мови»
- ВЛЛ «Українська ономастика»

- 30 ВЛЛ «MONDILEX» для укладання двомовних словників у системі мов: болгарська, польська, російська, словацька, словенська, українська.
- ВЛЛ «Українсько-кримськотатарський словник»
- ВТЛ «ЗВАРЮВАННЯ»
- ВТЛ «ФІЗИКА»
- Інтегрована лексикографічна система «Словники України»
- Українсько-російський та Російсько-український словник зі зварювання
- Українсько-російсько-англійський словник зі зварювання
- Українсько-російсько-англійський термінологічний словник-довідник «Зварювання»
- Російсько-українсько-англійський словник з механіки
- Російсько-українсько-англійський словник з радіоелектроніки
- Шестимовний словник металургійних термінів (українсько-грузинсько-російсько-англійсько-французько-німецький)

Для сучасних досліджень в галузі прикладної лінгвістики, Big Data та інформаційних потоків

необхідним є такий інструмент як

Лінгвістичний Корпус.

Зокрема, Лінгвістичний Корпус Google`а, на якому проводяться основні дослідження, має обсяг приблизно 15 трильйонів слововживань.

ЩО ТАКЕ КОРПУС?

Лінгвістичний корпус – інформаційно-довідкова система, яка складається із певних текстів в електронній формі. Це безкоштовний сайт, створений групою вчених Українського мовно-інформаційного фонду НАН України та Національного центру Мала академія наук. Корпус постійно поповнюється й розвивається.

Лінгвістичні корпуси

- **Корпус Google:**

Понад 1 000 000 000 000 000 слововживань

- **Національні корпуси – декілька сотень мільйонів (до 10^9) слововживань**

- **Український національний лінгвістичний корпус – 180 млн. слововживань (віртуалізована платформа ULISS)**

Особливості корпусу:

Корпус має дві важливі особливості, які відрізняють його від інших електронних бібліотек :

1. **Репрезентативність та збалансованість складу текстів** - корпус містить практично всі типи текстів:

- класична та сучасна художня література - проза й поезія;
- словники (близько 240);
- мемуарно-біографічна література;
- журнальна публіцистика й літературна критика;
- газетна публіцистика й новини;
- наукові, науково-популярні й навчальні тексти;
- релігійні й релігійно-філософські тексти;
- виробничо-технічні, офіційно-ділові і юридичні тексти.

2. **Має розмітку.** Розмітка надає особливу додаткову інформацію про властивості текстів, які входять до корпусу. Професійно зроблена розмітка дозволяє швидко та ефективно знайти ті слова, форми та конструкції, які потрібні дослідникові.

Як можна використовувати лінгвістичний корпус?

- підготовка навчальних матеріалів для занять: розроблення вправ, тестів та контрольних;

- організація самостійної роботи учнів;
- організація науково-дослідної роботи.

Як можна використовувати лінгвістичний корпус?

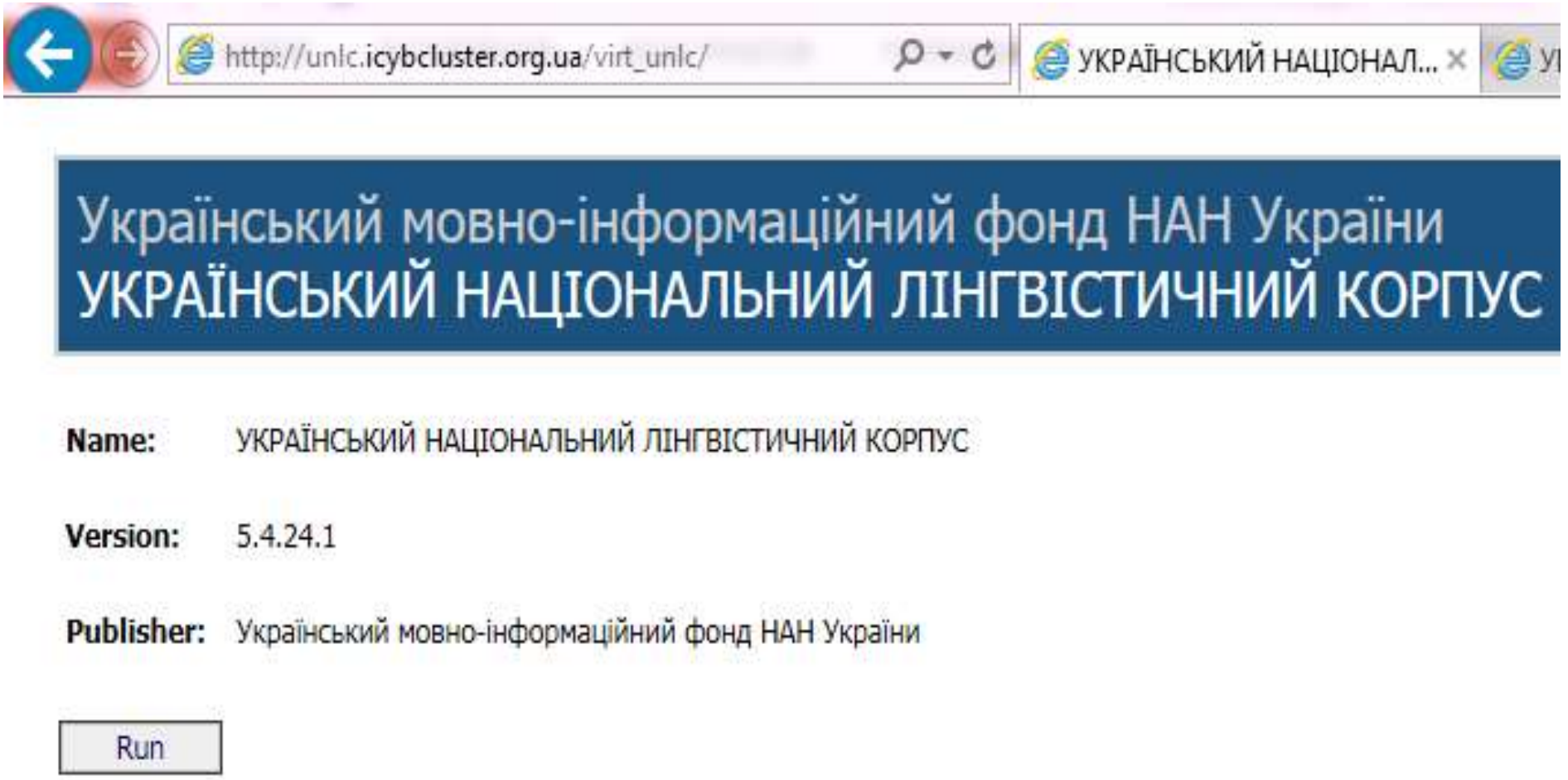
- - наводити приклади з корпусу в роздаткових матеріалах;
- - продемонструвати пошукові можливості корпусу;
- - роздавати приклади для інтерпретації;
- - домашнє завдання - провести пошук у корпусі й представити зібраний матеріал;
- - наукові дослідження учнів – визначити тему, зібрати в корпусі матеріал й інтерпретувати його.

- Можна вибрати короткі приклади;
- Користувачі розуміють, що вони можуть обрати тему, зібрати матеріал і інтерпретувати його, - так вони стануть самостійними вченими
- Можна проілюструвати тими або іншими прикладами вживання слів у контексті

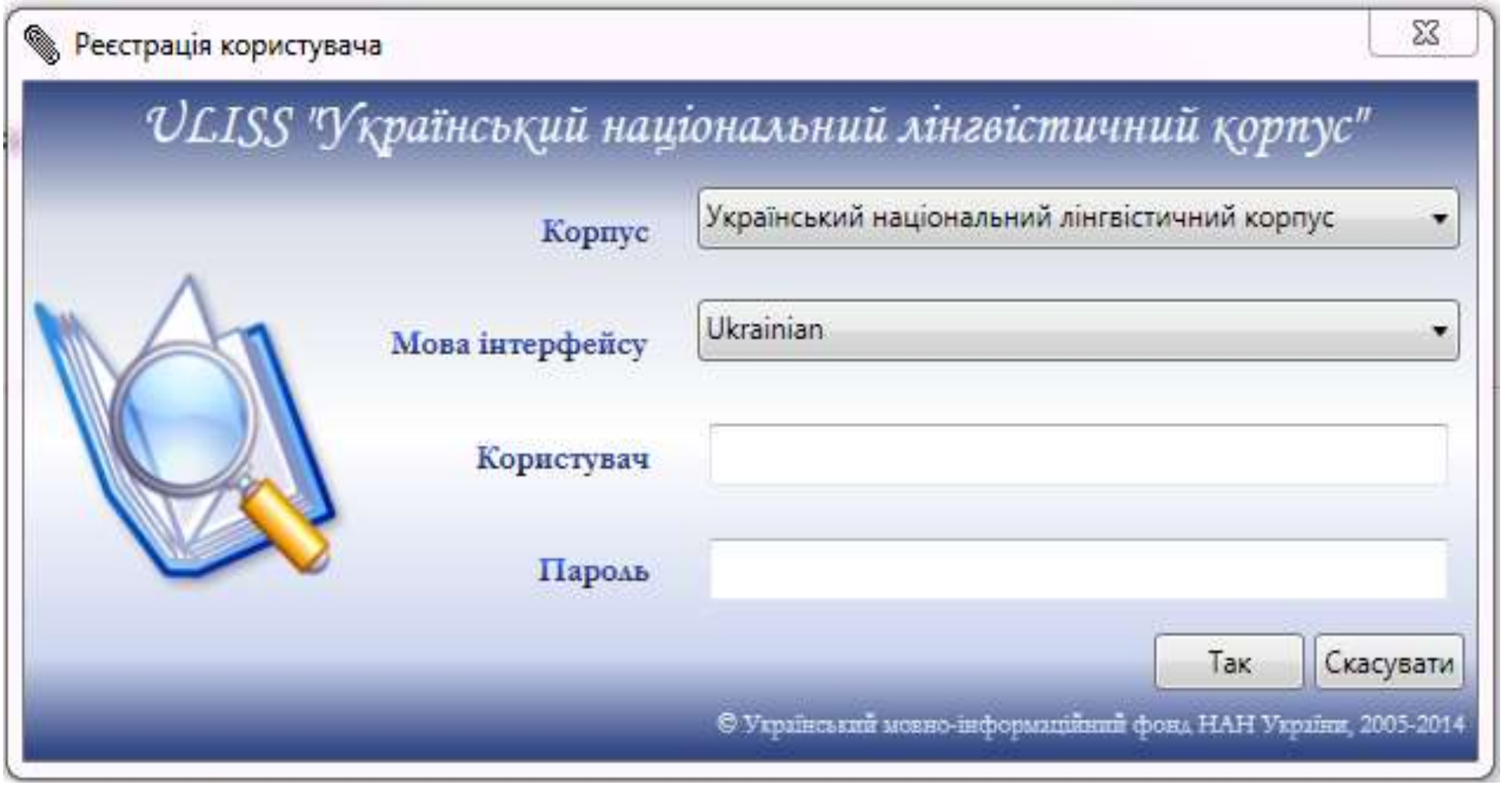
Які можливості надає корпус?

- пошук текстів;
-пошук додаткового матеріалу;
- довідково-інформаційна робота;
- дослідження творчості окремих авторів;
- перевірка правильності вживання слів;
- матеріали для науково-дослідної роботи.

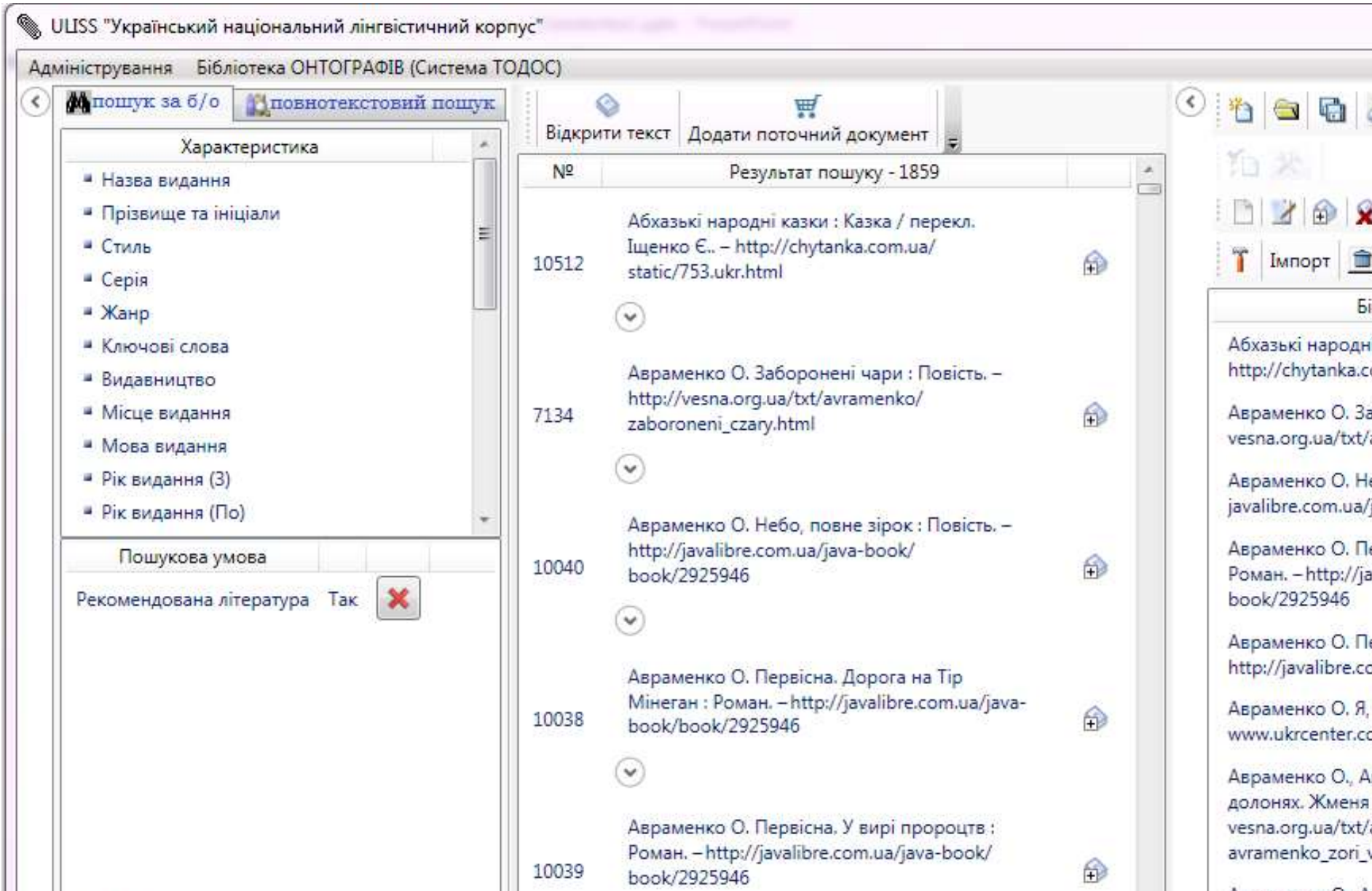
Лінгвістичний корпус



Лінгвістичний корпус



Лінгвістичний корпус



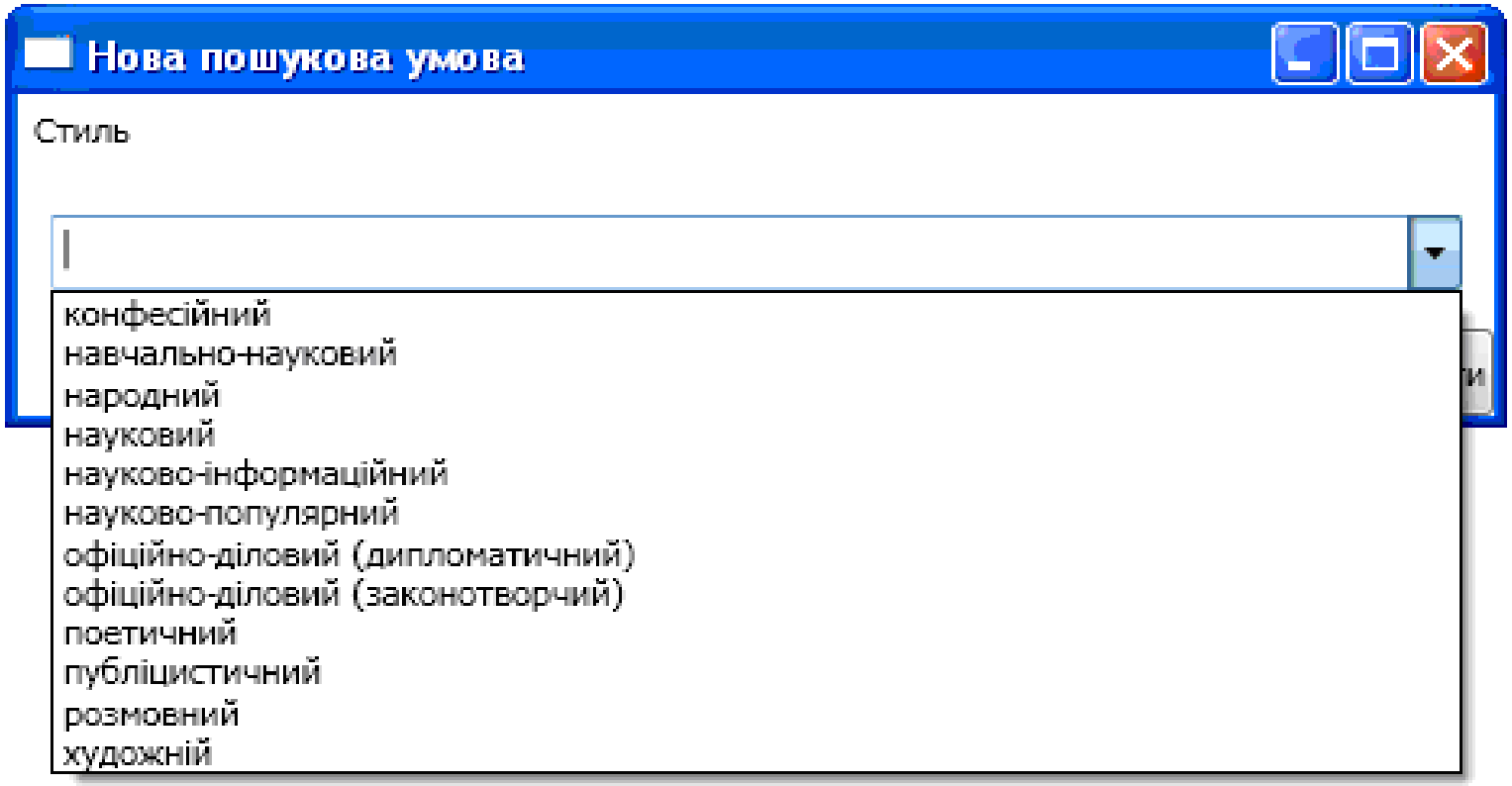
Лінгвістичний корпус: Як здійснюється пошук?

За бібліографічною ознакою:

- Назва видання
- Прізвище та ініціали
- Стиль
- Серія
- Жанр
- Ключові слова
- Видавництво
- Місце видання
- Мова видання
- Рік видання (З)
- Рік видання (По)

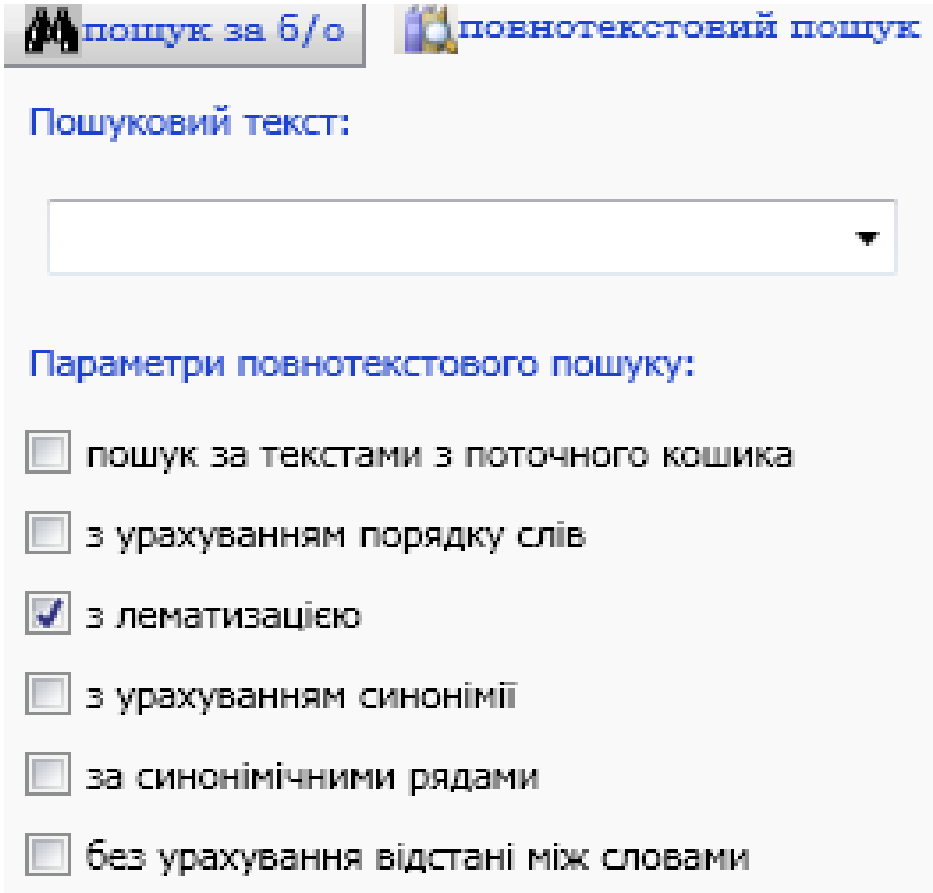
Лінгвістичний корпус: Як здійснюється пошук?

Для кожної ознаки передбачені свої пошукові умови :



Лінгвістичний корпус: Як здійснюється пошук?

Повнотекстовий пошук слова або словосполучення за параметрами:



Приклад завдання з використанням корпусу

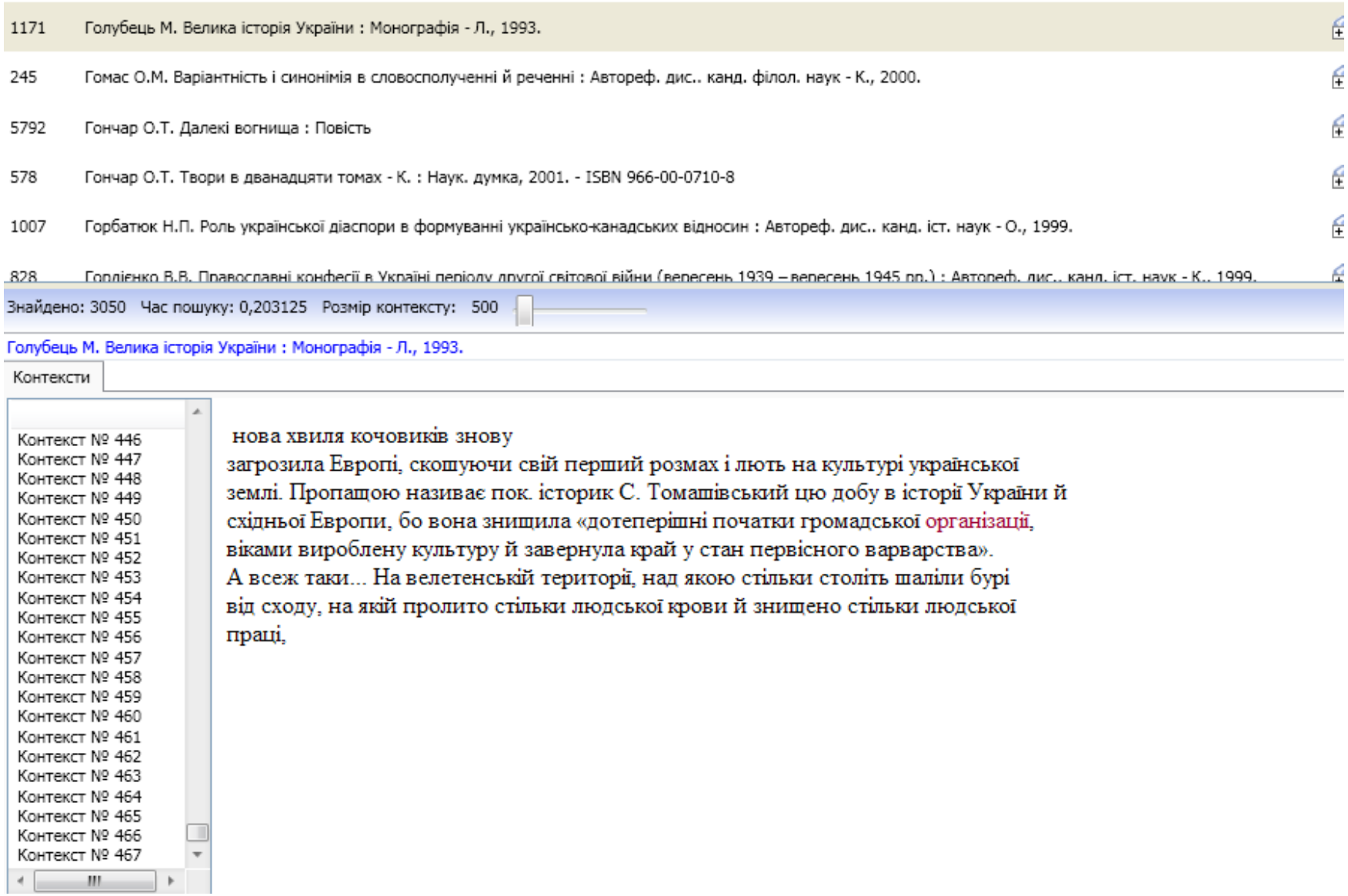
Синоніми, пароніми у діловому мовленні

Пароніми — слова близькі за звуковим складом і звучанням, та різні за значенням. Різниця може бути в одній букві.

Слова «компанія» — «кампанія».

Приклади використання цих слів у текстах

Результати пошуку слова «компанія» з урахуванням синонімії



Результати пошуку слова «кампанія» з урахуванням синонімії

1790	Білоус Д. Чари Барвінкові : Поезії / Бібліотека з української літератури для школярів - К. : Студія "Негоціант", 2000. - (Золота скарбниця України). - ISBN 966-7423-50-6
3337	Бондарчук Ю.П. Залізничний транспорт України в умовах утвердження адміністративно-командних методів управління народним господарством (кін. 20-х — 30-ті рр.) : Автореф. дис.. канд. іст. наук / Дніпропетровський державний університет - К., 1999.
3334	Борисенко В.М. Літературні організації в суспільно-політичному житті України (1920-1932рр.) : Автореф. дис.. канд. іст. наук / Київський університет імені Тараса Шевченка - К., 1999.
1539	Боротьба з організованою злочинністю і корупцією (теорія і практика)т.1 : Журнал; Науково-практичний журнал / Міжвідомчий науково-дослідний центр - К., 2000.
1538	Боротьба з організованою злочинністю і корупцією (теорія і практика)т.2 : Науково-практичний журнал / гол. ред. Кондратьєв Я.Ю.; Міжвідомчий науково-дослідний центр - К., 2000.
1541	Боротьба з організованою злочинністю і корупцією (теорія і практика)т.4 : Науково-практичний журнал / Міжвідомчий науково-дослідний центр - К., 2001.

Знайдено: 704 Час пошуку: 0,015625 Розмір контексту: 500

Білоус Д. Чари Барвінкові : Поезії / Бібліотека з української літератури для школярів - К. : Студія "Негоціант", 2000. - (Золота скарбниця України). - ISBN 966-7423-50-6

Контексти

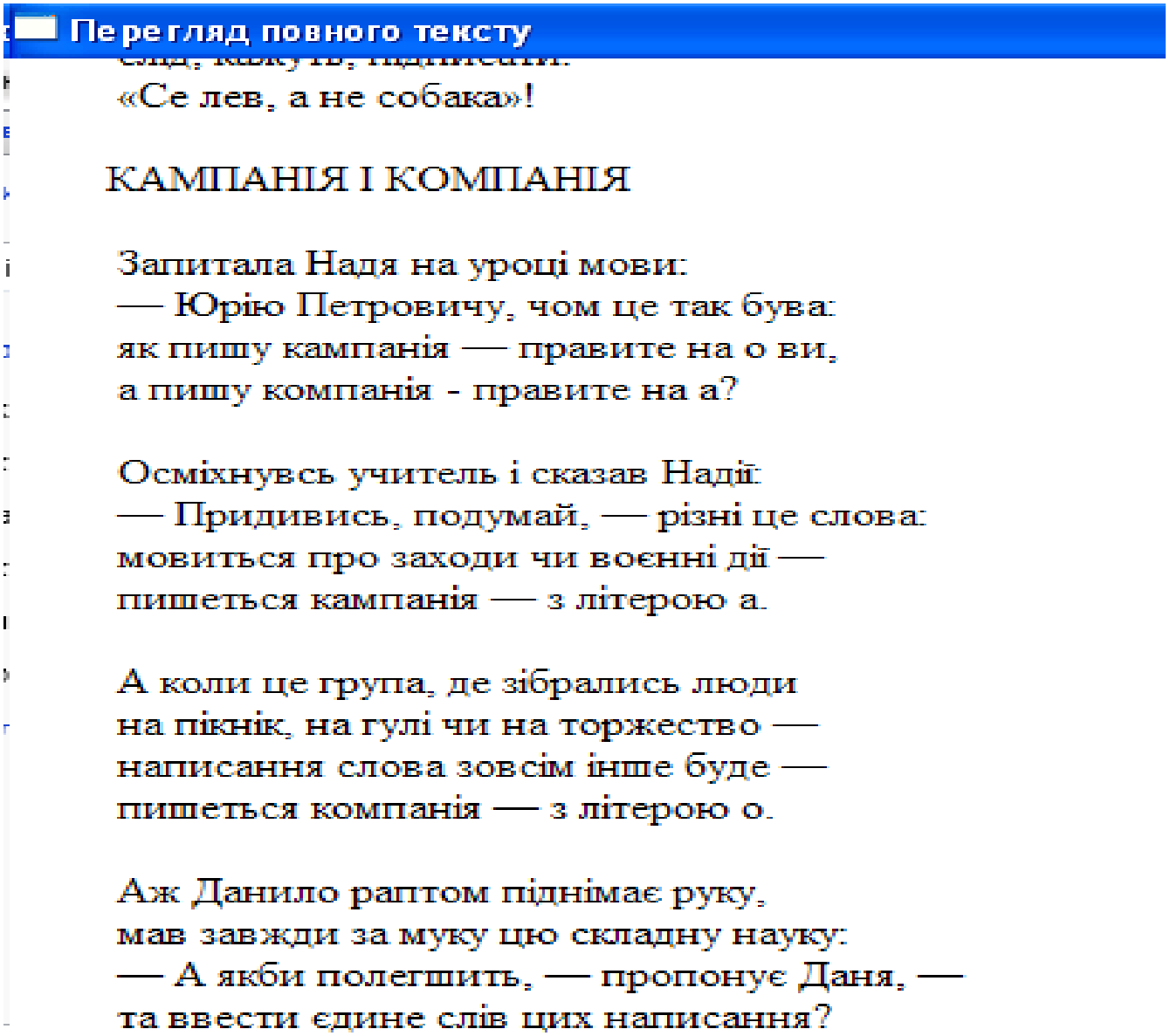
Контекст № 1
Контекст № 2
Контекст № 3
Контекст № 4

Мовляв, щоб розпізнати,
що втнув якийсь чудака,
слід, кажуть, підписати:
«Се лев, а не собака»!

КАМПАНІЯ І КОМПАНІЯ

Запитала Надя на уроці мови:
— Юрію Петровичу, чом це так бува:
як пишу кампанія — правите на о ви,

Перегляд повного тексту



Середовище ТОДОС

Середовище ТОДОС (Трансдисциплінарні Онтологічні Діалоги Об'єктно-орієнтованих Систем) - створене з використанням корпусних технологій та лінгвістичних онтологій

Трансдисциплінарність – спосіб розширення наукового світогляду, що полягає в розгляді того чи іншого явища поза рамками будь-якої однієї наукової дисципліни. За допомогою Т. здійснюється класифікація і систематизація формального взаємозв'язку окремих дисциплінарних знань.

Онтологія - є відображенням певної теорії й може бути представлена як активна система знань, яка включає в себе множину об'єктів (позначених певними термінами), що пов'язані з описами, а також формальні аксіоми, які обмежують інтерпретацію і спільне вживання цих термінів.

Діалог - форма і організація спілкування, комунікації, обміну висловлюваннями.

Об'єктно-орієнтований підхід - спеціальним способом організовані множини, об'єкти яких мають загальні властивості.

Середовище ТОДОС

- забезпечує можливість для автоматизованого подання знань у вигляді динамічної онтології – сукупності термінів дисципліни (чи будь-якої їх кількості) та зв’язків між ними.

ВІДЕО

ДЯКУЮ ЗА УВАГУ!