УДК 81’33(075)

Рекомендовано до друку вченою радою філологічного факультету Донецького національного університету імені Василя Стуса (протокол №12 від 17 червня 2020 р.).

Рецензенти:

Анатолій Приходько, доктор філологічних наук, професор (Запорізький національний технічний університет).

Михайло Торчинський, доктор філологічних наук, професор (Хмельницький національний університет).

Жанна Краснобаєва-Чорна, доктор філологічних наук, професор (Донецький національний університет імені Василя Стуса).

Гарбера І.В.

Прикладна морфологія: основи автоматичного морфологічного аналізу тексту: Навчально-методичний посібник. Вінниця, 2020. 194 с.

Пропонований посібник подає основи автоматичного морфологічного аналізу тексту, який є важливим напрямом теоретичних і прикладних досліджень сучасної комп’ютерної лінгвістики. Посібник містить три змістові модулі «Теорія автоматичного морфологічного аналізу», «Основні методи автоматичного морфологічного аналізу» та «Сучасні автоматичні морфологічні аналізатори», що включають відповідні теоретичні відомості, супроводжувані ілюстраціями, завдання для самоконтролю та вказівки до виконання лабораторних робіт.

Посібник адресований студентам-філологам, які здобувають спеціальність «Прикладна лінгвістика», викладачам вузів, аспірантам, усім, хто цікавиться проблемами автоматичного опрацювання текстової інформації.

ПЕРЕДМОВА

Автоматичний морфологічний аналіз тексту становить собою ключовий етап комп'ютерного опрацювання текстової інформації, є основою ефективного здійснення морфемного, семантичного та синтаксичного аналізу. Створення й використання сучасних автоматизованих систем опрацювання тексту неможливі без усвідомлення принципів, закономірностей та особливостей побудови алгоритмів автоматичного морфологічного аналізу.

Пропонований навчально-методичний посібник покликаний допомогти усім зацікавленим узагальнити теоретичні знання з прикладної морфології та навчитися застосовувати їх практично. З цією метою видання структуровано відповідним чином: змістовий модуль I «Теорія автоматичного морфологічного аналізу» містить розгляд основних теоретичних положень автоматичного морфологічного аналізу (його визначення, завдання, перспективи, основні системи, проблеми доморфологічного аналізу, лематизації та стемінгу, граматичної омонімії тощо), змістовий модуль II «Основні методи автоматичного морфологічного аналізу» подає опис найбільш розповсюджених алгоритмів автоматичного морфологічного аналізу (на основі графемного і флективного аналізу, на основі словників словоформ і основ, операції логічного множення), змістовий модуль III «Сучасні автоматичні морфологічні аналізатори» презентує опис і характеристику функціоналу найбільш розповсюджених сьогодні програмних засобів, що використовують морфологічні модулі (LanguageTool, libmorphukr, AOT, StarLing, MCR DLL v2.0).

Кожна з розглянутих тем супроводжена контрольними питаннями, домашніми завданнями та відповідним планом лабораторної роботи. У додатках містяться завдання до модульної контрольної роботи, індивідуальної роботи, словник термінів з прикладної морфології, перелік питань до заліку з прикладної морфології, робоча програма з прикладної морфології.

ЗМІСТ

Змістовий модуль І

ТЕОРІЯ АВТОМАТИЧНОГО МОРФОЛОГІЧНОГО АНАЛІЗУ

Змістовий модуль ІІ

ОСНОВНІ МЕТОДИ АВТОМАТИЧНОГО МОРФОЛОГІЧНОГО АНАЛІЗУ

Змістовий модуль ІІІ

СУЧАСНІ АВТОМАТИЧНІ МОРФОЛОГІЧНІ АНАЛІЗАТОРИ

Змістовий модуль І

ТЕОРІЯ АВТОМАТИЧНОГО МОРФОЛОГІЧНОГО АНАЛІЗУ

ТЕМА 1. АВТОМАТИЧНИЙ МОРФОЛОГІЧНИЙ АНАЛІЗ СЕРЕД ДИСЦИПЛІН ПРИКЛАДНОЇ ЛІНГВІСТИКИ

Опорний конспект

1. Основне завдання прикладної (комп'ютерної) лінгвістики – автоматизоване опрацювання текстової інформації. Виокремлюють два види такого опрацювання: 1) формальне – перетворення фрагментів тексту без аналізу його змісту; 2) змістове (смислове) – перетворення фрагментів тексту з аналізом його змісту, встановленням логіко-семантичних відношень між його компонентами (використовується додаткова, семантична інформація, виражена в тексті імпліцитно). Для цього необхідно навчити комп'ютер аналізувати текст на різних рівнях представлення інформації – морфологічному, синтаксичному, семантичному, для чого створюються системи автоматичного перероблення тексту (АПТ) / автоматизовані системи опрацювання тексту (АСОТ) – один із різновидів лінгвістичних інтелектуальних комп'ютерних систем, що моделюють розумову діяльність людини у процесі розв'язання теоретичних і практичних завдань.

2. Виділяють дві технології опрацювання текстової інформації: словникову (створення допоміжних лінгвістичних баз даних словників, правил ідентифікації мовних одиниць) і безсловникову, або «незалежну» (представлення всіх необхідних відомостей про мовні одиниці у вигляді алгоритмічних правил). Обрання однієї із них залежить від типу мови тексту, характеру виконуваних завдань, можливостей комп'ютерної техніки (пам'яті, швидкодії), особливостей програмного забезпечення. Найбільш ефективними є системи, що вдало поєднують обидві технології, допоміжні бази даних із достатньо складними лінгвістичними алгоритмами.

3. Початкова частина автоматичного опрацювання текстів – автоматичний морфологічний аналіз (АМА). Його визначають як: аналіз слів, застосовуваний з метою їх членування на морфеми або сполучення морфем та отримання граматичної інформації, необхідної на наступних етапах опрацювання (Герольд Бєлоногов); визначення леми (базової, канонічної форми слова) та її граматичних характеристик (Віктор Бочаров); в обчислювальній лінгвістиці – аналіз окремих словоформ поза контекстом, у результаті якого послідовність словоформ вхідного тексту замінюється послідовністю інформації про ці словоформи (Володимир Волошин); процес, у результаті якого кожна словоформа тексту набуває свого коду частини мови і значення граматичних категорій (рід, число, відмінок, вид, час, особа тощо) та який забезпечує проведення морфемного, синтаксичного й семантичного аналізів, неможливих без попереднього визначення частин мови (Наталія Дарчук); вихідний модуль систем АПТ (АСОТ), у результаті здійснення якого комп'ютер для кожного слова в тексті визначає його граматичний клас, або частиномовну належність, та в межах граматичних класів – граматичний підклас, або граматичні підкласи, тобто розряди слів зі спільними змістовими, формальними та функціональними властивостями (здебільшого це слова, належні до різних граматичних категорій у межах окремих частин мови) (Євгенія Карпіловська); в автоматичній обробці тексту природною мовою за допомогою комп'ютера – процедуру, унаслідок якої з форми, зовнішнього оформлення слова в тексті можна одержати відомості про будь-які рівні мовної структури (Юрій Марчук).

4. Завданнями АМА є: визначення частиномовної приналежності текстових одиниць; ідентифікація словоформ однієї лексеми (за Наталією Дарчук); розщеплення словоформи (тобто слова в будь-якій з його граматичних форм) на морфеми; її лематизація (тобто перетворення її у вихідну форму: дієслова – в інфінітив; іменника, займенника – в форму Н.в. одн.; прикметника – у форму ч.р. Н.в. одн. тощо); визначення її граматичних характеристик (приналежність до певної частини мови, роду, числа, відмінка,

часу, особи, виду тощо) (за Зіновієм Партико). Автоматичне кодування слів тексту (приписування їм кодів граматичних класів) пов'язане з усвідомленням принципів традиційної граматичної класифікації. Необхідно виділити формальні ознаки слів, за допомогою яких можна було б автоматично визначати приналежність лексеми до певної частини мови. Формальні граматичні ознаки дозволяють провести первинну семантичну характеристику, адже іменник узагальнено – предмет, дієслово – дія, прикметник – ознака, прислівник – ознака дії тощо. Подальший АМА (визначення відмінку й числа для іменників, часу й особи для дієслова тощо) дозволяє визначити синтаксичну будову речення: підмет, присудок, додаток, обставину (суб'єкт і об'єкт дії, детермінанти). Тобто морфологічна інформація – основа для визначення смислу речення.

5. Будь-який різновид АМА ґрунтується на наукових узагальненнях і законах традиційної морфології. Морфологія – це розділ лінгвістики, який досліджує структуру слів та їх морфологічні характеристики. Комп'ютерна морфологія аналізує слова програмними засобами (Віктор Бочаров). Морфологія – це частина граматичної будови мови, що охоплює частини мови (граматичні класи слів), морфологічні (граматичні) категорії цих частин мови та їх форми. Серед основних завдань морфології як науки: 1) з'ясування принципів класифікаційного виділення частин мови; 2) виокремлення т.зв. морфологічної семантики слова; 3) опис формальних засобів, закріплених за певними частинами мови. Основною одиницею морфології є слово. Предметом вивчення морфології є граматичне значення – показник тих відношень, у які вступає слово в межах словосполучення, речення, тексту. Об'єктом вивчення морфології є структура слова, форми словозміни, способи вираження граматичних значень.

6. Ключовими поняттями як традиційної, так і прикладної морфології є: словоформа, морфологічне слово, граматичне значення, граматична категорія. Словоформи – це граматичні форми одного слова, тотожні лексично (спільне лексичне значення), але протиставлені граматичним значенням.

Морфологічне слово – це сукупність усіх словоформ (граматичних форм слів). Якщо така сукупність упорядкована, вона утворює парадигму. Граматичне значення виявляється в певній граматичній формі (одне й те ж граматичне значення може мати різні матеріальні засоби вираження, або форми: *писатим**у*** й ***буду** писати* – для майб. часу; *студент**у***, *студент**ові*** – для Д.в.). Граматична форма – це мовний знак, за допомогою якого виражається граматичне значення. Способи вираження граматичного значення: 1) синтетичний (граматичні значення в межах морфологічного слова виражаються за допомогою афіксів: закінчення, суфікс, префікс, інтерфікс тощо): *дядьк**ові***, *кошен**яти***, ***від**бігти*; 2) аналітичний (показник граматичних значень – службове слово): ***буду** мріяти*, *мріяв **би***, ***хай** мріє*; 3) аналітико-синтетичний (поєднання двох попередніх – афіксальне граматичне оформлення слова + аналітичні елементи): ***на** вікн**і**, **в** університет**і*** (грамема М.в.); 4) суплетивний (творення граматичних форм від різних коренів): ***т**и* – ***теб**е* – ***тоб**і*, ***поган**ий* – ***гір**ший*. Засоби вираження граматичного значення: флексія: *поле – пол**я***; суфікс: *ягня – ягн**ят**и*; префікс: *читати – **про**читати*; постфікс: *будувати – будуват**ися***; чергування: *сон – **сн**у*; наголос: *вода – в**о́**ди*; службові слова: ***хай** думає*; суплетивізм: *я – **мене***; порядок слів: *День змінює ніч. / Ніч змінює день*. Для позначення класів однотипних граматичних значень вживається термін «граматична категорія». До морфологічних граматичних категорій належать категорії роду, числа, відмінка, особи, часу, способу, стану, виду.

7. АМА у першу чергу зосереджується на дослідженні окремо взятого слова, поза контексту його уживання, зокрема аналізується структура слова, виокремлюються різні типи основ і граматичних класів. Так, слово є мінімальною формально виокремлюваною одиницею зв’язного письмового тексту, але воно – не мінімальна змістова одиниця і може складатися з однієї або кількох морфем. У складі слова розрізняють кореневі морфеми (корені), префікси (морфеми, що стоять перед коренем) та суфікси (морфеми, що стоять після кореня). Основне значеннєве навантаження несе корінь, а префікси й суфікси виступають у ролі модифікаторів змісту. У мовознавстві у словах (у

процесі функціонування мови, у різних контекстних оточеннях, вони можуть набувати різноманітних форм, межа між якими умовна), крім афіксів (префікси, інтерфікси, суфікси, флексії та постфікси), виділяють також словотвірну й словозмінну основи. Словотвірна основа – це така основа, з якої додаванням суфіксів і закінчень можна отримати правильні (тобто наявні у словнику) словоформи. Іноді словотвірна й словозмінна основи можуть збігатися (Зіновій Партико). Володимир Волошин зазначає, що морфологія слова – тільки те, що належить до його форми: закінчення, суфікси, флексії, корені й інші частини словоформи. У мовах з розвиненою морфологією аналіз окремо взятої словоформи надає можливість одержати велику кількість різноманітної інформації про різні граматичні категорії на основі аналізу їх синтаксичних функцій і систем відмінкових, особових і родових закінчень. Зміни форм слів можуть мати різний характер, бути пов'язаними як зі зміною основи слова (зміни складу префіксів і суфіксів, різноманітні чергування голосних і приголосних), так і зі зміною його закінчення. Словотвірні класи – класи слів, що характеризуються однаковим переліком суфіксів і сполучень суфіксів, поєднуваних з їх словотвірною основою. Словотвірна основа – початкова частина буквеного коду, що лишається після відсікання максимальної кількості суфіксів та відповідає умові продуктивності. Умова продуктивності – здатність виокремленої основи утворювати осмислені слова у сполученні з іншими суфіксами. Зміна граматичних закінчень – основний спосіб утворення різних форм слів зі зміною їх роду, числа, відмінка й особи. За характером зміни граматичних закінчень (флексій) і за своєю синтаксичною функцією слова можуть бути розбиті на ряд класів, які називаються флективними. Флективні класи відмінюваних слів виокремлюються на основі аналізу їх синтаксичної функції та систем відмінкових, особових та родових закінчень. Флективний клас слів характеризується набором ознак на основі слів-репрезентантів, що є носіями цих ознак. Ознаками, за якими змінюване слово може бути віднесене до певного класу, є: належність до однієї з синтаксичних груп або підгруп; система закінчень (тип словозміни). За

синтаксичною функцією змінювані слова об’єднані в такі групи: іменники; прикметники; дієслова в особовій формі; дієслова минулого часу, прикметники і дієприкметники; кількісні числівники. Класи невідмінюваних слів – тільки за синтаксичною функцією (Герольд Бєлоногов).

Контрольні питання

1. Що ви розумієте під автоматизованим опрацюванням текстової інформації? Наведіть приклади формального та змістового АОТ.
2. Яка із двох технологій опрацювання текстової інформації – словникова чи безсловникова – є, на вашу думку, більш ефективною і чому саме?
3. Дайте визначення автоматичному морфологічному аналізу. Чому його вважають початковою частиною автоматичного опрацювання текстів?
4. Назвіть основні завдання АМА.
5. Які формальні граматичні ознаки свідчать про приналежність лексеми до певної частини мови?
6. Дайте визначення морфології. Чим традиційна морфологія відрізняється від прикладної (комп’ютерної)?
7. Назвіть об’єкт, предмет і завдання традиційної та прикладної (комп’ютерної) морфології.
8. Дайте визначення словоформі. Наведіть приклади.
9. Дайте визначення морфологічному слову. Наведіть приклади.
10. Дайте визначення граматичному значенню. Чим лексичне значення відрізняється від граматичного? Наведіть приклади.
11. Назвіть основні способи вираження граматичного значення. Наведіть приклади.
12. Назвіть основні засоби вираження граматичного значення. Наведіть приклади.
13. Дайте визначення граматичній категорії. Наведіть приклади.
14. Дайте визначення лексемі. Основною одиницею вивчення яких розділів науки про мову є слово?
15. Дайте визначення морфології слова.

16. Чим відрізняються словотвірна і словозмінна основи? Наведіть приклади.
17. Дайте визначення словотвірних, флективних, невідмінюваних класів слів. Наведіть приклади.

Домашнє завдання

1. Як ви розумієте твердження Юрія Марчука: *«У комп'ютерній лінгвістиці поняття морфологічного аналізу є операційним. Якщо у традиційній лінгвістиці до морфологічного аналізу належить те, що характеризує форму і відповідає на питання «що» класифікують, то у прикладній лінгвістиці важливо не «що», а «як» отримують ту чи ту інформацію»*. Напишіть коротке висловлювання (5-7 речень).

2. Визначте частини мови в поданому реченні / тексті. Виконайте їх синтаксичний розбір.

А) *Студенти старанно виконують складне завдання*.

Б) *Морфологія як розділ описової граматики почала формуватися у надрах античної мовознавчої традиції. Було сформульовано основи традиційної класифікації частин мови і граматичних категорій, протиставлення «субстанції» (вихідної форми слова) і «акциденції» (парадигми слова), явищ аналогії. В епоху Відродження з'являється система понять, що стосуються структури слова (корінь, афікс, суфікс). Термін «морфологія» пов'язують з іменем Йоганна-Вольфганга фон Гете, який називав ним розділ біології про форми живих організмів. У 19 ст. цей термін поширюється і в мовознавстві на означення тієї галузі науки про мову, яка вивчає форми слова. У лінгвістичному значенні його вперше ввів Август Шлейхер*.

3. Охарактеризуйте формальні засоби вираження граматичних значень (флексія, суфікс, префікс, постфікс, чергування, наголос, службове слово, суплетивізм, порядок слів). З конкретними прикладами.

4. Наведіть повну парадигму для таких лексем: *парта*, *розумний*, *мій*, *дев'яносто п'ять*, *високо*, *учити*, *освічений*, *читаючи*, *у*, *і*, *хай*, *агов!* За якими

граматичними категоріями вони відмінюються? До яких граматичних класів слів (словотвірних / флективних / невідмінюваних) належать? Які способи і засоби вираження граматичних значень можна виокремити?

Лабораторна робота №1

АВТОМАТИЧНИЙ МОРФОЛОГІЧНИЙ АНАЛІЗ СЕРЕД ДИСЦИПЛІН ПРИКЛАДНОЇ ЛІНГВІСТИКИ

1. Прочитайте текст. Знайдіть у ньому іменники (виділіть червоним), прикметники (виділіть зеленим), прислівники (виділіть жовтим), займенники (виділіть фіолетовим), дієслова (виділіть синім).

Варіант 1

Граматичне значення – це таке абстраговане поняття, яке оформляє лексичне значення й виражає різні його відношення за допомогою граматичної форми. Граматичні значення, на відміну від лексичних, не співвідносяться безпосередньо з предметами, явищами й ознаками об'єктивної дійсності; вони відображають різні типи відношень між словами з лексичним значенням. Тобто лексичне значення пов'язане насамперед з номінацією, з називанням відчуттів, уявлень, понять, що, у свою чергу, відображають предмети й ознаки об'єктивного світу; граматичне значення характеризується вираженням відношень між словами – носіями лексичних значень для оформлення повідомлень про процеси і явища, що відбуваються в об'єктивній дійсності. Отже, усі так звані службові слова, що використовуються для вираження взаємозв'язку між повнозначними словами, є носіями граматичних значень. Повнозначні слова, крім лексичного значення, мають також граматичне, оскільки в процесі мовної діяльності вони так чи інакше пов'язані з іншими повнозначними словами. Тобто граматичне значення відповідно оформляє лексичне значення слова. Граматичне значення тісно пов'язане з лексичним; лексичне значення є субстратом, підосновою існування відповідного граматичного значення. Граматичне значення

посередньо, через лексичне, співвідноситься з об'єктивною дійсністю. Відмінність між граматичним і лексичним значенням виявляється також у плані їх вираження: граматичне значення виражається за допомогою відповідних граматичних засобів, що входять до складу слова (афікси); лексичне ж значення передається цілим словом.

Варіант 2

Одним із найважливіших і при цьому найскладніших центральних понять граматики (морфології і синтаксису) є граматична категорія. Термін «граматична категорія» у нашу граматичну науку ввів Олександр Опанасович Потебня. Граматична категорія – це найбільш узагальнене поняття, яке об'єднує однорідні граматичні значення, виражені різними мовними засобами. Так, однина і множина становлять категорію числа; недоконаний і доконаний вид – категорію виду; дійсний, умовний і наказовий способи – категорію способу. Граматичні категорії за своїм способом не однакові. Категорія числа охоплює багато частин мови, так само й роду, а категорії часу, способу властиві тільки дієсловам. Поняття граматичної категорії є родовим стосовно кожного ряду підпорядкованих їй взаємозалежних граматичних значень. Наприклад, граматична категорія відмінка – це таке родове загальне граматичне поняття, якому підпорядкована вся система відмінків як видових понять, виражених за допомогою системи відповідних граматичних форм. До складу морфологічних категорій сучасної української літературної мови належать такі: роду, числа, відмінка, ступенів порівняння, перехідності / неперехідності, особи, часу, способу, виду, стану. Морфологічні категорії характерні для змінних частин мови: іменника, прикметника, займенника, числівника, дієслова.

Варіант 3

Частини мови – лексико-граматичні класи слів, кожен із яких характеризується морфологічними, синтаксичними, лексико-семантичними особливостями. Принципи віднесення слів до різних частин мови досі не усталені, тому в різних граматичних розробках визначається неоднакова

кількість частин мови. Загальноприйнята класифікація, в основу якої покладено морфологічний принцип, що доповнюється синтаксичним і лексико-семантичним. Кожна частина мови має певний граматичний стрижень, який об'єднує усі належні до неї слова. Для іменників – це предметність, для прикметників – ознака, що супроводжує предметність, для дієслів – процесуальність та ін. Головною морфологічною ознакою, покладеною в основу лексико-граматичної класифікації слів, є здатність або нездатність їх до формотворення (словозміни). За цією ознакою виділяють змінювані і незмінювані слова. За співвіднесеністю з поняттям слова поділяються на повнозначні (співвідносні з поняттям) і неповнозначні (не співвідносні з поняттям). До змінюваних повнозначних слів належать іменник, прикметник, числівник, займенник, дієслово, до незмінюваних – прислівник. Неповнозначні слова не мають парадигми словозміни, граматично поєднуються з іншими словами в реченні, не співвідносяться з поняттями. Це клас службових слів, серед яких виділяють прийменники, сполучники, частки. Поза частинами мови перебуває вигук, який виступає як еквівалент слів і речень.

2. Згрупуйте знайдені словоформи в таблицю й проаналізуйте за такими параметрами (див. Табл. 1):

Таблиця 1. Параметри словоформ

Слово у формі, поданій у тексті	**Початкова форма слова**	**Лексичне значення слова**	**Граматичне значення слова**	**Засіб вираження граматичного значення**
значення	значення	суть чого-небудь; зміст	імен., с.р., одн., Н.в. (Р.в., З.в., К.в.) / мн.,	флексія ***-я***

			Н.в. (З.в., К.в.)	
граматичне	граматичний	прикметник до «граматика» (лінгвістична наука, що вивчає будову мови, тобто будову і форми слова, речення і словосполучення; будова і форми слова, речення і словосполучення певної мови або групи мов)	прикм., с.р., одн., Н.в. (З.в.)	флексія ***-е***
безпосередньо	безпосередньо	прислівник до «безпосередній» (який не має проміжних ланок, здійснюється без посередництва кого-, чого-небудь; прямий)	присл., сп. дії	суфікс ***-о***
таке	такий	який указує на загальну якість, властивість і т.ін. чого-небудь	займ., с.р., одн., Н.в. (З.в.)	флексія ***-е***
оформляє	оформляти	надавати чому-небудь певного	дієсл., недок.в.,	флексія ***-є***

		вигляду, певної форми; доводити що-небудь до потрібної форми	дійсн.сп., теп.ч., одн., 3 ос.	

Примітка

* Для визначення лексичного значення використовуйте ресурс http://sum.in.ua/.
* Для визначення граматичного значення використовуйте ресурс http://lcorp.ulif.org.ua/dictua/.

3. Зробіть висновок, який засіб вираження граматичних значень є для української морфології основним, визначальним. Як зазначені інтернет-ресурси допомагають автоматизувати морфологічне опрацювання текстової інформації?

Література до теми:

1. Баранов А. Введение в прикладную лингвистику. М., 2001. 360 с.
2. Белоногов Г. Компьютерная лингвистика и перспективные информационные технологии. М., 2004. 159 с.
3. Волошин В. Комп'ютерна лінгвістика. Суми, 2004. 382 с.
4. Гельбух А., Сидоров Г. К вопросу об автоматическом морфологическом анализе флективных языков. Ел. режим доступу: www.dialog-21.ru/Archive/2005.
5. Грязнухіна Т., Нікула М. Система автоматичного морфологічного аналізу українського наукового тексту. Проблеми українізації комп'ютерів. Матеріали 2-ї міжнародної конференції. Київ, 1993. С. 42-46.
6. Данилюк І. Прикладна морфологія. Донецьк, 2010. 216 с.
7. Дарчук Н. Комп'ютерна лінгвістика. К., 2008. 351 с.
8. Карпіловська Є. Вступ до прикладної лінгвістики: Комп'ютерна лінгвістика. Донецьк, 2006. 188 с.
9. Марчук Ю. Компьютерная лингвистика. М., 2007. 317 с.

10. Морфологический анализ научного текста на ЭВМ. К., 1989. 262 с.
11. Николаев И., Митренина О., Ландо Т. Прикладная и компьютерная лингвистика. М., 2016. 315 с.
12. Ножов И. Морфологическая и синтаксическая обработка текста (модели и программы). М., 2003. 140 с.
13. Партико З. Прикладна і комп'ютерна лінгвістика: Вступ до спеціальності. Львів, 2008. 224 с.
14. Романюк Ю. Прикладна морфологія. Черкаси, 2009. 117 с.
15. Соснина Е. Введение в прикладную лингвистику. Ульяновск, 2012. 110 с.
16. Сучасна українська літературна мова. Морфологія. К., 1969. 250 с.

ТЕМА 2. ЕКСПЕРИМЕНТАЛЬНІ ТА ПРОМИСЛОВІ СИСТЕМИ АВТОМАТИЧНОГО МОРФОЛОГІЧНОГО АНАЛІЗУ

Опорний конспект

1. Як зазначає Володимир Волошин, АМА у системі автоматичної обробки текстів визначається такими чинниками: типом природної мови (аналітична чи флективна); типом алгоритму автоматичної обробки тексту; основозмінними класами флективних слів. Вибір принципів АМА, на думку Наталії Дарчук, обумовлений кількома факторами: 1) системою мови (якщо найбільше розвинений синтетичний спосіб вираження граматичного значення (словозміна), то початковим етапом АМА є аналіз структури словоформи; якщо найбільше розвинений аналітико-синтетичний або аналітичний спосіб вираження граматичного значення (сполучення різних слів або словоформ), то аналіз слова являє собою пошук за словником заздалегідь визначених

морфологічних характеристик кожного слова чи словоформи (подальший аналіз відбувається за допомогою оточення (дистрибуції) слова)); 2) системою письма і друку (адже АМА призначений для писемного тексту, у якому має значення, буквене це чи складове письмо; як співвідносяться усне / писемне мовлення; спосіб членування тексту спеціальними засобами); 3) тематикою тексту як результату мовленнєвої діяльності й засобу комунікації.

2. Перші системи АМА (у 50-60-х рр.) спочатку створювалися як експериментальні та включали такі етапи: автоматичне виділення основи у словоформі тексту; пошук основи у словнику основ; порівняння структури словоформи з даними про її основу, які містяться у словнику основ. Тобто, кожна словоформа тексту аналізувалася за допомогою заздалегідь укладених словників основ, коренів, префіксів, суфіксів, флексій. Омонімія словоформ не розрізнялася. З часом і розвитком комп'ютерних технологій почали використовувати текстові закономірності уживання словоформи, поєднувати морфологічний аналіз із синтаксичним та семантичним.

3. В одній із перших систем АМА, у системі ECAIT (Експериментальна система автоматичного індексування текстів) здійснюється аналіз і синтез рефератів (з журналу «Вопросы информационной теории и практики»). Інструментами аналізу є: словники основ, закінчень, омонімічних основ; таблиці семантико-синтаксичної сполучуваності компонентів прийменникових конструкцій; зняття лексичної омонімії; семантичний аналіз іменних безприйменникових конструкцій; таблиці семантичної сполучуваності іменників і прикметників; алгоритми АМА, які визначають певну послідовність перевірок і звертань до словника і таблиць. Процедура автоматичного індексування розбита на послідовні блоки: морфологічний аналіз, синтаксичний аналіз, семантико-синтаксичний аналіз прийменникових конструкцій та варіювання смислового запису запиту.

4. У 70-80-ті рр. XX ст. почали створюватися промислові системи опрацювання текстової інформації: системи автоматичного перекладу,

системи інформаційного пошуку, системи автоматичного редагування текстів, системи автоматизованого анотування й реферування літератури.

5. Однією з розробок засобів індексування тексту документів є система SMART, описана Герардом Селтоном. У її основі: система поділу слів тексту на флексію і основу; словник еквівалентностей (тезаурус), призначений для заміни еквівалентних слів одним або кількома номерами понять, які слугують ідентифікаторами змісту замість основ слів; тезаурус у вигляді ієрархії понять, що забезпечує пошук для даного поняття загальнішого або вужчого чи асоційованого з ним поняття; словники статистичних і синтаксичних словосполучень; система обслуговування словників.

6. Систему автоматичного індексування РЕФЕРАТ утворюють такі етапи: виокремлення найбільш інформативних слів і словосполучень із тексту; розшифрування абревіатури; заміна слів, основи яких мають дескриптори у машинному словнику, на код дескриптора; зняття омонімії.

7. Метод індексування на основі семантичного аналізу розглянуто в роботі Ніни Леонтьєвої та Світлани Вишнякової. Його утворюють два етапи: індексування за дескрипторним словником (див. Табл. 2) у режимі «Слово», який супроводжується частковим морфологічним аналізом і лематизацією; індексування за допомогою автоматичного інформаційно-пошукового тезауруса.

Таблиця 2. Структура дескрипторного словника

1 колонка	2 колонка	3 колонка
основа слова	набір дескрипторів (дескриптор – елементарне поняття, кожне слово – набір дескрипторів)	граматичні ознаки дескрипторів

8. В Україні над створенням АМА російського тексту із середини 80-х рр. працював колектив співробітників відділу структурно-математичної лінгвістики Інституту мовознавства ім. О.О. Потебні НАНУ під керівництвом

Валентини Перебийніс (монографія «Морфологический анализ научного текста на ЭВМ»). З 90-х рр. працюють над АМА української мови, створивши систему орфографічного контролю «РУТА». Обидві системи АМА української та російської мов покладено в основу системи машинного перекладу «ПЛАЙ».

9. Сьогодні у розробці систем АМА можна виокремити кілька основних напрямів: 1-й моделює класичну схему аналізу шляхом поділу словоформи на основу та потенційну флексію з подальшою перевіркою на сумісність флексії та основи; 2-й використовує інформацію, що міститься в кінцевих буквосполученнях (після попередньої статистичної обробки словника); 3-й створює універсальні математичні моделі морфології у формі відкритих систем рівнянь, що дозволяють шляхом обчислення здійснювати нормалізацію словоформ, отримувати граматичну інформацію та синтезувати словоформи; 4-й в основі побудови алгоритмів морфологічного аналізу має поділ усіх слів на класи, які визначають характер зміни буквеного складу форм слів (Володимир Волошин). Також виокремлюють: підходи на основі правил (rule-based methods); статистичні методи (statistical methods), пов'язані в основному з машинним навчанням (machine learning); гібридні підходи (hybrid methods), які поєднують правила і статистику (Віктор Бочаров).

10. Зіновій Партико основними завданнями сучасних і майбутніх систем АМА визначає: 1) створення таблиць словозміни; 2) укладання списків морфем (префіксів, коренів, інтерфіксів, суфіксів, закінчень, постфіксів); 3) визначення продуктивності й частотності тих чи тих морфем; 4) визначення частот реалізації в текстах різних граматичних категорій (роду, відмінка, числа, часу, способу дії, виду тощо); 5) автоматичне визначення морфологічної будови слів у текстах; 6) автоматична лематизація словоформ із текстів; 7) автоматичне визначення морфологічних характеристик словоформ із текстів; 8) автоматичне ймовірнісне визначення для нових невідомих слів їх морфологічної будови, способу відмінювання й приписування їм граматичних характеристик.

Контрольні питання

1. Дайте визначення аналітичного та флективного типів природної мови. Наведіть приклади.
2. Як залежить майбутній АМА від системи мови, системи письма / друку, тематики тексту?
3. Коли виникли і які основні етапи включали перші, експериментальні, системи АМА?
4. Коли виникли і які основні етапи включали промислові системи АМА?
5. Які напрями, підходи й виконувані завдання можна виокремити в розробці найсучасніших систем АМА?

Домашнє завдання

1. Укладіть аналітико-порівняльну таблицю *«Принципи АМА для української / російської / англійської мов»*.

2. Сформулюйте, чим промислові системи АМА відрізняються від експериментальних? Опишіть кількома реченнями принцип роботи однієї із них (SMART, РЕФЕРАТ, метод індексування на основі семантичного аналізу, «ПЛАЙ» – на вибір).

Лабораторна робота №2
ЕКСПЕРИМЕНТАЛЬНІ ТА ПРОМИСЛОВІ СИСТЕМИ АВТОМАТИЧНОГО МОРФОЛОГІЧНОГО АНАЛІЗУ

1. Перейдіть за посиланням http://www.mova.info/corpus.aspx?l1=209. Натисніть на меню поруч з заголовком «КОРПУС УКРАЇНСЬКОЇ МОВИ». Ознайомтеся з розділом «Про корпус» / підрозділом «Що таке корпус» (див. Рис. 1-2).

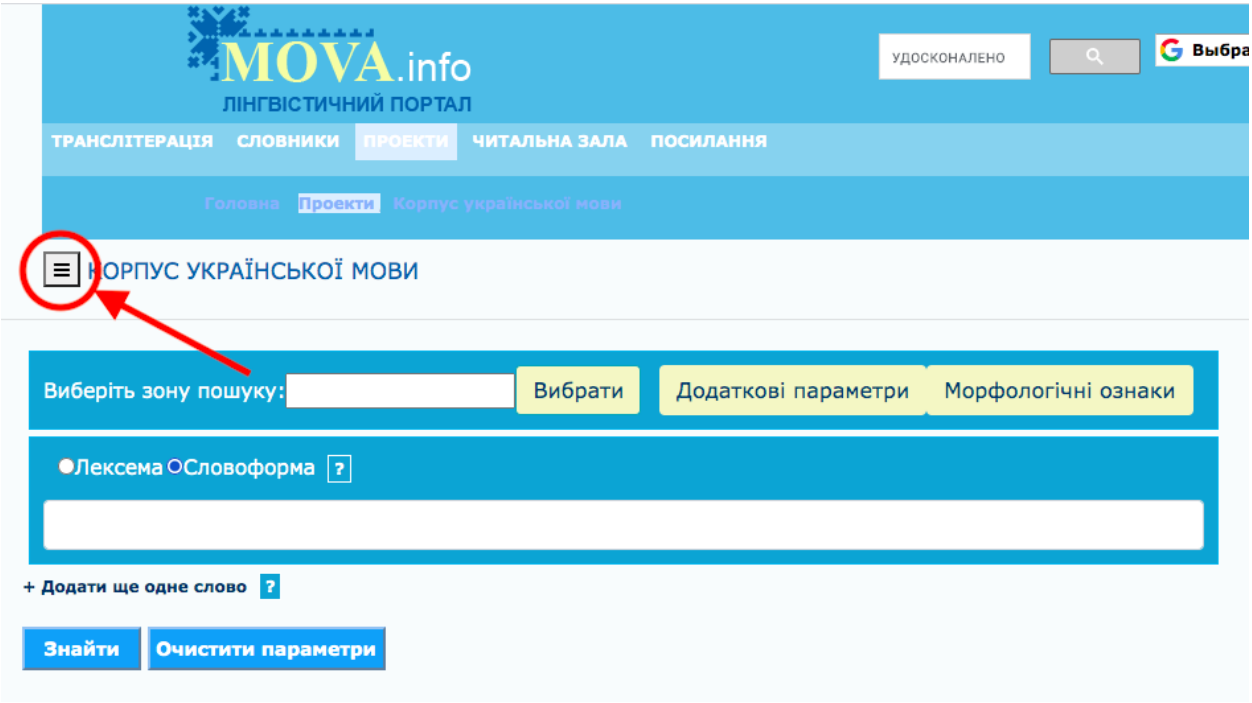


Рисунок 1. Інтерфейс Корпусу української мови

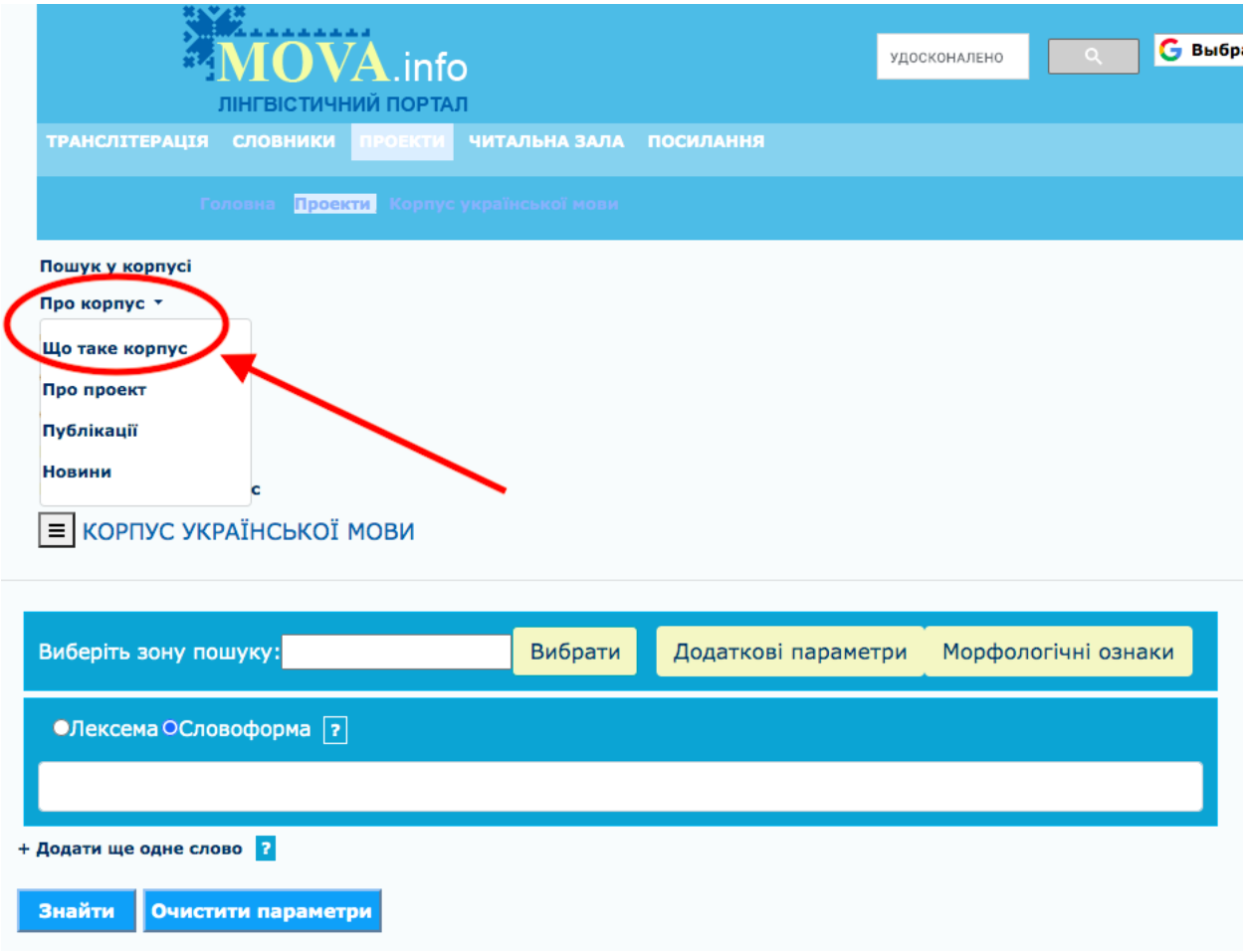


Рисунок 2. Інструкція Корпусу української мови

2. Ознайомтеся з програмними можливостями Корпусу – параметрами «Зона пошуку», «Додаткові параметри», «Морфологічні ознаки» (див. Рис. 3-4).

* Обов’язково натискайте на всі віконечка зі знаком питання [?], де подано додаткові роз’яснення до певних параметрів.

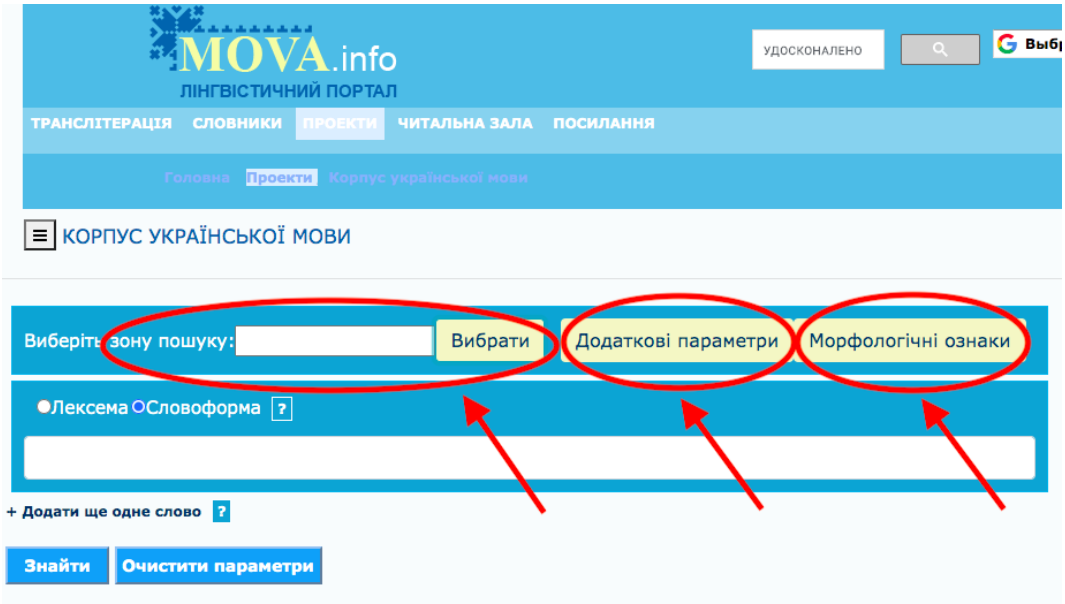


Рисунок 3. Параметри «Зона пошуку», «Додаткові параметри», «Морфологічні ознаки»

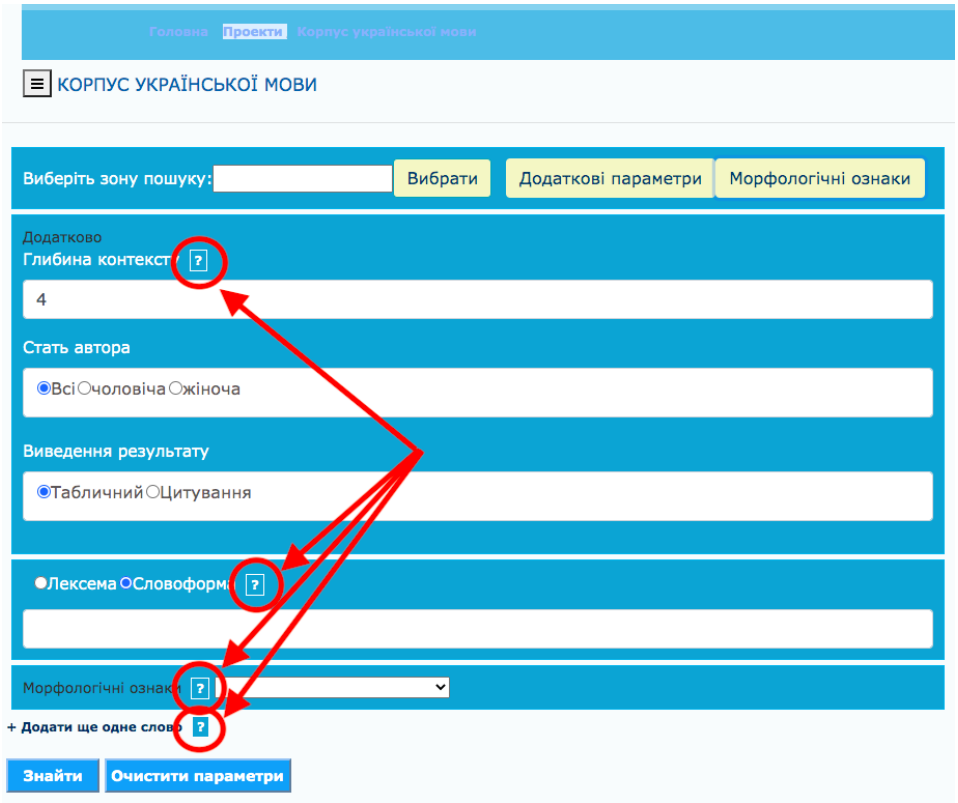


Рисунок 4. Додаткові роз’яснення щодо параметрів

3. Користуючись Корпусом української мови, здійсніть пошук таких:

– ***лексем і словоформ*** (див. Табл. 3)

Таблиця 3. Пошук лексем і словоформ

Зона пошуку	Лексема для пошуку	Словоформа для пошуку
ЗАКОНОДАВЧІ ТЕКСТИ	закон, країна, покарання	закону, країні, покаранням
НАУКОВІ ТЕКСТИ	термін, дисципліна, знання	терміном, дисципліну, знанню
ПОЕТИЧНА МОВА	герой, Україна, кохання	герої, Україною, коханні
ПУБЛІЦИСТИКА	журналіст, новина, обговорення	журналістові, новини, обговорень
ФОЛЬКЛОРНІ ТЕКСТИ	козак, доля, поле	козака, долею, полем
ХУДОЖНЯ ПРОЗА	чоловік, жінка, життя	чоловіки, жінці, життів

– ***словосполук (за допомогою функції «Додати ще одне слово»)*** (див. Табл. 4)

Таблиця 4. Пошук словосполук

Зона пошуку	Сполука лексем для пошуку	Сполука словоформ для пошуку
ЗАКОНОДАВЧІ ТЕКСТИ	юридичний + відповідальність	юридичної відповідальності
НАУКОВІ ТЕКСТИ	ключовий + поняття	ключові поняття
ПОЕТИЧНА МОВА	мій + любов	моєї любові
ПУБЛІЦИСТИКА	вибір + народ	вибором народу
ФОЛЬКЛОРНІ ТЕКСТИ	рідний + мати	рідну матір
ХУДОЖНЯ ПРОЗА	їх + розмова	їхньою розмовою

4. Користуючись Корпусом української мови, самостійно побудуйте три вибірки за різними морфологічними ознаками (використовуючи параметр «Морфологічні ознаки»).

5. Результат кожного виконаного завдання з вибірки (3-4-е завдання лабораторної роботи) фіксуйте у вигляді скріна екрана, іменуйте їх номерами 1, 2…, заархівуйте та надішліть викладачеві.

6. Зробіть висновок, чи можна вважати, що в Корпус української мови вбудована промислова система автоматичного морфологічного аналізу? Які граматичні характеристики про словоформи подає ця система?

Література до теми:

1. Баранов А. Введение в прикладную лингвистику. М., 2001. 360 с.
2. Белоногов Г. Компьютерная лингвистика и перспективные информационные технологии. М., 2004. 159 с.
3. Волошин В. Комп'ютерна лінгвістика. Суми, 2004. 382 с.
4. Гельбух А., Сидоров Г. К вопросу об автоматическом морфологическом анализе флективных языков. Ел. режим доступу: www.dialog-21.ru/Archive/2005.
5. Грязнухіна Т., Нікула М. Система автоматичного морфологічного аналізу українського наукового тексту. Проблеми українізації комп'ютерів. Матеріали 2-ї міжнародної конференції. Київ, 1993. С. 42-46.
6. Данилюк І. Прикладна морфологія. Донецьк, 2010. 216 с.
7. Дарчук Н. Комп'ютерна лінгвістика. К., 2008. 351 с.
8. Дарчук Н. Морфологічне анотування Корпусу української мови. *Комп'ютерна лінгвістика: сучасне і майбутнє. Матеріали Міжнародної науково-практичної конференції*. К., 2012. С. 16-19.
9. Карпіловська Є. Вступ до прикладної лінгвістики: Комп'ютерна лінгвістика. Донецьк, 2006. 188 с.
10. Марчук Ю. Компьютерная лингвистика. М., 2007. 317 с.

11. Міщенко Н. Система програм морфологічного аналізу науково-технічних текстів. *Наукові записки. Серія «Філологічні науки»*. Випуск 95(2). К., 2011. С. 538-542.
12. Морфологический анализ научного текста на ЭВМ. К., 1989. 262 с.
13. Николаев И., Митренина О., Ландо Т. Прикладная и компьютерная лингвистика. М., 2016. 315 с.
14. Ножов И. Морфологическая и синтаксическая обработка текста (модели и программы). М., 2003. 140 с.
15. Партико З. Прикладна і комп'ютерна лінгвістика: Вступ до спеціальності. Львів, 2008. 224 с.
16. Романюк Ю. Прикладна морфологія. Черкаси, 2009. 117 с.
17. Соснина Е. Введение в прикладную лингвистику. Ульяновск, 2012. 110 с.
18. Сучасна українська літературна мова. Морфологія. К., 1969. 250 с.
19. Antworth E. PC-KIMMO: a two-level processor for morphological analysis. *Occasional Publications in Academic Computing*. No. 16. Dallas, 1990. 273 p.

ТЕМА 3. ДОМОРФОЛОГІЧНИЙ АНАЛІЗ ЯК ПОЧАТКОВИЙ ЕТАП АВТОМАТИЧНОГО МОРФОЛОГІЧНОГО АНАЛІЗУ

Опорний конспект

1. Будь-якому різновиду АМА передує підготовчий – доморфологічний етап: аналізований текст необхідно розбити на речення, у кожному реченні виокремити слова, розділові знаки, інші елементи тексту (числа, формули, таблиці, смайлики тощо). Віктор Бочаров називає цей етап токенізацією

(tokenization), а виокремлені в результаті одиниці – токенами (tokens). На доморфологічному етапі вирішуються такі завдання: 1. Що таке слово? 2. Чи будуть пунктуаційні знаки або аналітична форма майбутнього часу словом? 3. Яке значення великої літери на початку речення, у складі власної назви, в абревіатурі? 4. Яке функціональне значення крапки: кінець речення, позиція після скорочення чи рубрикації тощо?

2. У письмових текстах є багато умовностей і елементів формалізації: пробіли – для виділення меж між словами, великі літери й розділові знаки – для виділення меж між реченнями й складовими частинами речень, абзацні відступи – для виділення меж між зв'язаними за смислом групами речень тощо.

3. Машинне слово – ланцюжок графем від пробілу до пробілу, у тому числі пунктуаційні знаки.

4. Крапка репрезентує кілька ситуацій: типові скорочені слова, задані у словнику; рубрикацію після цифри або букви; ініціали (*В.І.Іванов* або *Іванов В.І.*); кінець речення. Початок речення – слово після крапки і пробілу. У кінці назв текстів ставиться умовна крапка і фіксується умовний кінець речення (наприклад, у процесі автоматичного реферування). Дробові числа мають написання з комою і з крапкою. Усі слова, що мають у своєму складі одну або кілька великих букв у сполученні зі знаками дефіс, тире або скісна риска, розглядаються окремо, тобто аналізуються як окремі лексичні одиниці. Членуються й окремо аналізуються слова-композити (*семантико-синтаксичний, диван-ліжко*), слова з першою, переддефісною частиною у вигляді цифр або латинських букв (*n-розрядний, 10-го*). Якщо в тексті певною мовою помилково чи спеціально вживаються літери іншої мови, то їм присвоюється окремий код-знак (?).

5. Алгоритм доморфологічного аналізу може бути прийнятий за основу коректора тексту, який виявляє помилки вже на початковому етапі опрацювання.

Контрольні питання

1. Дайте відповіді на питання *1. Що таке слово? 2. Чи будуть пунктуаційні знаки або аналітична форма майбутнього часу словом? 3. Яке значення великої літери на початку речення, у складі власної назви, в абревіатурі? 4. Яке функціональне значення крапки: кінець речення, позиція після скорочення чи рубрикації тощо?*, керуючись принципами традиційної лінгвістики.
2. Дайте визначення машинному слову. Наведіть приклади.
3. Яке функційне навантаження традиційно виконують крапка, кома, крапка з комою, знак оклику, знак питання, тире, дефіс; пробіл, абзацний відступ.

Домашнє завдання

1. Наведіть 10 прикладів слів-композитів. Чи є їх частини окремими машинними словами?

2. Наведіть приклади 5 формул. Порахуйте в кожній кількість машинних слів.

3. Яка кількість машинних слів, на вашу думку, є оптимальною для заголовків наукових статей? Продемонструйте на 10 прикладах заголовків.

Лабораторна робота №3
ДОМОРФОЛОГІЧНИЙ АНАЛІЗ ЯК ПОЧАТКОВИЙ ЕТАП АВТОМАТИЧНОГО МОРФОЛОГІЧНОГО АНАЛІЗУ

1. Відредагуйте поданий текст, розділивши його на слова, речення, абзаци тощо. За необхідності виправте помилки.

Варіант 1

Іменник–змінюванаповнозначначастинамови,якамаєзначенняпредметності,щовираж аєтьсявграматичнихкатегоріяхроду,числаівідмінка.Уреченніименникинайчас тішевживаютьсяуроліпідмета,додатка,іменноїчастинискладеногоприсудка,

обставини.Однієюзнайістотнішихознак,якоюіменниквідрізняєтьсявідінших частинмови,єкатегоріяроду–
самостійна,синтаксичнонезалежнакатегорія.Належністьтогоабоіншогоіме нникадопевногородувизначаєтьсялексичнимзначенням,морфологічнимиознак ами(х-
ромосновитавідмінковимизакінченнямиврізнихтипахвідмін),синтаксичнимиз в'язкамизіншимисловамивреченні(узгодж.прикм,дієприкм,займ.тощо).Катег оріяродувім.–назвахістоттаім.–назвахпредметіввир-сяпо-
різному.Віменниківназвахістоткатегоріяродумаєсемантичнийхарактер(бат ько–мати,дід–
баба).Віменників,щоозначаютьнеживіпредмети,категоріяродуасемантична. Пор.компонентилекс-сем.груп,якідиференціюютьсязародами:назвимеблів– стілець(ч.р.),полиця(ж.р),ліжко(с.р.).Іменниковийрідвиз-
сязазакінченнямназ.в.одн.:ч.р.:–
нульовафлексія(будинок,обрій,ячмінь,свинець);–
о(соловейко,батько,Дмитро);–е(вовчище,комарище,кабанище);–а,–
я(староста,воєвода).Ж.р.:–а,–я(калина,вода,земля);–
нульовафлексія(велич,капость,мідь,мазь).С.р.:–о(вікон,дно,добро);–
е(горе,море,поле);–а(дівча,лоша,ведмежа);–
я(картання,орля).Двоїстаграматичнаприродаузгоджувальнихможливостейв иявляєтьсявіменниківспільногороду,якіозначаютьосібчоловічоїтажіночоїст аті.Здебільшогоцекласназв,якідаютьекспресивнухарактеристикуозначуваної особи(вереда,забіяка,блудяга,волоцюга).Іменникиподвійного(спільного)родуха рактеризуєяскраваекспресивність,їхуживаютьздебільшоговрозмовномумовл енні,зокремауфамільярномупросторіччі:Титаказанудаі Титакийзануда.Доних наближаютьсяіменникижіночогороду,яківживаютьщодочоловіків,таіменни кисереднього роду,яківживаютьщодоосібчоловічоїтажіночоїстаті.Ціімення киузгоджуютьсявреченнітількиграматично,тобтоіменникижіночогородуузг оджуютьсятількивжіночомуроді(Вінтакасвиня),аіменникисереднього роду– тількивсередньомуроді(Оленко!Тимоєкошенятко!;Андрійку!Тимоєсонечко!).

Варіант 2

Формичоловічогойжіночогородуєпаралельнимивпозначенніосібжіночоїстаті заїхпрофесією,посадою,званнямтаін.Колисьрізніпрофесії,посади,званнямогли матитількичоловіки,томуйназвиїхпередавалимовііменникичоловічогороду(академік,асистент,чумак,суддя).Цепозначилосяйнановихназвах(президент,ГоловаВерховноїРадиУкраїни).Формачоловічогородутутнехарактеризуєзаттю,томуїїзастосовуютьіщодожінок(академікСтеповаВ.І.).Відбагатьохтакихформєпаралельніформижіночогороду(лікарка,аспірантка),якіширокovикористовуютьурозмовному,художньомумовленні,вокремихжанровихрізновидах публіцистичногостилю,алевофіційно-діловомустиліназвипосад,професій,званьтаін.уживаютьздебільшоговчоловічомуроді(лікарВесельськаТ.К.,аспірантЛесечківськаН.Ю.).Багатоіменниківчоловічогороду–назвзапрофесією,званням,посадоютощо–немаютьпаралельнихформжіночогороду:адвокат,менеджер,муляр,міністр,нотаріус,прокурор,тренер,хірург.Немаєвідповідниківжіночогородувусіхскладенихназвпосад,званьтаін.:головнийбухгалтер,молодшийдержавнийінспектор, старшийвикладач,провіднийтехнолог,стажист-дослідник.Загаломіменникижіночогородунапозначенняпосад,професій,званьвиниклипереважноврозмовномумовленніійомувластиві.Підупливомросійськоїмовивиниклиіменникиізсуфіксами-ш(а),-их(а),якієстилістичнознижениминазвамижінок.Крімтого,завживанняцихіменниківвиникаєдвозначність,оскількивониможутьтакожназиватидружинутого,хтомаєпозначуванупрофесію(генеральша–жінка-генералчидружинагенерала?).Напозначеннягрупосібчоловічоїтажіночоїстаті,щомаютьспільнупрофесію,видзаняттятощо,уживаютьіменникчоловічогороду:Шановністуденти!Ане:Шановністуденткиістуденти!Останнюформувикористовуютьзпевноюстилістичноюметою(длярозрізненнястаті).Однакценестосуєтьсяофіційнихінапівофіційнихзвертаньнакшталт«Панове!Паніпанове!Шановнідобродії!Шановнідобродійкиідобродії!»,уякихєцілкомрівноцінниминобидваваріанти;ітількиврозмовномумовленніперевагуінколинадаютьдругим

варіантам.Єіменникижіночогороду,якінемаютьвідповідниківчоловічогороду, оскількипозначаютьтрадиційножіночіпрофесіїтавидизанять:домогосподарка,покоївка,праля.Іноділексичнізначенняспільнокореневихслів– назвосібчоловічоїтажіночоїстаті–незбігаються:друкар–друкарка,машиніст–машиністка.Якщоіменникчоловічогородувживаютьнапозначенняжінки,алеім'яособиприцьомуненазивають,тоузгодженеозначенняйприсудокмаютьформу чоловічогороду:Аспірантнаписавцікавупрацю.Сполученнянакшталтдоцент Стенишинавимагають,щобприсудокмавформужіночогороду.Означенняжутак омуразімаєформучоловічогороду:МійшефНаталяСтепанишинаприйшлавчасно.

Варіант 3

Удеякихіменниківпростежуємовагання(хитання)увживанніформроду:зал–зала,клавіш–клавіша,мандаринчик–мандаринка,абрикос–абрикоса,африкат–африката,вольєр–вольєра,кахель–кахля,лангуст–лангуста,мотузок–мотузка,перифраз–перифраза,спазм–спазма,чинар–чинара.Формародуможе:–пов'язуватисязрізнимзначеннямслів(кар'єр–кар'єра,друкар–друкарка);–бутинейтральноювзмістовомутастилістичномуплані(жираф–жирафа);–матистилістичнезначення(туфляітуфель–останнєзаст.;методіметода–останнєзаст.абонаук.;жилетіжилетка–останнєрозм.).Однокореневііменникиврізнихмовахможутьнезбігатисявроді(болем–болью,ступенем–степенью,рукописом–рукописью).Сплутуванняродутакихіменниківукраїнськіймовіспричинененасампередупливомросійськоїмови.Рідневідмінюванихабревіатурвизначаютьзародомстрижневогослова(іменникавН.в.):ЗОК(Західнеоперативнекомандування).БЮТ(БлокЮліїТимошенко),ЮНЕСКО(ОрганізаціяОб'єднанихНаційзпитань освіти,наукиікультури),ОБСЄ(ОрганізаціязбезпекиіспівробітництвавЄвропі),ЮНІСЕФ(ДитячийфондООН).Рідневідмінюванихіменниківіншомовногопоходженнявизначаютьтак.Ціименники–назвиістот–

маютьчоловічийрід,назвинеістот–
середнійрід.Колиневідмінюванііменникипозначаютьтількижінок(фрейлейн,міс,місіс,мадам,фрау,фрекен),товони,зрозуміло,належатьдожіночогороду.Якщовжитоіменник–назвутварини–
івтекстівказанонасамицю,тоцейіменникпоєднуютьякіменникжіночогороду(Кенгуругодуваласвоєдитя).Винятки:–
іменникицеце,івасі,авеню,кольрабі,салямі,бере(груша),бері-бері(хвороба),страдиварі,альма-матер,фейхоаналежатьдожіночогороду;–
словасироко,памперо,майстро,греготаін.назвивітру,сулугуні(сир),шимі(танець),кабукі(театр),кавасакі(бот),бефстроганов,євроєіменникамичоловічогороду;–іменникамичоловічогойсереднього родуєсловаекю,ескудо,па-де-де,па-де-труа,сиртакі,мачете,статус-кво,бренді;–
словаесперанто,афгані(грошоваодиниця)єіменникамижіночогойсереднього роду;–
невідмінюванібагатозначніслованабуваютьформитогочиіншогородузалежно відзначення:альпака–ч.іж.(тварина)іс.р.(шерсть);каберне–
ч.р.(сортвинограду)іс.р.(вино);контральто,сопрано–
с.р.(голос)іж.р.(співачка).Зародовоюназвоювизначаютьрідневідмінюванихіменників–географічнихтаумовнихназв:зеленеТбілісі(місто),ця«Нью-Йорктайме»(газета).

2. Порахуйте кількість машинних слів.

3. На конкретних прикладах із тексту проаналізуйте функції ужитих у ньому розділових знаків.

Література до теми:

1. Баранов А. Введение в прикладную лингвистику. М., 2001. 360 с.
2. Белоногов Г. Компьютерная лингвистика и перспективные информационные технологии. М., 2004. 159 с.
3. Волошин В. Комп’ютерна лінгвістика. Суми, 2004. 382 с.

4. Гельбух А., Сидоров Г. К вопросу об автоматическом морфологическом анализе флективных языков. Ел. режим доступу: www.dialog-21.ru/Archive/2005.
5. Грязнухіна Т., Нікула М. Система автоматичного морфологічного аналізу українського наукового тексту. Проблеми українізації комп'ютерів. Матеріали 2-ї міжнародної конференції. Київ, 1993. С. 42-46.
6. Данилюк І. Прикладна морфологія. Донецьк, 2010. 216 с.
7. Дарчук Н. Комп'ютерна лінгвістика. К., 2008. 351 с.
8. Дарчук Н. Морфологічне анотування Корпусу української мови. *Комп'ютерна лінгвістика: сучасне і майбутнє. Матеріали Міжнародної науково-практичної конференції*. К., 2012. С. 16-19.
9. Карпіловська Є. Вступ до прикладної лінгвістики: Комп'ютерна лінгвістика. Донецьк, 2006. 188 с.
10. Марчук Ю. Компьютерная лингвистика. М., 2007. 317 с.
11. Міщенко Н. Система програм морфологічного аналізу науково-технічних текстів. *Наукові записки. Серія «Філологічні науки»*. Випуск 95(2). К., 2011. С. 538-542.
12. Морфологический анализ научного текста на ЭВМ. К., 1989. 262 с.
13. Николаев И., Митренина О., Ландо Т. Прикладная и компьютерная лингвистика. М., 2016. 315 с.
14. Ножов И. Морфологическая и синтаксическая обработка текста (модели и программы). М., 2003. 140 с.
15. Партико З. Прикладна і комп'ютерна лінгвістика: Вступ до спеціальності. Львів, 2008. 224 с.
16. Романюк Ю. Прикладна морфологія. Черкаси, 2009. 117 с.
17. Соснина Е. Введение в прикладную лингвистику. Ульяновск, 2012. 110 с.
18. Сучасна українська літературна мова. Морфологія. К., 1969. 250 с.

19. Antworth E. PC-KIMMO: a two-level processor for morphological analysis. *Occasional Publications in Academic Computing*. No. 16. Dallas, 1990. 273 p.

ТЕМА 4. ЛЕМАТИЗАЦІЯ ТА СТЕМІНГ

Опорний конспект

1. Автоматичне ототожнення різних словозмінних форм одного слова передбачає їх зведення до вихідної, словникової форми / угруповання словоформ одного слова (нормалізація словникової форми, лематизація). Цей процес має важливе прикладне значення: для автоматичного індексування в розробці інформаційно-пошукової системи, для створення машинних словників різних типів і призначення, для автоматичного реферування.

2. Усі розробки у сфері лематизації можна розділити на дві групи залежно від покладеного в основу принципу: 1) аналіз лише графемної структури слова; 2) залучення словникової або контекстної інформації, вихід за межі слова. Найбільшого розповсюдження набули розробки із застосуванням різноманітних словників (словоформ, основ тощо), оскільки дозволяють відмовитися від алгоритмічного морфологічного аналізу. У словниках слів словоформи групуються під словниковою формою слова або супроводжуються таблицями чергувань. Актуальні такі таблиці для словників основ (для основи вказується номер чергування у таблиці і граматична форма, у якій воно відбувається, а також частина мови відповідної лексеми). У деяких системах (наприклад, Лідії Кнориної) здійснюється класифікація типів основ, для кожного з яких наводяться списки закінчень, установлюється сумісність типу основи й закінчення, для деяких слів у словнику наводяться кілька типів

основ. Є спроби лематизації морфологічних форм на основі словника флексій (робота Федора Рибакова), однак вони не були достатньо ефективними через велику омонімію флексій та неврахування варіантів основ, які виникають через морфонологічні зміни в них у процесі словозміни. Також є спроби лематизації морфологічних форм з урахуванням елементів синтаксичного аналізу: позиції слова у реченні, ролі службових слів. У роботі Анни Піотровської описується приведення до канонічної форми іменних слововживань на основі списку трибуквених кінцівок слів. Установлюється тверда і м'яка словозміна, основи на російську літеру **и**, для кожного з типів відмінювання наводяться списки кінцівок слів. Здійснення алгоритму ускладнює велика кількість винятків (приблизно 20% аналізованих словоформ).

3. В Україні алгоритм і програму лематизації російської мови розробила Валентина Перебийніс. У роботі зведення парадигм російської мови базується на аналізі кінцівок словоформ. Укладено два алгоритми: для тексту, у якому кожному слововживанню присвоюється код граматичного класу (частини мови) і граматичного підкласу (рід, число, відмінок, особа, час), і для нерозміченого тексту. Зведення парадигм включає в себе кілька часткових задач, які розв'язуються в такій послідовності: виявлення флексії і відокремлення її від основи слова; виявлення варіантів основи, якщо такі є; об'єднання всіх форм слова в одну групу (парадигму); виділення або реконструкція словникової форми слова. Оскільки кожний граматичний клас слів, який має словозміну, має і свої особливості формоутворення й об'єднання форм у парадигми, то в алгоритмі доцільно передбачити таку кількість блоків, яка відповідає кількості граматичних класів, що характеризуються словозміною. Відповідно до прийнятого переліком граматичних класів слів виділяються такі блоки: зведення парадигм іменників (ч., ж., с.р., pluralia tantum); зведення парадигм атрибутивних класів (прикметників, дієприкметників, порядкових числівників); зведення дієслівних парадигм; зведення парадигм предикативних класів (скорочених

форм прикметників і дієприкметників); зведення займенникових парадигм (займенників-іменників, займенників-прикметників); зведення парадигм кількісного числівника; зведення варіативних форм прийменників. Алгоритм працює як на однозначно встановлених, так і на омонімічних (диз'юнктивних) класах. В алгоритмі діють два види правил: 1) правила, які встановлюють відповідність між граматичним кодом словоформи і довжиною флексії: якщо перед словоформою є даний код, то флексія складається з певної кількості букв; 2) правила, що враховують умови таких типів: умови, відповідно до яких змінюється довжина флексії (імен. с.р. *решени-е*, *множеств-о*, *пол-е*, *числ-о*, *врем-я*, але *цел-ое*); умови визначення варіантів основи; множини зі вставленням **о** або **е** перед **к**, **л** або заміною попередніх **ь**, **й** на **е** (*ошибк-а – ошибок*, *стрелк-у – стрелок*, *точк-а – точек*, *ячейк-а – ячеек*, *кальк-а – калек*, *петл-я – петел-ь* тощо) – відповідно, в імен. ж.р. Р.в. є умова: якщо перед флексією **ь** або в кінці слова є **к** або **л**, то в усіх відмінках **о** або **е** перед **к**, **л**; **о**, **е** або випадне, або міняється на **ь** чи **й** (*оценок – оценк-а*, *строек – стройк-а*, *земел-ь – земл-я*, *калек – кальк-а*). Особливі правила встановлені для різних граматичних класів на **ся**, займенникових класів на **то**, **либо**, **нибудь**. Підготовчим етапом до роботи алгоритму зведення парадигм є укладання частотного словника словоформ, який істотно обмежує словник як лексичний, так і граматичний. Починається зведення парадигм із сортування всіх словоформ за класами, тобто за першою буквою коду, яка визначає вихід на правила певного блоку. Розглянутий алгоритм базується на великій попередній роботі з визначення класів і підкласів слів шляхом як аналізу їхньої графемної структури, так і контекстного аналізу.

4. Існує необхідність розробки алгоритму зведення парадигм у тексті з непроставленими кодами граматичних характеристик слів. Принцип лишається традиційним: аналіз графемної структури словоформ тексту, не спираючись на попередньо заданий словник основ або словоформ. Однак стратегія цього алгоритму інша: оскільки немає опори на код слова, аналіз словоформи починається відразу з її кінця. Для кожної графеми слова

розробляється послідовність правил (блок), за якими аналізуються словоформи з кінця до початку не більше, ніж на 4-5 графем, тобто не більше двох графем від флексії. Складність кожного блоку залежить від кінцевої графеми. Правила чотирьох видів: аналізовану кінцеву букву або ланцюжок букв вважати флексією і межу основи встановлювати перед нею (*тем-а*); кінцеву букву вважати кінцем основи, слово не членувати (*алгоритм*); кінцева буква або ланцюжок букв може бути і флексією, і кінцем основи, і сполученням флексії з кінцем основи, тому необхідно здійснювати подвійне членування, кожна з гіпотетичних основ зіставляється з уже наявною внаслідок аналізу, знайдена основа вважається правильною, друга відкидається (наприклад, із *сист-ем* і *систем-* правильна друга); перевірити кінцеву букву виділеної основи для виявлення можливих варіантів основи (*ошибк-а* і *ошибок*, *конец* і *конц-а*, *мож-ет* і *мог-ут*). Тобто, алгоритм членує словоформу на основу і флексію, об'єднує словоформи з однією і тією ж основою і підраховує частоту кожної словоформи. Описаний алгоритм є ефективним для аналізу будь-якого тексту незалежно від його стильової і тематичної спрямованості. Його недолік – громіздкість і неможливість зняти граматичну та лексико-граматичну омонімію (імен. *стал-и* та дієсл. *ста-ли*), рішення вважати нечленованими всі словоформи, які зустрілися лише раз.

5. Було розроблено інший підхід, що враховує аналіз і синтез, які відтворюють у формалізованому вигляді процес ідентифікації словозмінних характеристик кожної лексеми за її морфологічним статусом. Аналіз текстової словоформи передбачає лематизацію – виведення канонічної словникової форми, а синтез, навпаки, – розгортання парадигми кожного слова, що має словозміну. Парадигматичний клас – множина слів з однаковим характером відношень між основою вихідної форми й основами інших словоформ парадигми, а також з однаковим набором кінцевих афіксів. Різні парадигматичні класи – це різні моделі формотворення. При визначенні парадигматичного класу ключовими є дві диференційні ознаки: флективний набір і графемні зміни, які відбуваються у процесі словозміни. Кожному слову

зі словозміною приписується парадигматичний клас, для якого за машинними правилами здійснюється розгортання парадигми (зв'язується інформація двох таблиць – таблиці основ з номером парадигматичного класу і таблиці змінної частини словоформ, що відповідають номеру парадигматичного класу), приписується відповідна граматична інформація. Зворотна процедура (коли за словоформою виводиться вихідна форма) передбачає інші машинні правила: текстова словоформа, синтезована з двох таблиць, отримує за відповідним кодом вихідну форму. Позитивним є те, що автоматичний словник створюється раз і назавжди, його можна лише поповнювати новими словами.

6. Стемінг (*stemming*) – це процес скорочення слова до основи шляхом відкидання допоміжних частин, таких як закінчення чи суфікс. Результати стемінгу іноді дуже схожі на визначення кореня слова, але його алгоритми базуються на інших принципах. Тому слово після обробки алгоритмом стемінгу (стематизації) може відрізнятися від морфологічного кореня слова. Стемінг застосовується в лінгвістичній морфології та в інформаційному пошуку. Багато пошукових систем використовують стемінг для об'єднання слів, у яких збігаються форми після стематизації, вони вважають такі слова синонімами. Цей процес називають злиттям. Комп'ютерна програма, що реалізує алгоритм стемінгу, називається стемером. Перші алгоритми стемінгу були створені Джулі Бет Ловінс (1968), Мартіном Портером (1980). Саме алгоритм останнього визнано стандартним для англійської мови. Пізніше науковець побудував Snowball – спеціальне середовище для написання алгоритмів стемінгу англійської та ін. мов.

7. Існують такі варіанти алгоритмів стемінгу: 1) пошук за таблицею (у якій зібрані всі можливі варіанти слів та їх форми після стемінгу); 2) відсічення закінчень та суфіксів; 3) лематизація (перший крок – POS tagging – визначення частин мови у реченні, другий крок – застосування правил стемінгу відповідно до частини мови); 4) стохастичні алгоритми (базуються на наборі логічних правил та таблиць пошуку, як результат – на ймовірнісному визначенні основи слова); 5) гібридний підхід (комбінація

вищенаведених алгоритмів – наприклад, пошуку за таблицею та відсічення закінчень і суфіксів); 6) відсічення префіксів; 7) пошук відповідності (ґрунтується на базі знань, що містить у собі лише основи слів та дозволяє знайти для слова найвідповіднішу форму з бази знань).

8. Типовими помилками стемінгу є: надстемінг (*overstemming*) – коли під час стематизації два слова скорочуються до однієї основи, хоча це не мало б статися; недостемінг (*understemming*) – протилежна помилка, коли слова отримують різні основи, хоча б мали мати одну спільну.

Контрольні питання

1. Дайте визначення лематизації. Які принципи лежать у її основі?
2. Яке практичне й теоретичне значення розв'язання проблеми автоматичного зведення парадигми слова?
3. Які основні алгоритми лематизації були розроблені й ким саме?
4. Опишіть алгоритм лематизації у розміченому тексті.
5. Опишіть алгоритм лематизації у нерозміченому тексті.
6. Опишіть алгоритм лематизації, що враховує аналіз і синтез словоформ.
7. Дайте визначення парадигматичному класу. Наведіть приклади.
8. Дайте визначення стемінгу. Хто і коли розробив перші алгоритми стемінгу?
9. Які варіанти алгоритмів стемінгу існують?
10. Які типові помилки стемінгу можна виокремити?

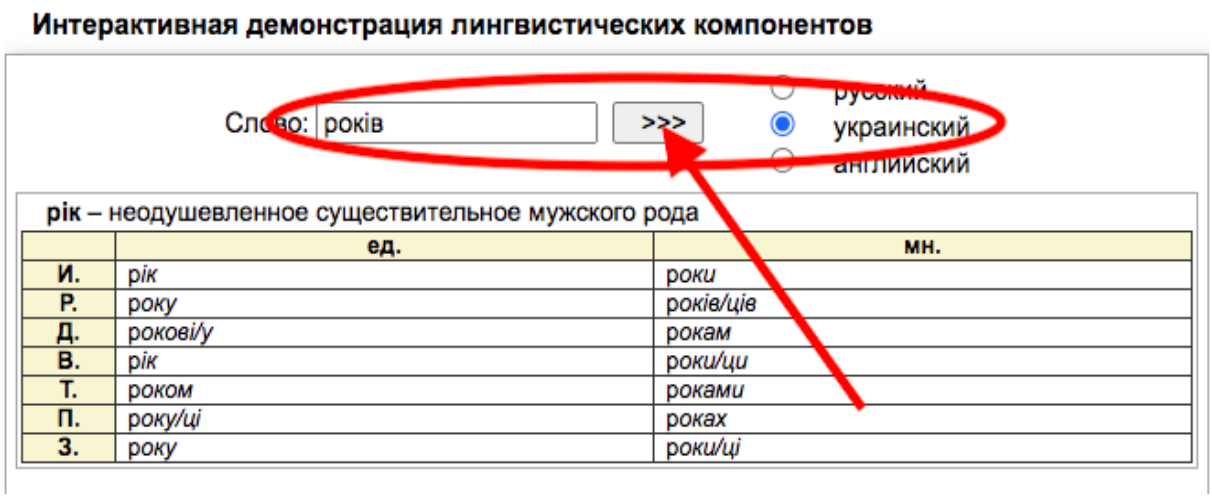
Домашнє завдання

1. Оберіть один із описаних алгоритмів лематизації і проілюструйте його роботу на конкретному прикладі.

2. Оберіть один із описаних алгоритмів стемінгу і проілюструйте його роботу на конкретному прикладі.

Лабораторна робота №4
ЛЕМАТИЗАЦІЯ ТА СТЕМІНГ

1. Користуючись ресурсом http://www.keva.ru/?cat=ling-morph-dem, здійсніть лематизацію будь-яких 10 іменників, 10 прикметників, 10 дієслів із запропонованого тексту (див. Рис. 5-7):



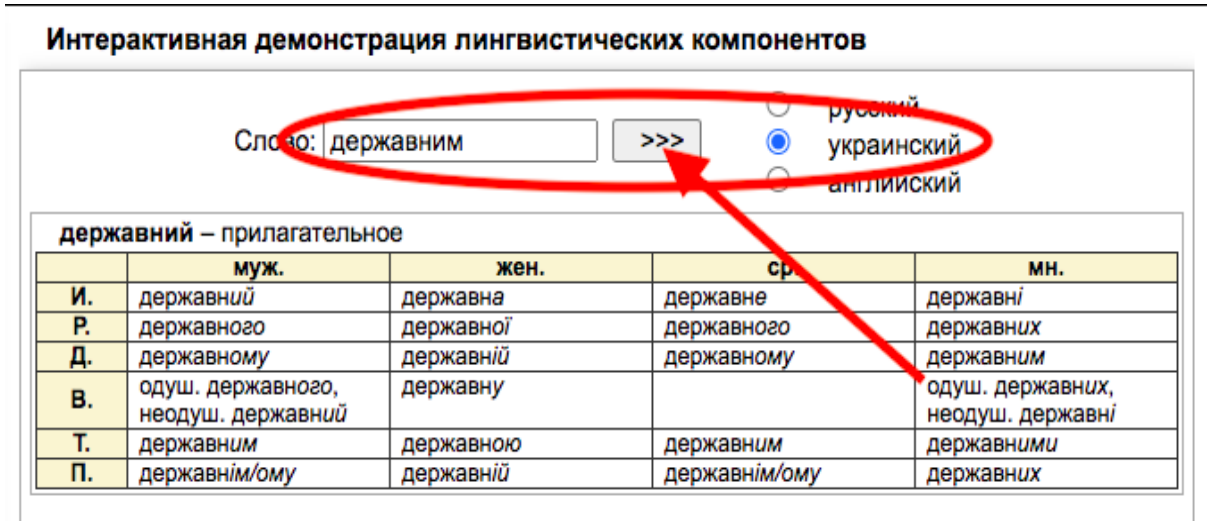
Интерактивная демонстрация лингвистических компонентов

Слово: років >>> русский / украинский / английский

рік – неодушевленное существительное мужского рода

	ед.	мн.
И.	*рік*	*роки*
Р.	*року*	*років/ців*
Д.	*рокові/у*	*рокам*
В.	*рік*	*роки/ци*
Т.	*роком*	*роками*
П.	*року/ці*	*роках*
З.	*року*	*роки/ці*

Рисунок 5. Пошук іменника



Интерактивная демонстрация лингвистических компонентов

Слово: державним >>> русский / украинский / английский

державний – прилагательное

	муж.	жен.	ср.	мн.
И.	*державний*	*державна*	*державне*	*державні*
Р.	*державного*	*державної*	*державного*	*державних*
Д.	*державному*	*державній*	*державному*	*державним*
В.	одуш. *державного*, неодуш. *державний*	*державну*		одуш. *державних*, неодуш. *державні*
Т.	*державним*	*державною*	*державним*	*державними*
П.	*державнім/ому*	*державній*	*державнім/ому*	*державних*

Рисунок 6. Пошук прикметника

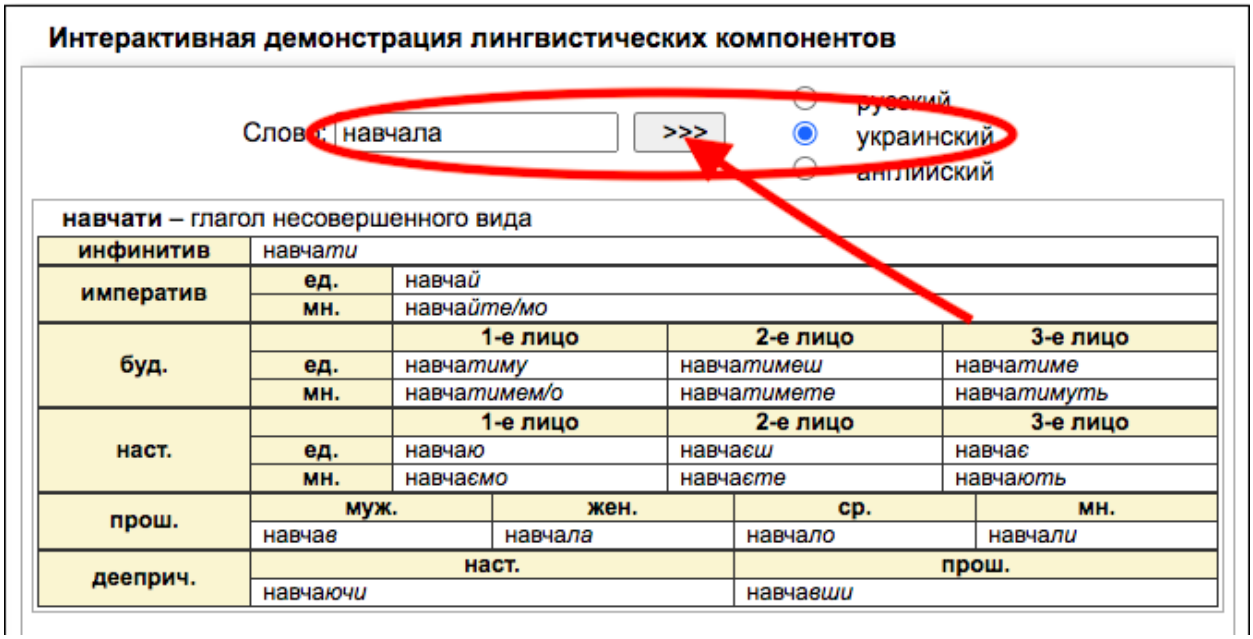


Рисунок 7. Пошук дієслова

Варіант 1

Упродовж років Україна державним коштом навчала десятки тисяч студентів, які в результаті жодного дня не працювали за спеціальністю. Українська академія лідерства виникла з бажання змінити цю ситуацію та створити альтернативу традиційній для України моделі навчання заради диплому. Випускникам УАЛ не видають дипломів-корочок. Натомість за 10 місяців навчання в академії молодь вчиться бути лідерами в будь-якій галузі – бізнесі, державотворенні чи громадській діяльності. За останні п'ять років в Україні з'явилися та активно розвиваються проєкти неформальної освіти, які значно доповнюють традиційну зашкарублу «освітянську ниву». З-поміж таких ініціатив особливо вирізняється Українська академія лідерства – платформа неформальної освіти для випускників шкіл та студентів віком від 16 до 20 років. Процес навчання в УАЛ триває 10 місяців, а програма передбачає розвиток у трьох напрямах: інтелектуальному, фізичному та емоційному. Академія не видає офіційного диплома, проте навчає молодь усвідомленому аналізу суспільних явищ, заохочує змінювати країну на краще та допомагає визначитися із вибором майбутнього фаху і розвинути лідерські якості.

Варіант 2

Своїми неформальними підходами до навчання УАЛ заперечує міцно вкорінені методи консервативної пострадянської системи освіти. Приміром, навчальний рік розпочинається тут не лінійками чи подібним офіціозом, а сходженням на Петрос – одну з найвищих гір українських Карпат. Для УАЛ це є символом спільно підкореної вершини, адже йдуть до неї усі разом. Загалом УАЛ акцентується на тому, що довгострокові трансформаційні зміни в суспільстві починаються з молоді, а девізом академії є «Творимо себе – творимо Україну! Кожне завтра розпочинається вже сьогодні». Керівник і засновник Української академії лідерства Роман Тичківський пояснює, що їхнім пріоритетом є розвиток особистості кожного студента. Команда менторів та студентська спільнота поділяють цінність бути справжнім; студентів заохочують приділяти увагу особистісному розвитку та розуміють, що кожен майбутній лідер потребує особливого підходу: Ми інвестуєм дуже багато в індивідуальну роботу, хочемо почути молоду людину: які прагнення, чого хоче, і не просто чого хоче – чи має здатність і бажання розвивати здібності до того, щоб досягати цих цілей.

Варіант 3

Для підтримки УАЛ команда академії об'єднує підприємців, державу, муніципалітети, інституційних донорів та українську діаспору. УАЛ активно підтримують декотрі із муніципалітетів, де діють осередки академії, зокрема в Миколаєві, Маріуполі, Харкові та Львові. Є донори у академії також і серед українських підприємців. Наприклад, родина Віктора та Ірини Іванчиків, засновників благодійного фонду «Повір у себе», свого часу ініціювала створення осередку на Полтавщині, а буковинський підприємець Роман Клічук – у Чернівцях. Аби стати студентом, потрібно пройти три відбіркових етапи: спочатку слід заповнити анкету, потім абітурієнта запрошують на регіональний відбір до найближчого осередку академії, і насамкінець відбувається фінальний національний відбір. Під час відбору зважають на особисту стійкість, логіку, витривалість, критичне мислення,

вміння працювати в команді. Навчальний тиждень в УАЛ триває шість днів замість звичних п'яти. Студенти проживають кожен цей день за детально розписаним розкладом, у якому поєднано всі три види розвитку. За особистісним поступом кожного студента стежить ментор, тобто досвідчений наставник, який супроводжує свого підопічного протягом усього навчання в академії. Ментори допомагають визначити цілі, проаналізувати вчинені дії, підтримують порадами та рефлексують разом із студентами.

2. Проаналізувавши отримані парадигми, визначте для кожного слова словотвірну основу. Випишіть її.

3. Здійсніть стемінг отриманого тексту, користуючись ресурсом https://gsgen.ru/tools/dlina-seo-text/.

4. Випишіть основи, відмінні від тих, які можна виділити шляхом традиційного граматичного виділення.

Література до теми:

1. Баранов А. Введение в прикладную лингвистику. М., 2001. 360 с.
2. Белоногов Г. Компьютерная лингвистика и перспективные информационные технологии. М., 2004. 159 с.
3. Волошин В. Комп'ютерна лінгвістика. Суми, 2004. 382 с.
4. Гельбух А., Сидоров Г. К вопросу об автоматическом морфологическом анализе флективных языков. Ел. режим доступу: www.dialog-21.ru/Archive/2005.
5. Грязнухіна Т., Нікула М. Система автоматичного морфологічного аналізу українського наукового тексту. Проблеми українізації комп'ютерів. Матеріали 2-ї міжнародної конференції. Київ, 1993. С. 42-46.
6. Данилюк І. Прикладна морфологія. Донецьк, 2010. 216 с.
7. Дарчук Н. Комп'ютерна лінгвістика. К., 2008. 351 с.
8. Дарчук Н. Морфологічне анотування Корпусу української мови. *Комп'ютерна лінгвістика: сучасне і майбутнє. Матеріали Міжнародної науково-практичної конференції*. К., 2012. С. 16-19.

9. Карпіловська Є. Вступ до прикладної лінгвістики: Комп'ютерна лінгвістика. Донецьк, 2006. 188 с.
10. Марчук Ю. Компьютерная лингвистика. М., 2007. 317 с.
11. Міщенко Н. Система програм морфологічного аналізу науково-технічних текстів. *Наукові записки. Серія «Філологічні науки»*. Випуск 95(2). К., 2011. С. 538-542.
12. Морфологический анализ научного текста на ЭВМ. К., 1989. 262 с.
13. Николаев И., Митренина О., Ландо Т. Прикладная и компьютерная лингвистика. М., 2016. 315 с.
14. Ножов И. Морфологическая и синтаксическая обработка текста (модели и программы). М., 2003. 140 с.
15. Партико З. Прикладна і комп'ютерна лінгвістика: Вступ до спеціальності. Львів, 2008. 224 с.
16. Романюк Ю. Прикладна морфологія. Черкаси, 2009. 117 с.
17. Соснина Е. Введение в прикладную лингвистику. Ульяновск, 2012. 110 с.
18. Сучасна українська літературна мова. Морфологія. К., 1969. 250 с.

ТЕМА 5. ПРОБЛЕМА ГРАМАТИЧНОЇ ОМОНІМІЇ У ПРОЦЕСІ АВТОМАТИЧНОГО МОРФОЛОГІЧНОГО АНАЛІЗУ

Опорний конспект

1. Із появою систем автоматичного опрацювання тексту, великих текстових корпусів та необхідності аналізу тексту на різних мовних рівнях, зокрема морфологічному, функційний аспект розгляду граматичної омонімії набув особливої актуальності. Це зумовлено тим, що змінне слово у тексті

репрезентовано тільки у вигляді певної словоформи, тобто під час граматичного аналізу тексту ми фактично маємо справу із граматичними формами слів у контексті (слововживаннями), а не вихідними (лексикографічними, канонічними) формами слів. Вирішення проблеми неоднозначності лексико-граматичного трактування реалізованого в контексті слова є одним із найактуальніших завдань прикладної лінгвістики.

2. Граматичні омоніми – формально тотожні граматичні форми або конструкції, що мають різне граматичне значення. Під час аналізу тексту засобами опрацювання природної мови кожній текстовій одиниці присвоюють набір граматичних характеристик. Морфологічний аналіз передбачає визначення щонайменше частини мови, а також значень відповідних граматичних категорій (грамем) і базової словоформи (леми). Якщо у процесі АМА не вдається однозначно проінтерпретувати словоформу, пропонується низка варіантів можливих лем і граматичних характеристик.

3. Внутрішньочастиномовні омоніми – омонімічні слововживання з однаковою частиномовною належністю. Внутрішньопарадигматичні омоніми – омонімічні слововживання в межах певної граматичної категорії (наприклад, відмінка). Внутрішньочастиномовні міжпарадигматичні омоніми – омонімічні слововживання, у яких лексико-граматичне значення збігається, але відрізняються леми.

4. Оскільки система морфологічного аналізу розглядає кожне слово ізольовано, вирішити граматичну неоднозначність на цьому рівні неможливо, треба звернутися до контексту. Це завдання можна вирішити вручну, однак наразі, через величезний обсяг сучасних корпусів текстів, воно потребує автоматизації. Частково завдання морфологічного уоднозначнення для української мови вирішував Олег Бугаков під час дослідження функціонування прийменників у тексті. Було створено алгоритм встановлення текстових умов зняття функціональної омонімії, коли одним із компонентів омоніма є прийменник; за допомогою дистрибутивного методу знято граматичну омонімію прийменників з іншими граматичними класами. Ольга

Шипнівська в межах свого дослідження сформувала лексикографічні бази даних міжчастиномовних омонімів та лінгвістичну базу даних для дослідження контекстів, де актуалізуються ті чи ті значення омонімічних одиниць; це стало основою для розробки правил автоматичного усунення міжчастиномовної морфологічної омонімії в українській мові.

5. Відповідно до способу отримання контекстної інформації, на основі якої відбувається розрізнення слів, виокремлюють такі основні підходи до граматичного уоднозначнення у процесі АМА: ймовірнісні (*stochastic, probabilistic*; також: статистичні, на основі машинного навчання); контрольованого навчання (*supervised*; супроводжуваного людиною); машинного самонавчання (*unsupervised*; без супроводу людини); на основі правил, розроблених вручну (*rule-based*).

6. Статистичні методи дають змогу обрати найвірогіднішу граматичну інтерпретацію словоформи в контексті на основі статистичних даних. Система машинного навчання аналізує певний текст, тренується на ньому, розпізнає деякі закономірності і на їх основі може робити певні узагальнення. Згідно з набутою інформацію про закономірні властивості словоформ у всіх контекстах, які система проаналізувала, вона може робити прогнози щодо найвірогіднішої граматичної інтерпретації словоформи у нових текстах. Методи контрольованого машинного навчання роблять ймовірнісний прогноз після тренування на корпусі текстів повністю або частково розміченого інформацією про словоформи, а методи машинного самонавчання дозволяють працювати з нерозміченим корпусом. Системи морфологічного аналізу на основі цього методу створено для англійської мови (Brill tagger, RDRPOSTagger) та адаптовано до деяких флективних мов. Ці аналізатори використовують методику трансформаційних правил, виведених в результаті машинного навчання. Загалом точність уоднозначнення в корпусах текстів англійської мови ймовірнісними аналізаторами перевищує 97%. Методи на основі правил полягають у використанні вручну розроблених та формалізовано представлених правил, обмежувальної граматики, у якій кожне

правило на основі контекстного оточення неоднозначної словоформи дозволяє або унеможливлює присвоєння їй певної грамеми. Правила можуть застосовуватися циклічно багато разів до максимального зменшення кількості варіантів граматичної інтерпретації словоформи. Перший великий аналізатор на основі контекстних правил TAGGIT містив 3300 правил і досягав уоднозначнення у Браунівському корпусі точністю 77%. Щоби скористатися цим методом, потрібна інформація щодо типів і моделей морфологічних омонімів та особливостей їх функціонування в тексті. Варто зважати на те, що: цей метод вимагає компетенції розробника; можливе замкнене коло, коли для морфологічного уоднозначнення потрібна інформація синтаксична чи лексична, а водночас для синтаксичного чи лексичного уоднозначнення потрібна морфологічна; характер проблеми й обсяг завдання залежить також від ефективності роботи самого граматичного аналізатора; якщо існують кілька аналізаторів для мови, то варто спробувати їх спільне використання. Методи на основі правил доцільно поєднувати з ймовірнісними: спробувати отримати правила уоднозначнення за допомогою статистичних методів і розробити правила для випадків, яких не було враховано за допомогою статистичних методів.

Контрольні питання

1. У чому полягає функційний аспект розгляду граматичної омонімії?
2. Дайте визначення граматичній омонімії. Наведіть приклади.
3. Дайте визначення внутрішньочастиномовним омонімам. Наведіть приклади.
4. Дайте визначення внутрішньопарадигматичним омонімам. Наведіть приклади.
5. Дайте визначення внутрішньочастиномовним міжпарадигматичним омонімам. Наведіть приклади.
6. Хто з науковців займався проблемою морфологічного уоднозначнення на матеріалі українськомовних текстів?

7. Опишіть основні підходи до граматичного уоднозначнення у процесі АМА.

Домашнє завдання

1. Прочитайте текст. Знайдіть і виділіть у ньому граматичні (морфологічні) омоніми.

Іменники мають два числа: однину й множину. Іменники в однині можна співвіднести із займенниками він, вона, воно або поєднати із цей, ця, це: цей світ, ця громада, це листя, це коріння. Іменники у множині можна співвіднести із займенником вони або поєднати із ці: ці дерева, ці корені, ці ножиці, ці Карпати. Однина й множина іменників звичайно різняться між собою закінченнями: будинок – будинки, вікно – вікна, дорога – дороги. Проте в іменниках середнього роду на -я (змагання, сузір'я) закінчення в обох числах збігаються. Число цих іменників визначаємо, орієнтуючись на слова, що стоять при них: змагання триває – змагання тривають, далеке сузір'я – далекі сузір'я. Частина іменників вживається, як правило, лише в однині. Вони означають: а) назви речовин: залізо, віск, кров, молоко, чорнило; б) збірні назви: студентство, молодь, дітвора, рідня, морква, бурячиння, листя, проміння, каміння, волосся; в) назви дій, якостей, почуттів: молотьба, хода, бджільництво, байдужість, поспішність, гнів, дружба; г) власні назви: Ольга, Степан, Іванченко, Луцьк, Куренівка.

2. Встановіть різновид виокремлених омонімів: внутрішньочастиномовний / внутрішньопарадигматичний / внутрішньочастиномовний міжпарадигматичний.

Лабораторна робота №5
ПРОБЛЕМА ГРАМАТИЧНОЇ ОМОНІМІЇ У ПРОЦЕСІ АВТОМАТИЧНОГО МОРФОЛОГІЧНОГО АНАЛІЗУ

1. Проаналізуйте отриманий текст. Знайдіть у ньому іменники (виділіть червоним), займенники (виділіть фіолетовим), дієслова (виділіть синім).

Варіант 1

Сучасну академію в Острозі можна вважати одним із найновіших вишів в Україні. У той же час академія, заснована тут ще у XVI ст., стала першим вишем Східної Європи. Тоді вона називалася Острозькою слов'яно-греко-латинською академією. Передумовою для культурного та освітнього життя міста стало князювання Василя-Костянтина Острозького. Князь переносить князівську резиденцію із Дубна до Острога та у 1576 році засновує тут навчальний заклад й починає будівництво приміщення для академії. За рік до того в Острозі князь засновує друкарню, куди запрошує найкращого книгодрукаря того часу – Івана Федоровича. Себе ж Василь-Костянтин оточує спільнотою кращих учених, теологів, публіцистів, іконописців того часу, яким дає доступ до найкращої із можливих на той час бібліотеки словників, граматик, грецької та європейської богословської літератури, передруків античних творів тощо. Саме це створює прецедент, незнаний доти: поєднується візантійська та західноєвропейська культури. Це стало можливим завдяки запозиченню із Західної Європи системи вивчення семи базових наук: граматики, риторики, діалектики, арифметики, геометрії, музики та астрономії. А також завдяки першій на цих територіях можливості вивчати вищі науки: філософію, богослів'я та медицину й опановувати 5 мов: слов'янську, польську, давньоєврейську, грецьку, латинську.

Варіант 2

Василь-Костянтин Острозький як полководець виграв 86 битв, більшість із яких були з татарами, кожна з цих битв не допустила їхнього входу до Центральної Європи, тому такі перемоги історично важко переоцінити стосовно збереження культури Заходу. Князь Острозький, якого ще називали «некоронованим королем Русі», магнат Князівства Литовського та сенатор Речі Посполитої, що має право ставити свою печатку на червоному воску (що робили тільки королі та особливої заслуги магнати), робить у Острозі те, що пізніше називатимуть ренесансом українського

народу. Він дбає про розвиток православ'я, але заради освіченості користується творами католицьких богословів, створюючи умови для якісного розвитку громади. Що ж до Острозької школи, яка стає пізніше академією, ресурсна база першої науково-освітньої установи була в пріоритеті князя, а тому доходи з довколишніх сіл автоматично йдуть на потреби закладу. Одним із перших вагомих вкладів у заклад був внесок племінниці князя – Гальшки. Острог набирає обертів із такою потужною базою. Тут у 1578 році видали першу в Україні «Грецько-руську церковнослов'янську читанку». Пізніше тут видаються Буквар, Новий Заповіт, а до нього й перший друкований в Україні алфавітно-предметний покажчик. Взагалі кількість українських першодруків, які побачили світ саме в Острозі вражає, втім успіхом та основним досягненням того часу недарма вважають саме Острозьку Біблію. Надрукована 1581 року, вона стала першим повним православним канонічним виданням усіх 76 книг Старого та Нового Заповітів церковнослов'янською мовою.

Варіант 3

Нинішній ректор Острозької академії Ігор Пасічник може розповідати про відродження академії багато та детально, адже саме він повністю співпереживав кожен етап становлення тут навчального закладу: «Я приїхав – це була руїна в повному розумінні цього слова. Тут було нічого: жодного стола, жодного стільця, жодної книжки, жодного приміщення і, я вже не говорю про кандидата наук». Тоді, на початку 90-х, він був єдиним доктором наук на Західній Україні, планував перебиратися жити до більшого міста. Втім, зустріч із тодішнім віце-прем'єром Миколою Жулинським та ректором Києво-Могилянської академії В'ячеславом Брюховецьким змінила і плани Пасічника, і подальші погляди на життя. Вони розповіли про Острог та історію академії, про князів та книговидавничу справу. Тоді Ігор Пасічник нічого не знав про цю історію, адже за період Радянського Союзу була ампутована найменша згадка й пам'ять, і про князів Острозьких, і про Острозьку академію, а Острог був відомим як районний центр із великою

психіатричною лікарнею: «Це була найбільш фантастична ідея, яка тільки могла прийти комусь у голову: відродити Острозьку академію в богом забутому містечку, не дивлячись на те, що це містечко називалось колись стольний град Острог. І я природньо почав тікати звідсіля. І більшість моїх друзів вважали, тут є психіатрична лікарня така, знаєте, потужна в Острозі. То мені говорили, що, мабуть, я скоро потраплю туди, бо погодитись на Острог і на ніщо – це було смішно на той час».

2. Знайдіть серед виділених іменників, займенників, дієслів граматичні омоніми. Зазначте, які саме морфологічні характеристики виступають показниками омонімічності.

3. Запропонуйте свій спосіб уоднозначнити граматичне значення тієї чи тієї словоформи.

Література до теми:

1. Баранов А. Введение в прикладную лингвистику. М., 2001. 360 с.
2. Белоногов Г. Компьютерная лингвистика и перспективные информационные технологии. М., 2004. 159 с.
3. Буньо Г. Сучасні методи вирішення проблеми граматичної омонімії в тексті. *Наукові записки. Серія: Філологічна*. 2014. Вип. 49. С. 12-16.
4. Волошин В. Комп'ютерна лінгвістика. Суми, 2004. 382 с.
5. Гельбух А., Сидоров Г. К вопросу об автоматическом морфологическом анализе флективных языков. Ел. режим доступу: www.dialog-21.ru/Archive/2005.
6. Грязнухіна Т., Нікула М. Система автоматичного морфологічного аналізу українського наукового тексту. Проблеми українізації комп'ютерів. Матеріали 2-ї міжнародної конференції. Київ, 1993. С. 42-46.
7. Данилюк І. Прикладна морфологія. Донецьк, 2010. 216 с.
8. Дарчук Н. Комп'ютерна лінгвістика. К., 2008. 351 с.

9. Дарчук Н. Морфологічне анотування Корпусу української мови. *Комп'ютерна лінгвістика: сучасне і майбутнє. Матеріали Міжнародної науково-практичної конференції*. К., 2012. С. 16-19.
10. Карпіловська Є. Вступ до прикладної лінгвістики: Комп'ютерна лінгвістика. Донецьк, 2006. 188 с.
11. Марчук Ю. Компьютерная лингвистика. М., 2007. 317 с.
12. Міщенко Н. Система програм морфологічного аналізу науково-технічних текстів. *Наукові записки. Серія «Філологічні науки»*. Випуск 95(2). К., 2011. С. 538-542.
13. Морфологический анализ научного текста на ЭВМ. К., 1989. 262 с.
14. Николаев И., Митренина О., Ландо Т. Прикладная и компьютерная лингвистика. М., 2016. 315 с.
15. Ножов И. Морфологическая и синтаксическая обработка текста (модели и программы). М., 2003. 140 с.
16. Партико З. Прикладна і комп'ютерна лінгвістика: Вступ до спеціальності. Львів, 2008. 224 с.
17. Романюк Ю. Прикладна морфологія. Черкаси, 2009. 117 с.
18. Соснина Е. Введение в прикладную лингвистику. Ульяновск, 2012. 110 с.
19. Сучасна українська літературна мова. Морфологія. К., 1969. 250 с.
20. Antworth E. PC-KIMMO: a two-level processor for morphological analysis. *Occasional Publications in Academic Computing*. No. 16. Dallas, 1990. 273 p.

Змістовий модуль ІІ

ОСНОВНІ МЕТОДИ АВТОМАТИЧНОГО МОРФОЛОГІЧНОГО АНАЛІЗУ

ТЕМА 1. АВТОМАТИЧНИЙ МОРФОЛОГІЧНИЙ АНАЛІЗ НА ОСНОВІ ГРАФЕМНОГО АНАЛІЗУ

Опорний конспект

1. Правильність і повнота здійсненого автоматичного морфологічного аналізу залежить від рівня: знань про мову і мовлення, тобто правильності лінгвістичної теорії, покладеної в основу АМА; формалізації цих знань у створюваній комп'ютерній граматиці. Принципи виведення морфологічних ознак слова за допомогою його структури: 1) здобуття граматичної інформації зі слова шляхом його графемного аналізу; 2) представлення граматичної інформації у словнику основ і словнику флексій.

2. Однією з найґрунтовніших робіт з АМА на основі графемного принципу був алгоритм визначення граматичних класів слів Герольда Бєлоногова. Він використовує списки кінцевих буквосполучень: 481 двобуквене буквосполучення; 1137 трибуквених буквосполучень; 3184 чотирибуквених буквосполучення; 3282 п'ятибуквені буквосполучення. При цьому виділяються такі граматичні класи слів: іменник, кількісний числівник; прикметник, порядковий числівник, повна форма дієприкметника; особова форма дієслова; дієслово минулого часу, скорочена форма прикметника і дієприкметника; інфінітив; прислівник, дієприслівник, порівняльні ступені прикметника; службові слова (прийменники, сполучники, частки). Пізніше цей досвід було узагальнено у книзі «Автоматизовані інформаційні системи», де представлено таблиці для визначення граматичних класів за кінцевими буквосполученнями словоформ. У наступних роботах ученого цей метод використаний як додатковий до аналізу на основі словника основ і словоформ.

3. Алгоритм виділення іменників та прикметників з тексту представлено у роботі Наталії Кравченко, яка наводить список усіх неіменникових кінцевих буквосполучень-закінчень і всіх іменникових, останні букви яких збігаються з якимось неіменниковим закінченням. Цей список містить 130 закінчень. Використовується додатковий список із 500 словоформ, які не мають формальних ознак приналежності до граматичного класу. Списки кінцевих буквосполучень застосовуються не лише для визначення частини мови, а і при виділенні умовної пошукової основи слова. При порівнянні словоформи зі списком флексій (кінцевих буквосполучень) флексія відсікається, а при пошуку використовується лише незмінна частина слова. Виділення основи методом усічення словоформи проводилося з метою підрахунку частоти слова або для полегшення пошуку словоформи за словником основ.

Контрольні питання

1. Назвіть принципи виведення морфологічних ознак слова за допомогою його структури. Як у цьому аспекті можуть прислужитися графеми, що утворюють словоформу?
2. Опишіть алгоритм визначення граматичних класів слів на основі графемного принципу (за методикою Герольда Бєлоногова).
3. Опишіть алгоритм виділення іменників та прикметників з тексту на основі списку кінцевих буквосполучень-закінчень (за методикою Наталії Кравченко).
4. Чому методика графемного аналізу виявилася доцільною для АМА текстів вторинних документів (рефератів, патентів тощо)?
5. Які тексти відрізняються обмеженою лексикою і стандартизованістю морфології (обмеженість у вживанні граматичних форм, синтаксичних структур)?

Домашнє завдання

Виділіть основні принципи укладання частотних словників на прикладі тих, що представлені за посиланням: http://www.mova.info/Page.aspx?l1=57.

Лабораторна робота №1
АВТОМАТИЧНИЙ МОРФОЛОГІЧНИЙ АНАЛІЗ НА ОСНОВІ ГРАФЕМНОГО АНАЛІЗУ

1. Розподіліть усі словоформи отриманого тексту за поданими нижче граматичними класами слів (за методикою АМА Герольда Бєлоногова – див. Табл. 1).

Варіант 1

Універсальним носієм знань є природна мова. Мова – не тільки форма вираження думок, а й єдиний і найважливіший засіб змістової організації та представлення знань. Функція мовної системи – бути засобом породження, збереження і передавання інформації. Це означає, що на перший план комп'ютеризації інтелектуальної людської діяльності виходять лінгвістичні аспекти. Будь-які проблеми інформатики, штучного інтелекту, пов'язані з цими аспектами, сприяють усвідомленню загальнонаукової й загальнодержавної важливості мовознавчої науки. У зв'язку з активними процесами інформатизації діяльності сучасного суспільства зростає значення комп'ютерної лінгвістики, яка поєднує мовознавство – глибинні знання про мову – та кібернетику – комп'ютерні технології. Взаємодія людини і комп'ютера сприяє інтенсифікації предметної галузі мовознавства – комп'ютерної лінгвістики. Прикладні лінгвістичні завдання переважно є соціальним замовленням: передавання усного мовлення різними каналами зв'язку, автоматичне розпізнавання змісту тексту, машинний переклад, автоматичне реферування, мережеве представлення даних, уніфікація і стандартизація термінології (у тому числі створення термінологічних словників, баз даних і знань), укладання корпусів текстів і проєктування автоматизованого робочого місця лінгвіста тощо. Водночас комп'ютеризація сприяє поглибленню теоретичного мовознавства. Від цієї взаємодії виграють як традиційна лінгвістика, так і комп'ютерна. Адаптація лінгвістичних знань для вирішення завдань штучного

(комп'ютерного) інтелекту, розроблення об'єктивних методів аналізу і формальне представлення мовних закономірностей, які можуть бути програмованими й адекватно відтворюваними в комп'ютері, є головними завданнями комп'ютерної лінгвістики. Деякі принципи цієї галузі як теоретичного, так і практичного характеру вже усталилися, отже, можна вважати, що вони цілком забезпечують потреби порівняно нової предметної галузі – комп'ютерної лінгвістики, яка в Україні на сьогодні перебуває на стадії розвитку. Оскільки основною формою представлення і збереження інформації є природна мова в усній і писемній формах, ефективність використання комп'ютера залежить значною мірою від розв'язання комплексу завдань автоматичного опрацювання текстів (АОТ), кінцева мета якого – розпізнавання їх змісту. В системах АОТ розглядаються два рівні опрацювання тексту залежно від мети і зумовленої нею глибини: 1) формальне опрацювання – аналіз і систематизація фрагментів тексту безвідносно до його змісту; 2) смислове опрацювання – розпізнавання змісту окремих елементів і логіко-семантичних відношень між ними з метою побудови семантичного представлення повідомлення. На цьому рівні виникає необхідність у використанні додаткової семантичної інформації, яка експліцитно не виражена в тексті. Якщо перший рівень – формальне опрацювання – є основою всіх наявних інформаційних технологій у діючих системах АОТ, то другий рівень нині є полем для теоретичних та експериментальних досліджень.

Варіант 2

Оскільки інформація організована засобами природної мови, її реальне засвоєння можливе лише за умови автоматичного смислового опрацювання текстів. Потреба в лінгвістичному забезпеченні обумовлена необхідністю створення систем «людина-машина-людина»: оперативна, зручна кооперація людини і машини повинна спиратися на природну мову. В соціальному плані значущість лінгвістичних проблем комп'ютеризації пов'язана з такими основними напрямками індустрії опрацювання знань, як збирання, зберігання,

систематизація, поширення, інтерпретація інформації, для чого створюється спеціальне лінгвістичне забезпечення. Лінгвістичне забезпечення автоматизованих систем – сукупність засобів для здійснення комп'ютеризації мовної діяльності – необхідне практично для будь-якої інтелектуальної діяльності людини. З технологічної точки зору йдеться про створення того чи іншого типу автоматичної системи опрацювання інформації, на вході і виході якої наявна текстова інформація природною мовою. Типи систем різноманітні й можуть бути спрямовані на моделювання різних мовних завдань, зокрема таких, як діалогова взаємодія, стиснення інформації, реферування тексту, логічне опрацювання змісту, переклад іншою мовою тощо. Прикладні системи, які створює лінгвіст у цій галузі, – це лінгвістично осмислені метамови – моделі представлення знань, кожна з яких базується на фундаментальних положеннях мовознавства і реалізується за допомогою методів структурно-математичної лінгвістики. Втілювана у прикладні завдання діалектична тріада «традиційна лінгвістика – структурно-математична лінгвістика – комп'ютерна лінгвістика» сприяє високому рівню лінгвістичного забезпечення автоматичних систем. Комп'ютерна граматика – це системний, строго впорядкований, формалізований, лінгвостатистичний, інтегральний опис знакових одиниць певної мови у вигляді структурних моделей із необхідною і достатньою аналітикою для виконання завдань штучного інтелекту, які відтворюють та імітують дослідницьку діяльність лінгвіста. Комп'ютерна граматика АГАТ має такі особливості. Перш за все, у ній дотримано рівневий підхід – рівні взаємодіють між собою від нижнього до верхнього, кожний наступний рівень використовує результати аналізу попереднього. Друга особливість – відкритість стратифікаційної структури граматики, що є принциповим моментом, оскільки дозволяє досить вільно розширювати обсяг лінгвістичного забезпечення, ускладнювати словникове та модифікувати програмне забезпечення без перебудови всієї системи.

Варіант 3

Обов'язковою частиною комп'ютерної граматики є автоматичний морфологічний аналіз (АМА) словоформ, тому що ані морфемний, ані синтаксичний, ані семантичний аналізи не можуть обійтися без визначення для словоформи її частини мови та словозмінних форм. До завдань АГАТ-морфології входять: автоматичне визначення для одиниць тексту граматичної інформації про місце їх у морфологічній системі мови; автоматична ідентифікація словоформ однієї лексеми. Для створення АГАТ-морфології української мови в теоретичному плані виконувалися дослідження, пов'язані: 1) з принципами частиномовної класифікації в українській мові; 2) з формальним обґрунтуванням морфологічних граматичних значень; 3 з принципами опису формальних засобів, характерних для відповідних частин мови з їх морфологічними значеннями. Морфологічна аксіоматика була налаштована на можливість алгоритмічного оперування граматичними даними. У прикладному аспекті створено словник квазіоснов, який налічує 210 тис. одиниць, і, відповідно, словник словоформ, які породжуються поєднанням інформації, взятої з таблиці основ і допоміжної таблиці, – близько 3,2 млн. слововживань, що забезпечує автоматичне приписування морфологічної інформації словоформам практично на 97%. Отже, АГАТ-морфологія української мови є автоматичним формально-морфологічним процесором з елементами морфолого-синтаксичного аналізу. Особливу увагу при створенні АГАТ-морфології приділено визначенню мовленнєвих умов, у яких реалізуються актуалізовані граматичні значення одиниці-омоніма. У теоретичному плані здійснено дослідження словниковозорієнтованих умов виникнення граматичних і лексико-граматичних омонімів в українській мові, що дало можливість укласти Граматичний словник омоформ. У прикладному аспекті визначено мовленнєві умови для реалізації значень досліджуваної словоформи, сформульовані за допомогою лінгвістичного методу – контекстного аналізу. В основу КА покладено твердження про те, що багатозначні елементи мови функціонують у своїх конкретних значеннях у

певному лексико-граматичному контексті. Реалізація цієї ідеї знайшла відображення у створенні автоматичного конкордансу, теоретичною основою якого є: 1) наявність таких визначників, за якими кожне значення словоформи (граматичне, лексичне) детермінується в контексті іншими словоформами, їх сполученнями або іншими текстовими ознаками; 2) текстоцентричний підхід до його створення: він укладається на певному масиві текстів для певної словоформи або лексеми. Такий словник-конкорданс вичерпно ілюструє використання певної лексеми і всіх її ЛСВ з лексико-граматичними значеннями.

Таблиця 1. Граматичні класи слів (за методикою АМА Герольда Бєлоногова)

№	Граматичний клас	Словоформи із тексту
1	іменник, кількісний числівник	
2	прикметник, порядковий числівник, повна форма дієприкметника	
3	особова форма дієслова	
4	дієслово минулого часу, скорочена форма прикметника і дієприкметника	
5	інфінітив	
6	прислівник, дієприслівник, порівняльні ступені прикметника	

7	службові слова (прийменники, сполучники, частки)	

2. У кожній словоформі визначте мінімальні кінцеві буквосполучення, що дозволяють віднести ту чи іншу словоформу до певного граматичного класу слів.

* Правильність визначених кінцевих буквосполучень та віднесеність словоформи до певного граматичного класу можете перевірити за Граматичним словником української мови http://www.mova.info/grmasl.aspx (див. Рис. 1-2):

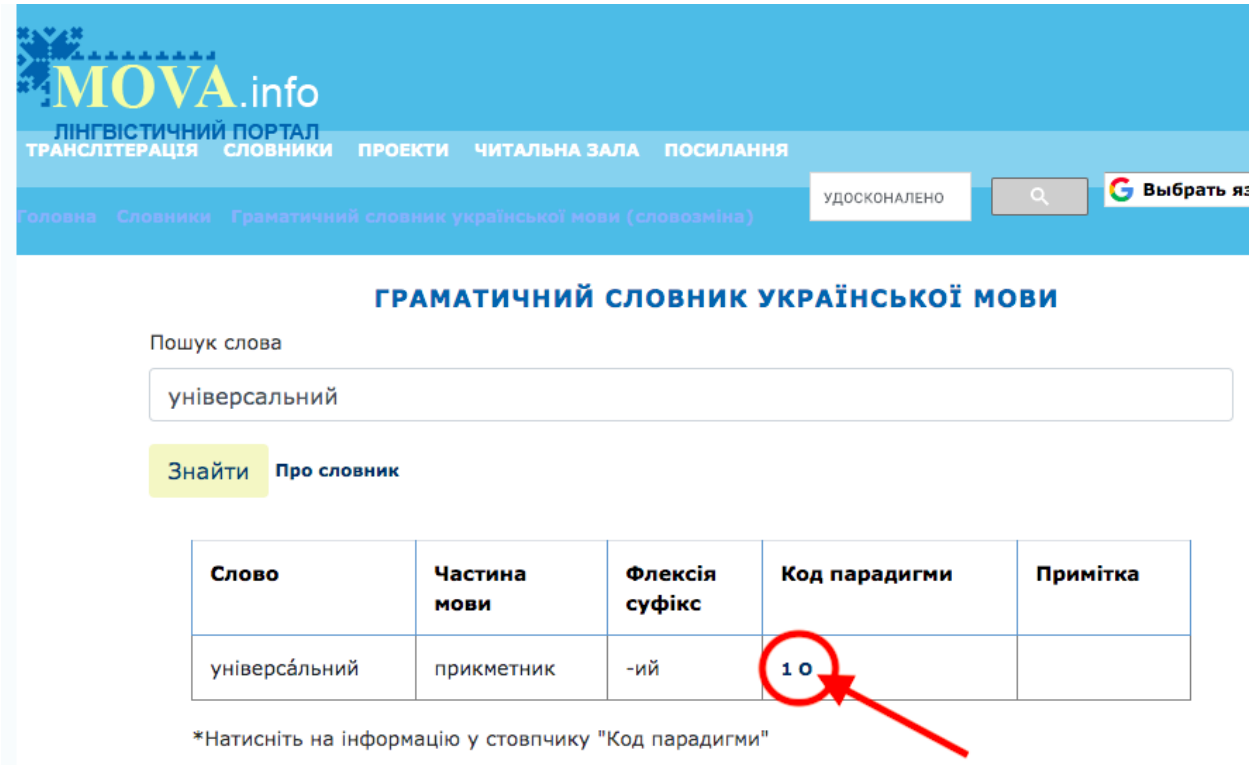


Рисунок 1. Пошук у Граматичному словнику української мови

Зразок парадигми

Код парадигми			1 О
Число	Рід	Відм.	**вели́кий**
одн.	ч.	Н	велик-ий
		Н	
		Р	велик-ого
		Д	велик-ому
		З	велик-ий
		Зі	велик-ого
		О	велик-им
		М	велик-ому
		М	велик-ім
		К	велик-ий
	с.	Н	велик-е
		Р	велик-ого
		Д	велик-ому
		З	велик-е
		О	велик-им
		М	велик-ому
		М	велик-ім
		К	велик-е
	ж.	Н	велик-а
		Р	велик-ої
		Д	велик-ій
		З	велик-у
		О	велик-ою
		М	велик-ій
		К	велик-а
мн.		Н	велик-і
		Р	велик-их
		Д	велик-им
		З	велик-і
		Зі	велик-их
		О	велик-ими

Рисунок 2. Зразок парадигми у Граматичному словнику української мови

Література до теми:

1. Бабина О. Корпусный метод автоматического морфологического анализа флективных языков. *Вестник Южно-Уральского государственного университета. Серия «Лингвистика»*. №25(284), выпуск 15. Челябинск, 2012. С. 38-44.
2. Баранов А. Введение в прикладную лингвистику. М., 2001. 360 с.
3. Белоногов Г. Компьютерная лингвистика и перспективные информационные технологии. М., 2004. 159 с.
4. Буньо Г. Сучасні методи вирішення проблеми граматичної омонімії в тексті. *Наукові записки. Серія: Філологічна*. 2014. Вип. 49. С. 12-16.
5. Волошин В. Комп'ютерна лінгвістика. Суми, 2004. 382 с.
6. Гельбух А., Сидоров Г. К вопросу об автоматическом морфологическом анализе флективных языков. Ел. режим доступу: www.dialog-21.ru/Archive/2005.
7. Грязнухіна Т., Нікула М. Система автоматичного морфологічного аналізу українського наукового тексту. Проблеми українізації

комп’ютерів. Матеріали 2-ї міжнародної конференції. Київ, 1993. С. 42-46.

8. Данилюк І. Прикладна морфологія. Донецьк, 2010. 216 с.
9. Дарчук Н. Комп’ютерна лінгвістика. К., 2008. 351 с.
10. Дарчук Н. Комп’ютерне анотування українського тексту: результати і перспективи. К., 2013. 543 с.
11. Дарчук Н. Морфологічне анотування Корпусу української мови. *Комп’ютерна лінгвістика: сучасне і майбутнє. Матеріали Міжнародної науково-практичної конференції*. К., 2012. С. 16-19.
12. Карпіловська Є. Вступ до прикладної лінгвістики: Комп’ютерна лінгвістика. Донецьк, 2006. 188 с.
13. Марчук Ю. Компьютерная лингвистика. М., 2007. 317 с.
14. Міщенко Н. Система програм морфологічного аналізу науково-технічних текстів. *Наукові записки. Серія «Філологічні науки»*. Випуск 95(2). К., 2011. С. 538-542.
15. Морфологический анализ научного текста на ЭВМ. К., 1989. 262 с.
16. Николаев И., Митренина О., Ландо Т. Прикладная и компьютерная лингвистика. М., 2016. 315 с.
17. Ножов И. Морфологическая и синтаксическая обработка текста (модели и программы). М., 2003. 140 с.
18. Партико З. Прикладна і комп’ютерна лінгвістика: Вступ до спеціальності. Львів, 2008. 224 с.
19. Романюк Ю. Прикладна морфологія. Черкаси, 2009. 117 с.
20. Соснина Е. Введение в прикладную лингвистику. Ульяновск, 2012. 110 с.
21. Сучасна українська літературна мова. Морфологія. К., 1969. 250 с.
22. Antworth E. PC-KIMMO: a two-level processor for morphological analysis. *Occasional Publications in Academic Computing*. No. 16. Dallas, 1990. 273 p.

ТЕМА 2. АВТОМАТИЧНИЙ МОРФОЛОГІЧНИЙ АНАЛІЗ НА ОСНОВІ ФЛЕКТИВНОГО АНАЛІЗУ

Опорний конспект

1. У мовах флективного типу (українська, російська), з розгалуженою системою словозміни, інформація про граматичні значення зосереджена в кінці слова і формально виражена флексією чи формотворчим суфіксом. Основним інструментом автоматичного морфологічного аналізу на основі флективного аналізу як засобу ідентифікації граматичної інформації є список квазіфлексій – кінцівок словоформ, що дозволяють однозначно встановлювати частиномовну приналежність словоформ тексту та їх граматичну характеристику.

2. У системі АМА список квазіфлексій можна укладати двома способами. Перший (принцип навчальної вибірки) передбачає автоматичне формування списку за вибіркою текстів, закодованих попередньо вручну в термінах граматичних класів. При цьому передбачається багаторазове автоматичне коригування списку за словоформами, доданими до вихідної вибірки, які були не розпізнані або розпізнані неправильно на незакодованому тексті. Другий (робочий принцип) передбачає формування списку квазіфлексій вручну на основі лінгвістичного аналізу оберненого словника словоформ, який укладається автоматично на певній вибірці текстів, з урахуванням даних Граматичного словника російської мови або Оберненого словника для української. У списку допускається вкладання квазіфлексії один в один, оскільки порівняння починається з найдовшої «кінцівки» при збігу останньої графеми у текстової словоформи й у квазіфлексії зі списку.

3. Для укладання списку квазіфлексій створюється дерево, яке репрезентує квазіфлексії з певною кінцевою графемою. З вершини виходить стільки стрілок, скільки графем може передувати за списком квазіфлексій графемі, що задана вершиною (Ø – вершина, яка вказує, що даній квазіфлексії у текстовій словоформі може передувати інша графема; : – словоформа задана

повністю; останні елементи гілок дерева – коди граматичних класів, які будуть приписуватися словоформам у тексті, якщо алгоритм встановить тотожність їх кінцівок з відповідними гілками дерева).

4. Аналіз графемної структури словоформи може бути використаний не лише як інструмент ідентифікації лексико-граматичних класів у тексті, а і при визначенні граматичних підкласів (у межах класу). Такий аналіз реалізується на тих же принципах флективного аналізу за допомогою списків квазіфлексій, які використовуються на етапі визначення класів слів. Його завданням є опрацювання текстових одиниць (узятих окремо, поза контекстом) для визначення набору можливих граматичних значень, які репрезентують словозміну й узгодження з іншими одиницями в тексті, представляючи підклас певного коду лексико-граматичного класу слів. Підкласи різних граматичних класів відрізняються за граматичною природою: у змінюваних класів слів вони є виразниками суто морфологічних характеристик; у службових частин мови (прийменника, сполучника) підклас вказує на синтаксичні особливості (код підкласу прийменника містить інформацію про відмінки, якими даний прийменник може керувати; код підкласу сполучника вказує на тип зв'язку (сурядний чи підрядний), що формується за участі сполучника).

5. Недоліками вищеописаного способу АМА є те, що: жодний аналіз вирваної з контексту словоформи не може в усіх випадках визначити однозначно належність її до певного лексико-граматичного класу слів; немає жодної словозмінної парадигми слова іменних лексико-граматичних класів слів, у якій була б відсутня омонімія словозмінних форм (омоформи слова). Ці чинники зумовили введення поняття диз'юнктивних кодів класів і підкласів, які будуть аналізуватися на наступному етапі контекстного аналізу, що по суті є синтаксичним (позиційним) аналізом, що базується на синтаксичних зв'язках словоформ у тексті.

Контрольні питання

1. Дайте визначення списку квазіфлексій.

2. Опишіть два принципи укладання списку квазіфлексій у процесі АМА на основі флективного аналізу.
3. Як виглядає дерево квазіфлексій? Наведіть приклади.
4. Як аналіз графемної структури словоформи може прислужитися у визначенні граматичних підкласів (у межах лексико-граматичного класу)?
5. Опишіть недоліки АМА на основі флективного аналізу. Як можна їх подолати?

Домашнє завдання

Напишіть коротке висловлення (5-7 речень) щодо питання, чи розпізнавальний потенціал квазіфлексій, за якими здійснюється ідентифікація граматичних підкласів, у різних частин мови різний? Проілюструйте на прикладах словозмінних парадигм дієслова, іменника, прикметника.

Лабораторна робота №2
АВТОМАТИЧНИЙ МОРФОЛОГІЧНИЙ АНАЛІЗ НА ОСНОВІ ФЛЕКТИВНОГО АНАЛІЗУ

1. Користуючись ресурсом http://lcorp.ulif.org.ua/dictua/ (див. Рис. 3-5), визначте парадигму поданих нижче слів та припишіть словоформам граматичні коди (наприклад, для іменника *особа* – ІЖОН, де І – імен., Ж – ж.р., О – одн., Н – Н.в.; для прикметника *білий* – ПЧОН, де П – прикм., Ч – ч.р., О – одн., Н – Н.в.; для дієслова *хотіти* – ДНІ, де Д – дієсл., Н – недок.в., І – інфінітивна форма):

– іменників жіночого роду (*особа, вада, голова, книга, ручка, відьма, кар'єра, радість, стаття*), чоловічого роду (*птах, пристрій, прохід, прогін, олівець, плюскіт, пролісок, стіл, день, тиждень, достаток, службовець, курінь, куркуль, жаль, корабель, журавель, ведмідь, товариш, прохід*), середнього роду (*мовчання, волосся, узбіччя, теля,*

яйце, кільце, щастя, життя, листя, почуття, сторіччя, місце, віконце, слівце, лице, щеня, ім'я, весілля, тім'я);

– прикметників (*білий, гарний, синій, третій, довгошиїй*);

– дієслів (*хотіти, любити, ненавидіти, дати, жити, давати, мережити, бити, вказувати*).

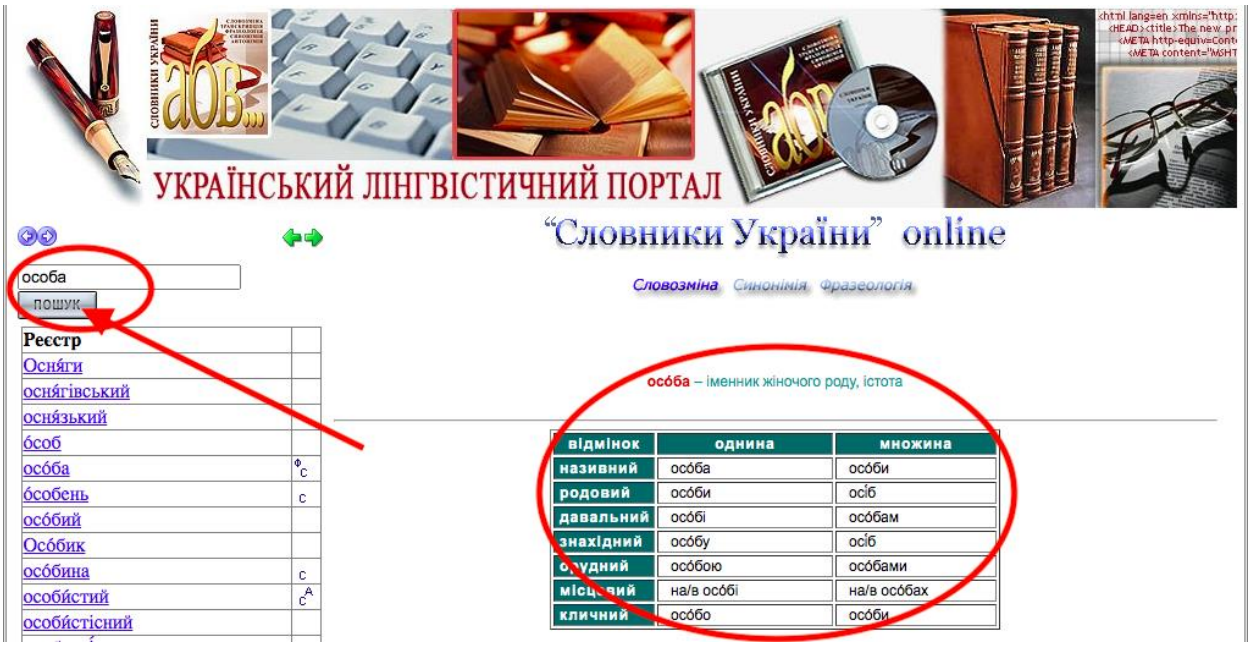


Рисунок 3. Пошук іменника

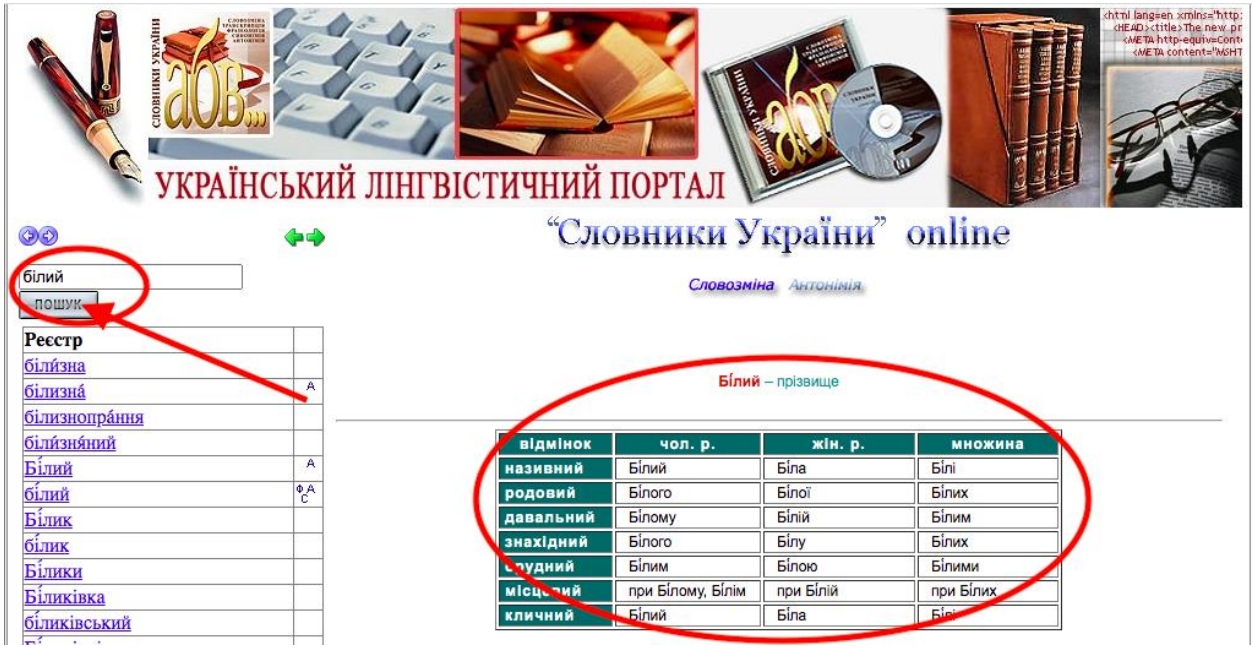


Рисунок 4. Пошук прикметника

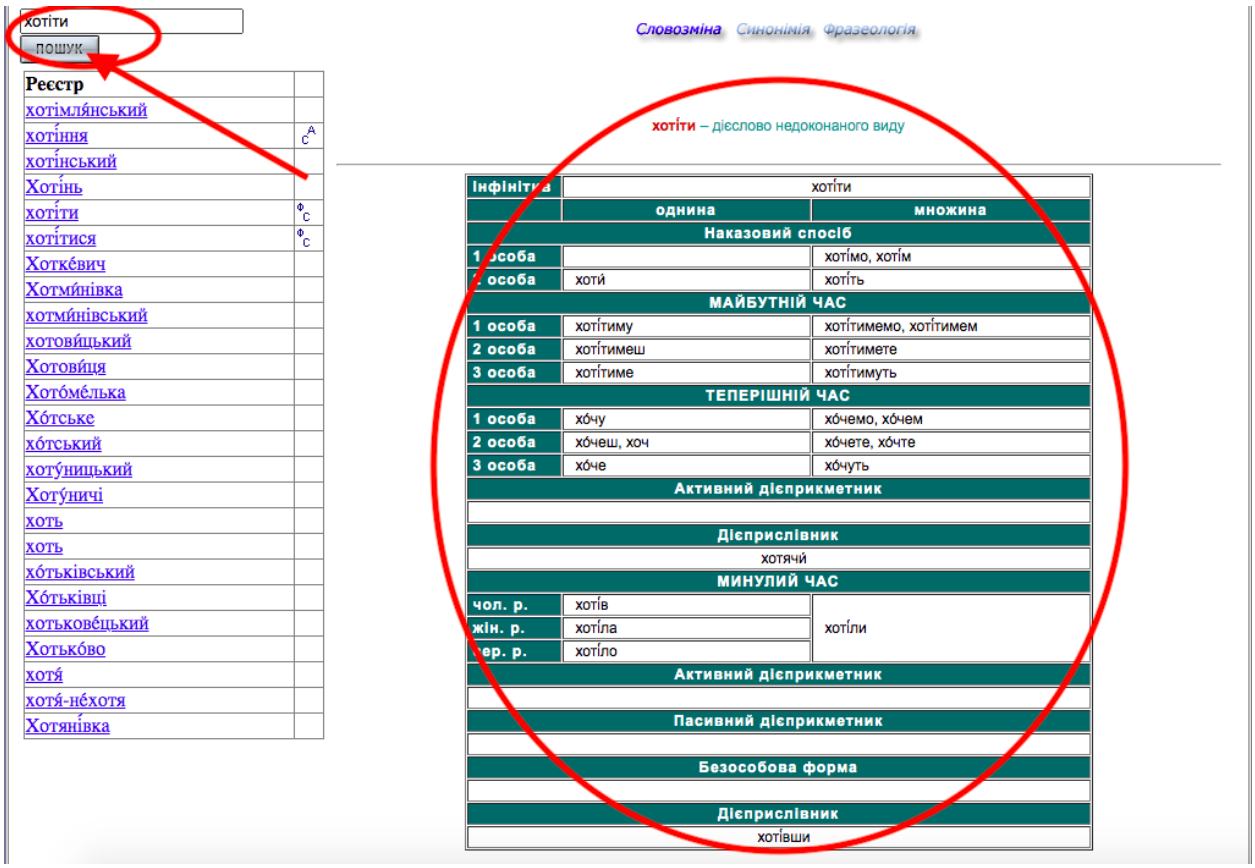


Рисунок 5. Пошук дієслова

2. Розташуйте отримані словоформи в інверсійному порядку. Як зразок можна використовувати словник https://drive.google.com/file/d/0B4gL7dSktTplc3dzdEtIN2dpelU/view (див. Рис. 6):

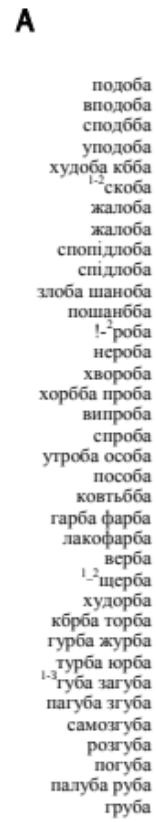


Рисунок 6. Зразок інверсійного розташування словоформ

3. За принципом дерева виділіть квазіфлексії, що дозволяють визначити граматичні підкласи зазначених слів (див. Рис. 7):

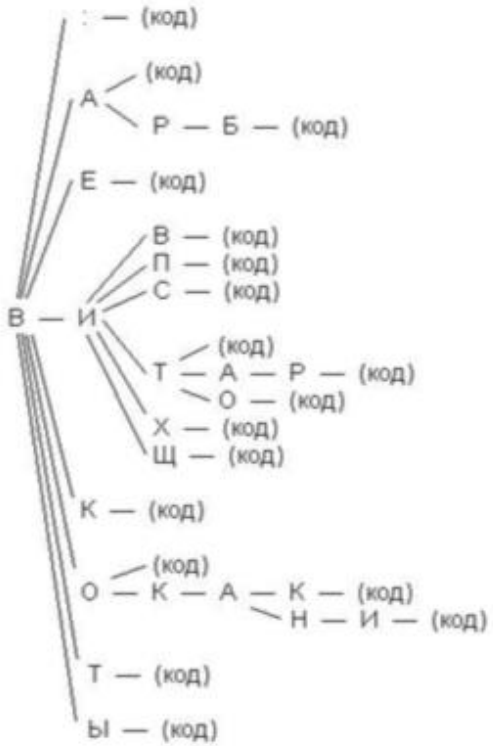


Рисунок 7. Фрагмент алгоритму визначення граматичних класів слів у вигляді дерева квазіфлексій з кінцевою літерою ***в***

Література до теми:

1. Бабина О. Корпусный метод автоматического морфологического анализа флективных языков. *Вестник Южно-Уральского государственного университета. Серия «Лингвистика»*. №25(284), выпуск 15. Челябинск, 2012. С. 38-44.
2. Баранов А. Введение в прикладную лингвистику. М., 2001. 360 с.
3. Белоногов Г. Компьютерная лингвистика и перспективные информационные технологии. М., 2004. 159 с.
4. Буньо Г. Сучасні методи вирішення проблеми граматичної омонімії в тексті. *Наукові записки. Серія: Філологічна*. 2014. Вип. 49. С. 12-16.
5. Волошин В. Комп'ютерна лінгвістика. Суми, 2004. 382 с.
6. Гельбух А., Сидоров Г. К вопросу об автоматическом морфологическом анализе флективных языков. Ел. режим доступу: www.dialog-21.ru/Archive/2005.
7. Грязнухіна Т., Нікула М. Система автоматичного морфологічного аналізу українського наукового тексту. Проблеми українізації комп'ютерів. Матеріали 2-ї міжнародної конференції. Київ, 1993. С. 42-46.
8. Данилюк І. Прикладна морфологія. Донецьк, 2010. 216 с.
9. Дарчук Н. Комп'ютерна лінгвістика. К., 2008. 351 с.
10. Дарчук Н. Комп'ютерне анотування українського тексту: результати і перспективи. К., 2013. 543 с.
11. Дарчук Н. Морфологічне анотування Корпусу української мови. *Комп'ютерна лінгвістика: сучасне і майбутнє. Матеріали Міжнародної науково-практичної конференції*. К., 2012. С. 16-19.
12. Карпіловська Є. Вступ до прикладної лінгвістики: Комп'ютерна лінгвістика. Донецьк, 2006. 188 с.
13. Марчук Ю. Компьютерная лингвистика. М., 2007. 317 с.

14. Міщенко Н. Система програм морфологічного аналізу науково-технічних текстів. *Наукові записки. Серія «Філологічні науки»*. Випуск 95(2). К., 2011. С. 538-542.
15. Морфологический анализ научного текста на ЭВМ. К., 1989. 262 с.
16. Николаев И., Митренина О., Ландо Т. Прикладная и компьютерная лингвистика. М., 2016. 315 с.
17. Ножов И. Морфологическая и синтаксическая обработка текста (модели и программы). М., 2003. 140 с.
18. Партико З. Прикладна і комп'ютерна лінгвістика: Вступ до спеціальності. Львів, 2008. 224 с.
19. Романюк Ю. Прикладна морфологія. Черкаси, 2009. 117 с.
20. Соснина Е. Введение в прикладную лингвистику. Ульяновск, 2012. 110 с.
21. Сучасна українська літературна мова. Морфологія. К., 1969. 250 с.
22. Antworth E. PC-KIMMO: a two-level processor for morphological analysis. *Occasional Publications in Academic Computing*. No. 16. Dallas, 1990. 273 p.

ТЕМА 3. АВТОМАТИЧНИЙ МОРФОЛОГІЧНИЙ АНАЛІЗ НА ОСНОВІ СЛОВНИКА СЛОВОФОРМ

Опорний конспект

1. Автоматичний морфологічний аналіз проводять точними й наближеними методами. Точні методи засновані на використанні словника основ слів або словоформ, наближені – на експериментально встановленому зв'язку між кінцевими буквосполученнями словоформ та їх граматичною

інформацією. Використання словника словоформ у точних методах дозволяє легко долати складнощі морфологічного аналізу, пов’язані з такими явищами в українській, російській та інших флективних мовах, як чергування голосних і приголосних, наявність суплетивних форм слів.

2. При здійсненні морфологічного аналізу на базі словника словоформ завдання одержати граматичні ознаки зводиться до пошуку у словнику і вибору відповідної інформації. У роботах Герольда Бєлоногова (1975) наведено алгоритм і результати морфологічного аналізу на базі словника словоформ. Процедура аналізу полягає в тому, що проводиться ототожнення вихідної словоформи за словником, і при позитивній відповіді для неї обираються граматичні й семантичні ознаки, у протилежному випадку словоформа заноситься (у режимі поповнення) до словника, а її ознаки визначаються за допомогою аналізу буквеного коду словоформи: граматичні – за словником п’ятибуквених кінцівок слів, семантичні – за словниками словоформ, флексій і сполучень суфіксів.

3. Юрій Марчук виділяє такі проблеми, які лишаються при здійсненні аналізу за словником словоформ: аналіз не знайдених у словнику словоформ, адже визначення інформації для словоформи, не знайденої у словнику, є необхідним для наступного етапу аналізу (синтаксичного), коли треба визначити щонайменше частину мови; ототожнення різних словоформ одного й того самого слова: якщо кожна словоформа буде реєструватися як самостійна лексична одиниця, то це істотно ускладнить весь наступний аналіз і синтез форм через граматичну і лексико-граматичну омонімію. Однак при досить повному словнику словоформ частка операцій з аналізу невелика, програмне забезпечення нескладне, що забезпечує швидке створення системи АМА.

<u>Контрольні питання</u>

1. У чому суть точних і наближених методів АМА?
2. Які складнощі морфологічного аналізу долає використання словника словоформ?

3. Опишіть процедуру морфологічного аналізу на базі словника словоформ (за методикою Герольда Бєлоногова).
4. Які проблеми лишаються при здійсненні АМА за словником словоформ? Як їх можна подолати?

Домашнє завдання

Наведіть приклади українських, російських словоформ з чергуванням голосних і приголосних, суплетивними основами.

Лабораторна робота №3
АВТОМАТИЧНИЙ МОРФОЛОГІЧНИЙ АНАЛІЗ НА ОСНОВІ СЛОВНИКА СЛОВОФОРМ

1. Ознайомтеся з інструкцією до Електронного граматичного словника української літературної мови: http://www.mova.info/Page.aspx?l1=222 (див. Рис. 8-9):

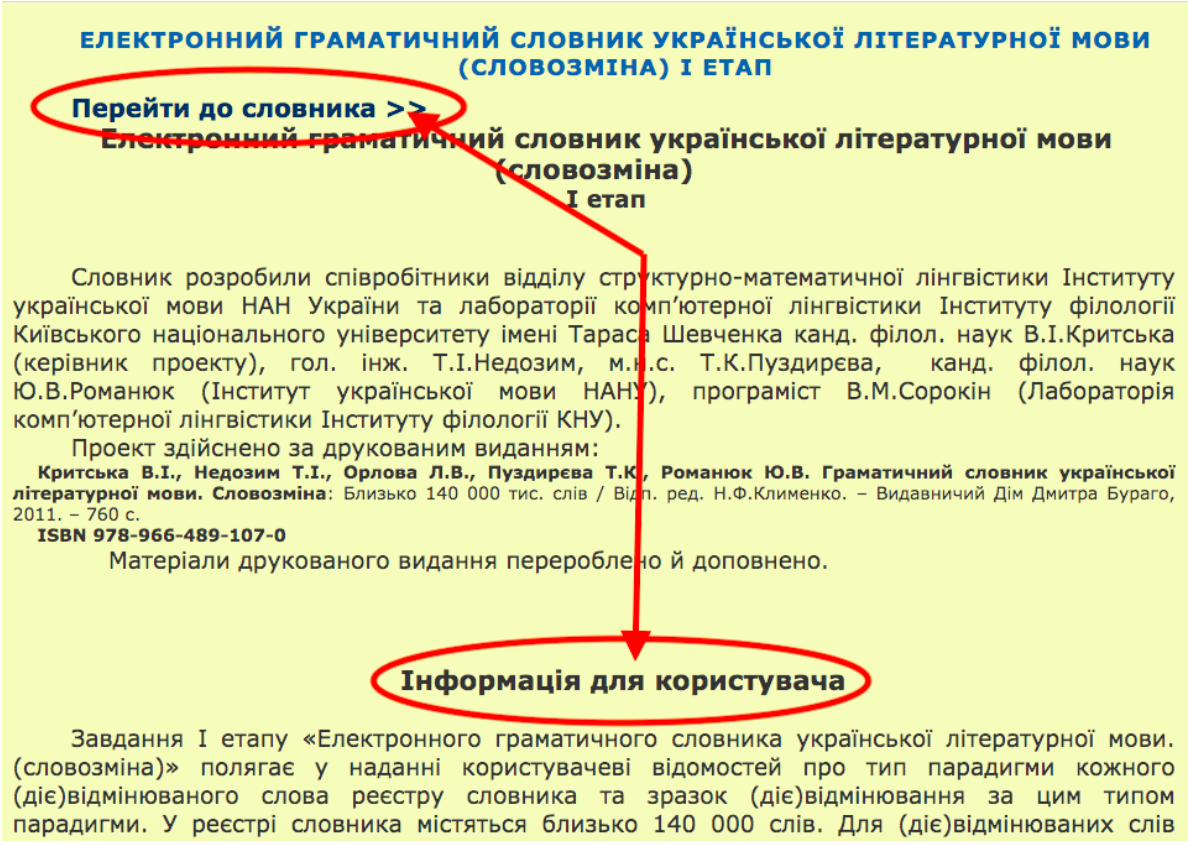


Рисунок 8. Інструкція до Електронного граматичного словника української літературної мови

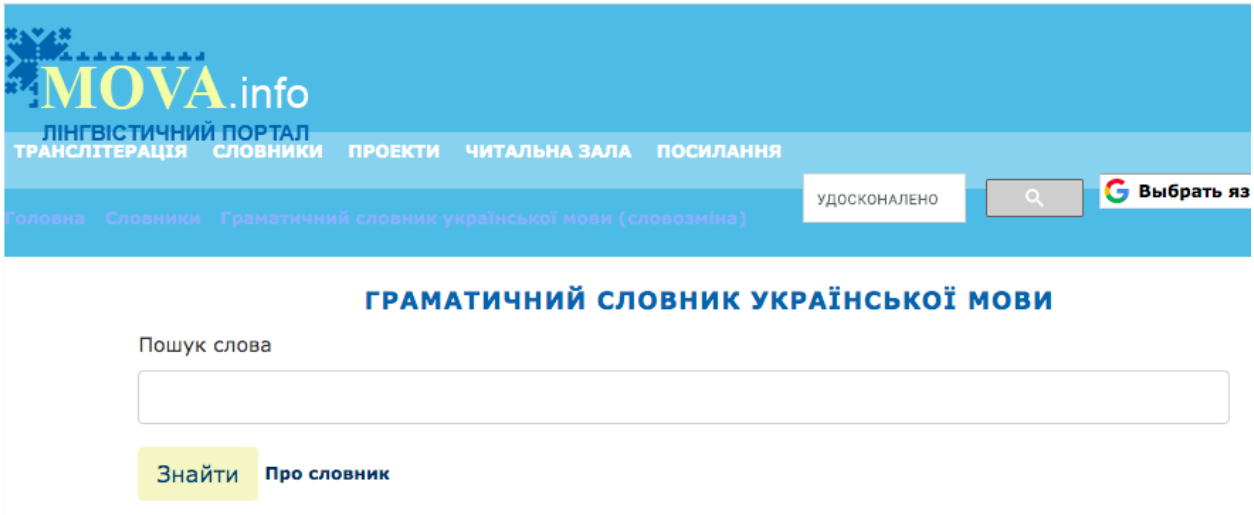


Рисунок 9. Пошук у Електронному граматичному словнику української літературної мови

2. Прокоментуйте, які граматичні ознаки словоформ можна взяти з поданого словника (на прикладі усіх частин мови – візьміть 10 різних лем).

3. Як, на вашу думку, можна вдосконалити функціонал Словника?

Література до теми:

1. Бабина О. Корпусный метод автоматического морфологического анализа флективных языков. *Вестник Южно-Уральского государственного университета. Серия «Лингвистика»*. №25(284), выпуск 15. Челябинск, 2012. С. 38-44.
2. Баранов А. Введение в прикладную лингвистику. М., 2001. 360 с.
3. Белоногов Г. Компьютерная лингвистика и перспективные информационные технологии. М., 2004. 159 с.
4. Буньо Г. Сучасні методи вирішення проблеми граматичної омонімії в тексті. *Наукові записки. Серія: Філологічна*. 2014. Вип. 49. С. 12-16.
5. Волошин В. Комп'ютерна лінгвістика. Суми, 2004. 382 с.
6. Гельбух А., Сидоров Г. К вопросу об автоматическом морфологическом анализе флективных языков. Ел. режим доступу: www.dialog-21.ru/Archive/2005.
7. Грязнухіна Т., Нікула М. Система автоматичного морфологічного аналізу українського наукового тексту. Проблеми українізації

комп’ютерів. Матеріали 2-ї міжнародної конференції. Київ, 1993. С. 42-46.

8. Данилюк І. Прикладна морфологія. Донецьк, 2010. 216 с.
9. Дарчук Н. Комп’ютерна лінгвістика. К., 2008. 351 с.
10. Дарчук Н. Комп’ютерне анотування українського тексту: результати і перспективи. К., 2013. 543 с.
11. Дарчук Н. Морфологічне анотування Корпусу української мови. *Комп’ютерна лінгвістика: сучасне і майбутнє. Матеріали Міжнародної науково-практичної конференції*. К., 2012. С. 16-19.
12. Карпіловська Є. Вступ до прикладної лінгвістики: Комп’ютерна лінгвістика. Донецьк, 2006. 188 с.
13. Марчук Ю. Компьютерная лингвистика. М., 2007. 317 с.
14. Міщенко Н. Система програм морфологічного аналізу науково-технічних текстів. *Наукові записки. Серія «Філологічні науки»*. Випуск 95(2). К., 2011. С. 538-542.
15. Морфологический анализ научного текста на ЭВМ. К., 1989. 262 с.
16. Николаев И., Митренина О., Ландо Т. Прикладная и компьютерная лингвистика. М., 2016. 315 с.
17. Ножов И. Морфологическая и синтаксическая обработка текста (модели и программы). М., 2003. 140 с.
18. Партико З. Прикладна і комп’ютерна лінгвістика: Вступ до спеціальності. Львів, 2008. 224 с.
19. Романюк Ю. Прикладна морфологія. Черкаси, 2009. 117 с.
20. Соснина Е. Введение в прикладную лингвистику. Ульяновск, 2012. 110 с.
21. Сучасна українська літературна мова. Морфологія. К., 1969. 250 с.
22. Antworth E. PC-KIMMO: a two-level processor for morphological analysis. *Occasional Publications in Academic Computing*. No. 16. Dallas, 1990. 273 p.

ТЕМА 4. АВТОМАТИЧНИЙ МОРФОЛОГІЧНИЙ АНАЛІЗ НА ОСНОВІ СЛОВНИКА ОСНОВ

Опорний конспект

1. Найбільш розповсюдженим видом автоматичного морфологічного аналізу є аналіз на базі словника основ, використовуваний для більшості європейських мов. У цьому виді аналізу використовується словник основ слів і низка допоміжних таблиць. Усі основи поділяються на 4 типи: незмінні (I тип); із чергуванням голосних (II тип); із чергуванням приголосних (III тип); незмінні основи, які не ввійшли до другого та третього типів (IV тип). Цей метод ґрунтується на флективному аналізі, мета якого – правильне виділення основи слова. Аналіз здійснюється за допомогою морфологічної таблиці. За номерами закінчення і флективного класу із таблиці обирається частиномовна і граматична (категоріальна) інформація. Структура таблиці передбачає перевірку правильності членування слова на основу й закінчення.

2. Ці принципи аналізу закладені в алгоритмі Лідії Кноріної. Розробники морфологічного аналізу, використовуваного в системі автоматичного індексування ПСИХЕЯ-2, застосовували комбінований спосіб, основою якого був флективний аналіз на базі основ і флексій у взаємодії з аналізом граматичних норм порядку слів у реченнях, написаних російською мовою, із семантичною обробкою інформації.

3. Ототожнення слів на етапі цього виду АМА відбувається за такими граматичними класами: основні класи (іменники; прикметники, включаючи порядкові числівники і дієприкметники у повній формі; особові форми дієслова); додаткові класи (кількісні числівники, прислівники, інфінітиви і дієприслівники). Основні класи розбито на підкласи за ознаками відмінка, роду, числа. Клас іменників розбито на 36 підкласів, кожний із який однозначно визначає рід, число, відмінок і має відповідно свій номер, вказаний у Таблицях 2, 3:

Таблиця 2. Список морфологічних кодів для іменників в однині (за методикою АМА Герольда Бєлоногова)

Число	Однина		
Рід	Середній	Жіночий	Чоловічий
Відмінок	И Р Д В Т П	И Р Д В Т П	И Р Д В Т П
Номер класу	1 2 3 4 5 6	7 8 9 10 11 12	13 14 15 16 17 18

Таблиця 3. Список морфологічних кодів для іменників у множині (за методикою АМА Герольда Бєлоногова)

Число	Множина		
Рід	Середній	Жіночий	Чоловічий
Відмінок	И Р Д В Т П	И Р Д В Т П	И Р Д В Т П
Номер класу	19 20 21 22 23 24	25 26 27 28 29 30	31 32 33 34 35 36

Клас прикметників розбитий на 24 підкласи (див. Табл. 4):

Таблиця 4. Список морфологічних кодів для ад'єктивних класів (за методикою АМА Герольда Бєлоногова)

Число	Однина			Множина
Рід	Середній	Жіночий	Чоловічий	-
Відмінок	И Р Д В Т П	И Р Д В Т П	И Р Д В Т П	И Р Д В Т П
Номер класу	37 38 39 40 41 42	43 44 45 46 47 48	49 50 51 52 53 54	55 56 57 58 59 60

Клас дієслів складається із 5 підкласів, визначуваних числом, родом та інфінітивом (див. Табл. 5):

Таблиця 5. Список морфологічних кодів для дієслова у минулому часі (за методикою АМА Герольда Бєлоногова)

Число	Однина			Інфінітив	Множина
Рід	Середній	Жіночий	Чоловічий		
Номер класу	61	62	63	64	65

Класи прислівників і кількісних числівників відповідно мають номери 66 та 67.

4. Словник флексій включає 157 закінчень тих граматичних класів і груп, характерних для слів галузевих текстів. Фрагмент словника флексій російської мови представлено у Таблиці 6:

Таблиця 6. Список флексій і граматичної інформації до них

Флексія	Номери граматичного класу
а	7, 14, 19, 22
на	7, 14
та	14
ана	62
ена	62
ята	62
ута	62
има	62
мена	7, 14
ства	2, 19, 22
ав	13,26
ев	32
ов	32
ств	20
е	9, 12, 1, 4, 6, 15, 18
ке	9, 12
не	9, 12
те	9, 12
ие	55, 58
ые	55, 58
ое	37, 40
ее	37, 40

5. Такий вид аналізу дає на стилістично і тематично обмеженому тексті непогані результати морфологічного аналізу. Однак наявність у другій колонці таблиці кількох номерів граматичних класів свідчить про омонімічність форм, яка буде знята на наступному, семантико-синтаксичному етапі аналізу.

6. В Україні розробники АМА української і російської мов використовували комбінований метод, у якому поєднувалися дві таблиці: таблиця основ (сталої, незмінної частини слова та змінної частини слова без флексії) – Таблиця 7 – і додаткова таблиця флексій (флексії або змінна частина слова із флексіями) із частиномовною і категоріальною характеристикою (рід, число, відмінок, особа, час). Кожній лексемі, яка має словозміну, приписувався номер парадигматичного класу, для якого у додатковій таблиці наводилися форми словозміни (Таблиця 8):

Таблиця 7. Фрагмент словника основ

Незмінна частина слова	Змінна частина слова	Код парадигматичного класу
ст	іл	й
став	ок	й

Таблиця 8. Фрагмент словника змінної частини слова із флексіями

Код пар. класу	Код частини мови (перша буква) і грам. характеристик (друга буква)	Клас словоформи	Словозмінний піджлас	Змінна частина слова
1	ЙИ	Ім. ч. р.	Наз. одн.	іл
1	ЙР	Ім. ч. р.	Род. одн.	ола
1	ЙД	Ім. ч. р.	Дав. одн.	олу
1	ЙВ	Ім. ч. р.	Знах. одн.	іл
1	ЙТ	Ім. ч. р.	Орудн. одн.	олом
1	ЙП	Ім. ч. р.	Місц. одн.	олі
1	ЙА	Ім. ч. р.	Наз. мн.	оли
1	ЙЕ	Ім. ч. р.	Род. мн.	олів
1	ЙО	Ім. ч. р.	Дав. мн.	олам
1	ЙУ	Ім. .ч. р.	Знах. мн.	оли
1	ЙЮ	Ім. ч. р.	Орудн. мн.	олами
1	ЙЯ	Ім. ч. р.	Місц. мн.	олах
2	ЙИ	Ім. ч. р.	Наз. одн	ок
2	ЙР	Ім. ч. р.	Род. одн	ка
2	ЙД	Ім. ч. р.	Дав. одн	ку
2	ЙВ	Ім. ч. р.	Знах. одн	ок
2	ЙТ	Ім. ч. р.	Орудн. одн	ком
2	ЙП	Ім. ч. р.	Місц. одн	ку
2	ЙА	Ім. ч. р.	Наз. мн	ки
2	ЙЕ	Ім. ч. р.	Род. мн	ків
2	ЙО	Ім. ч. р.	Дав. мн	кам
2	ЙУ	Ім. ч. р.	Знах. мн	ки
2	ЙЮ	Ім. ч. р.	Орудн. мн	ками
2	ЙЯ	Ім. ч. р.	Місц. мн	ках

За цими таблицями АМА відбувається в такій послідовності: 1) кожна текстова словоформа порівнюється з основами словника основ (Таблиця 7) на максимальний графемний збіг у колонці «незмінна частина основи»; 2) при позитивній відповіді аналіз продовжується за Таблицею 8 і в разі збігу зі словозмінною формою приписується інформація про граматичні значення слова, а у разі омонімії (кількох однакових графемно виражених слів) – ланцюжок граматичних значень у послідовності кодів – алфавітно-цифрових, цифрових, які в машинних системах використовуються для швидкості опрацювання комп'ютером (описова назва – ім. ч. р. наз. відм. мн. використовується лише при взаємодії з людиною).

Контрольні питання

1. Опишіть суть процедури АМА на базі словника основ. Які типи основ він використовує?

2. За якими граматичними класами відбувається ототожнення слів на етапі АМА на базі словника основ?
3. Опишіть алгоритм комбінованого методу АМА на базі словника основ (з використанням таблиці основ і додаткової таблиці флексій).

Домашнє завдання

За аналогією до Таблиць 7 та 8 подайте свої варіанти фрагменту словника основ та фрагменту словника змінної частини слова із флексіями для слів *кріт*, *кросівок*.

Лабораторна робота №4
АВТОМАТИЧНИЙ МОРФОЛОГІЧНИЙ АНАЛІЗ НА ОСНОВІ СЛОВНИКА ОСНОВ

1. Скопіюйте фрагмент тексту публіцистичного стилю довжиною в 2000 знаків без пробілів.

2. Визначте в тексті повнозначні слова. Розподіліть їх за такими граматичними класами: 1) ***основні:*** іменники; прикметники (+ порядкові числівники і дієприкметники у повній формі); особові форми дієслова; 2) ***додаткові:*** кількісні числівники, прислівники, інфінітиви і дієприслівники.

3. Для **іменників** (за потреби користуючись словником словозміни http://lcorp.ulif.org.ua/dictua/) визначте й пропишіть у таблиці такі граматичні ознаки (див. Табл. 9):

Таблиця 9. Граматичні ознаки іменників

Словоформа з тексту	Незмінна частина слова	Змінна частина слова	Рід	Число	Відмінок
стіл	ст	іл	ч	одн	Н. (З.)

4. Присвойте визначеним іменникам код парадигматичного класу: І (іменник) + С (словозмінний) / Н (не словозмінний).

5. Для **дієслів** (за потреби користуючись словником словозміни http://lcorp.ulif.org.ua/dictua/) визначте й пропишіть у таблиці такі граматичні ознаки (див. Табл. 10):

Таблиця 10. Граматичні ознаки дієслів

Словоформа з тексту	Незмінна частина слова	Змінна частина слова	Вид	Час	Число	Особа	Рід
робити	роб	ити	недоконаний	-	-	-	-

6. Присвойте визначеним дієсловам код парадигматичного класу: Д (дієслово) + Н (недоконаний) / Д (доконаний).

7. Для **прикметників** (за потреби користуючись словником словозміни http://lcorp.ulif.org.ua/dictua/) визначте й пропишіть у таблиці такі граматичні ознаки (див. Табл. 11):

Таблиця 11. Граматичні ознаки прикметників

Словоформа з тексту	Незмінна частина слова	Змінна частина слова	Рід	Число	Відмінок
розумний	розумн	ий	ч	одн	Н. (З.)

8. Присвойте визначеним прикметникам код парадигматичного класу: П (прикметник) + Ч (чоловічий) / Ж (жіночий) / С (середній) / М (множина).

9. На основі проаналізованих вище частин мови (іменників, дієслів, прикметників) укладіть таблицю флексій за поданим зразком (див. Табл. 12):

Таблиця 12. Таблиця флексій

Флексія (змінна частина слова)	Код парадигматичного класу
іл	ІС
ити	ДН
ий	ПЧ

Література до теми:

1. Бабина О. Корпусный метод автоматического морфологического анализа флективных языков. *Вестник Южно-Уральского государственного университета. Серия «Лингвистика»*. №25(284), выпуск 15. Челябинск, 2012. С. 38-44.
2. Баранов А. Введение в прикладную лингвистику. М., 2001. 360 с.
3. Белоногов Г. Компьютерная лингвистика и перспективные информационные технологии. М., 2004. 159 с.
4. Буньо Г. Сучасні методи вирішення проблеми граматичної омонімії в тексті. *Наукові записки. Серія: Філологічна*. 2014. Вип. 49. С. 12-16.
5. Волошин В. Комп'ютерна лінгвістика. Суми, 2004. 382 с.
6. Гельбух А., Сидоров Г. К вопросу об автоматическом морфологическом анализе флективных языков. Ел. режим доступу: www.dialog-21.ru/Archive/2005.
7. Грязнухіна Т., Нікула М. Система автоматичного морфологічного аналізу українського наукового тексту. Проблеми українізації комп'ютерів. Матеріали 2-ї міжнародної конференції. Київ, 1993. С. 42-46.
8. Данилюк І. Прикладна морфологія. Донецьк, 2010. 216 с.
9. Дарчук Н. Комп'ютерна лінгвістика. К., 2008. 351 с.
10. Дарчук Н. Комп'ютерне анотування українського тексту: результати і перспективи. К., 2013. 543 с.
11. Дарчук Н. Морфологічне анотування Корпусу української мови. *Комп'ютерна лінгвістика: сучасне і майбутнє. Матеріали Міжнародної науково-практичної конференції*. К., 2012. С. 16-19.
12. Карпіловська Є. Вступ до прикладної лінгвістики: Комп'ютерна лінгвістика. Донецьк, 2006. 188 с.
13. Марчук Ю. Компьютерная лингвистика. М., 2007. 317 с.

14. Міщенко Н. Система програм морфологічного аналізу науково-технічних текстів. *Наукові записки. Серія «Філологічні науки»*. Випуск 95(2). К., 2011. С. 538-542.
15. Морфологический анализ научного текста на ЭВМ. К., 1989. 262 с.
16. Николаев И., Митренина О., Ландо Т. Прикладная и компьютерная лингвистика. М., 2016. 315 с.
17. Ножов И. Морфологическая и синтаксическая обработка текста (модели и программы). М., 2003. 140 с.
18. Партико З. Прикладна і комп'ютерна лінгвістика: Вступ до спеціальності. Львів, 2008. 224 с.
19. Романюк Ю. Прикладна морфологія. Черкаси, 2009. 117 с.
20. Соснина Е. Введение в прикладную лингвистику. Ульяновск, 2012. 110 с.
21. Сучасна українська літературна мова. Морфологія. К., 1969. 250 с.
22. Antworth E. PC-KIMMO: a two-level processor for morphological analysis. *Occasional Publications in Academic Computing*. No. 16. Dallas, 1990. 273 p.

ТЕМА 5. АВТОМАТИЧНИЙ МОРФОЛОГІЧНИЙ АНАЛІЗ НА ОСНОВІ ОПЕРАЦІЇ ЛОГІЧНОГО МНОЖЕННЯ

Опорний конспект

1. Особливе місце серед різних видів аналізу посідає спосіб автоматичного морфологічного аналізу методом логічного множення, розроблений Сергієм Фітіаловим. За цим методом спочатку відбувається пошук слова у словнику основ. Якщо слова, які мають закінчення, не

знаходяться у словнику, тоді від кожного слова відкидається по одній графемі справа і пошук повторюється. При негативній відповіді відкидається наступна графема і т.д. Відкинуті графеми утворюють флексію і фіксуються. Кожна відкинута графема вважається елементарною одиницею морфологічного аналізу. Їй приписується бульовий вектор – сукупність нулів й одиниць компонентів цього вектора. Число компонентів даного вектора дорівнює числу граматичних категорій, які можуть бути виражені закінченням, частиною якого є дана графема. Оскільки попередньо відбувся пошук за словником основ і встановлено частину мови аналізованого слова, то є можливість однаковим графемам, які входять до складу флексій різних частин мови, приписувати різні вектори.

2. Наприклад, треба визначити, у якому числі і відмінку вживається словоформа *столом* (рос. мовою – бо автор методу апробував його на російськомовних текстах). Після пошуку у словнику встановлюється, що основа *стол* – іменник, графеми, які входять до складу закінчення, **о** і **м**. Графема **м** зустрічається серед графем закінчень іменника в орудному відмінку однини чоловічого і середнього роду, а також у давальному й орудному відмінку множини всіх трьох родів. Приписується графемі **м** такий бульовий вектор, у якому на місці компонентів, що відповідають відмінкам, у яких вони зустрічаються, стоять одиниці, а на місці інших компонентів – нулі. Такі ж дії відбуваються із другою графемою закінчення. Виконавши операцію логічного множення бульових векторів графем **о** і **м**, одержуємо в результуючому векторі одиницю на місці розряду тієї граматичної категорії, у флексії якої зустрічаються одночасно і графема **о**, і графема **м** (див. Табл. 13):

Таблиця 13. Бульові вектори закінчень

Графеми закінчень	Відмінок						Грам. знач. числа і роду
	Наз.	Род.	Дав.	Знах.	Орудн.	Місц.	
О	0	0	0	0	1	0	Однина, чоловічий рід
М	0	0	0	0	1	0	
О	0	0	0	0	1	0	Однина, жіночий рід
М	0	0	0	0	0	0	
О	1	0	0	1	1	0	Однина, середній рід
М	0	0	0	0	1	0	
О	0	0	0	0	0	0	Однина, жіночий рід
М	0	0	0	0	0	0	
О	0	0	0	0	0	0	Множина
М	0	0	1	0	0	0	

3. Незалежний аналіз відбувається без звертання до словника, лише за рахунок таблиць афіксів; це вивчення комбінаторики флексій та інших афіксів, ідея якого полягає у максимальному використанні інформації про флексію з урахуванням аломорфії та варіантності, щоб можна було б звести їх до однієї морфеми; результатом є укладання спеціальних словників (лексем, що не мають словозміни; службових слів; флексій, де кожна графема має вказівку на те, у які морфеми вона входить); алгоритм цього аналізу надає таку інформацію: вказівка на те, якою частиною мови є аналізована лексема; номер морфеми із граматичними характеристиками роду, числа, відмінка – для іменників, прикметників; роду, числа – для дієслів минулого часу; особи, числа – для дієслів теперішнього і майбутнього часу.

Контрольні питання

1. Опишіть суть процедури АМА методом логічного множення.
2. Дайте визначення бульовому вектору.
3. Опишіть процедуру незалежного АМА.

Домашнє завдання

Порівняйте АМА методом логічного множення з уже дослідженими методами (на основі графемного аналізу, на основі флективного аналізу, на основі словника словоформ, на основі словника основ). Укладіть перелік його переваг і перелік його недоліків у порівнянні з іншими.

Лабораторна робота №5

АВТОМАТИЧНИЙ МОРФОЛОГІЧНИЙ АНАЛІЗ НА ОСНОВІ ОПЕРАЦІЇ ЛОГІЧНОГО МНОЖЕННЯ

1. Користуючись ресурсом http://lcorp.ulif.org.ua/dictua/, визначте частину мови та основу слів: *син*, *вікно*, *парта*. Скопіюйте парадигму кожного слова.

2. У поданих нижче таблицях (див. Табл. 14, 15) пропишіть бульові вектори для флексій ***ом*** та ***ах***:

Таблиця 14. Бульові вектори для флексії ***ом***

Графеми флексій	**Відмінок**							**Число і рід**
	Н.	**Р.**	**Д.**	**З.**	**О.**	**М.**	**К.**	
О								**одн., ч.р.**
М								
О								**одн., ж.р.**
М								
О								**одн., с.р.**
М								
О								**мн.**
М								

Таблиця 15. Бульові вектори для флексії ***ах***

Графеми флексій	**Відмінок**							**Число і рід**
	Н.	**Р.**	**Д.**	**З.**	**О.**	**М.**	**К.**	
А								**одн., ч.р.**
Х								
А								**одн., ж.р.**
Х								
А								**одн., с.р.**
Х								

А								**мн.**
Х								

* *Дивлячись на всі три парадигми, знайдіть ті словоформи, у флексіях яких є графема **м**. Поставте одиничкі на перетині відповідних відмінка / числа і роду в першій таблиці. Решту жовтих клітинок заповніть нулями.*
* *Дивлячись на всі три парадигми, знайдіть ті словоформи, у флексіях яких є графема **о**. Поставте одиничкі на перетині відповідних відмінка / числа і роду в першій таблиці. Решту блакитних клітинок заповніть нулями.*
* *Дивлячись на всі три парадигми, знайдіть ті словоформи, у флексіях яких є графема **х**. Поставте одиничкі на перетині відповідних відмінка / числа і роду в другій таблиці. Решту зелених клітинок заповніть нулями.*
* *Дивлячись на всі три парадигми, знайдіть ті словоформи, у флексіях яких є графема **а**. Поставте одиничкі на перетині відповідних відмінка / числа і роду в другій таблиці. Решту рожевих клітинок заповніть нулями.*

3. Здійснивши операцію логічного множення бульових векторів, зробіть висновок, для яких словоформ (укажіть їх рід, число, відмінок) є характерними флексії ***ом*** та ***ах***.

Література до теми:

1. Бабина О. Корпусный метод автоматического морфологического анализа флективных языков. *Вестник Южно-Уральского государственного университета. Серия «Лингвистика»*. №25(284), выпуск 15. Челябинск, 2012. С. 38-44.
2. Баранов А. Введение в прикладную лингвистику. М., 2001. 360 с.
3. Белоногов Г. Компьютерная лингвистика и перспективные информационные технологии. М., 2004. 159 с.
4. Буньо Г. Сучасні методи вирішення проблеми граматичної омонімії в тексті. *Наукові записки. Серія: Філологічна*. 2014. Вип. 49. С. 12-16.

5. Волошин В. Комп’ютерна лінгвістика. Суми, 2004. 382 с.
6. Гельбух А., Сидоров Г. К вопросу об автоматическом морфологическом анализе флективных языков. Ел. режим доступу: www.dialog-21.ru/Archive/2005.
7. Грязнухіна Т., Нікула М. Система автоматичного морфологічного аналізу українського наукового тексту. Проблеми українізації комп’ютерів. Матеріали 2-ї міжнародної конференції. Київ, 1993. С. 42-46.
8. Данилюк І. Прикладна морфологія. Донецьк, 2010. 216 с.
9. Дарчук Н. Комп’ютерна лінгвістика. К., 2008. 351 с.
10. Дарчук Н. Комп’ютерне анотування українського тексту: результати і перспективи. К., 2013. 543 с.
11. Дарчук Н. Морфологічне анотування Корпусу української мови. *Комп’ютерна лінгвістика: сучасне і майбутнє. Матеріали Міжнародної науково-практичної конференції*. К., 2012. С. 16-19.
12. Карпіловська Є. Вступ до прикладної лінгвістики: Комп’ютерна лінгвістика. Донецьк, 2006. 188 с.
13. Марчук Ю. Компьютерная лингвистика. М., 2007. 317 с.
14. Міщенко Н. Система програм морфологічного аналізу науково-технічних текстів. *Наукові записки. Серія «Філологічні науки»*. Випуск 95(2). К., 2011. С. 538-542.
15. Морфологический анализ научного текста на ЭВМ. К., 1989. 262 с.
16. Николаев И., Митренина О., Ландо Т. Прикладная и компьютерная лингвистика. М., 2016. 315 с.
17. Ножов И. Морфологическая и синтаксическая обработка текста (модели и программы). М., 2003. 140 с.
18. Партико З. Прикладна і комп’ютерна лінгвістика: Вступ до спеціальності. Львів, 2008. 224 с.
19. Романюк Ю. Прикладна морфологія. Черкаси, 2009. 117 с.

20. Соснина Е. Введение в прикладную лингвистику. Ульяновск, 2012. 110 с.
21. Сучасна українська літературна мова. Морфологія. К., 1969. 250 с.
22. Antworth E. PC-KIMMO: a two-level processor for morphological analysis. *Occasional Publications in Academic Computing*. No. 16. Dallas, 1990. 273 p.

Змістовий модуль ІІІ

СУЧАСНІ АВТОМАТИЧНІ МОРФОЛОГІЧНІ АНАЛІЗАТОРИ

ТЕМА 1. АВТОМАТИЧНИЙ МОРФОЛОГІЧНИЙ АНАЛІЗАТОР LANGUAGETOOL

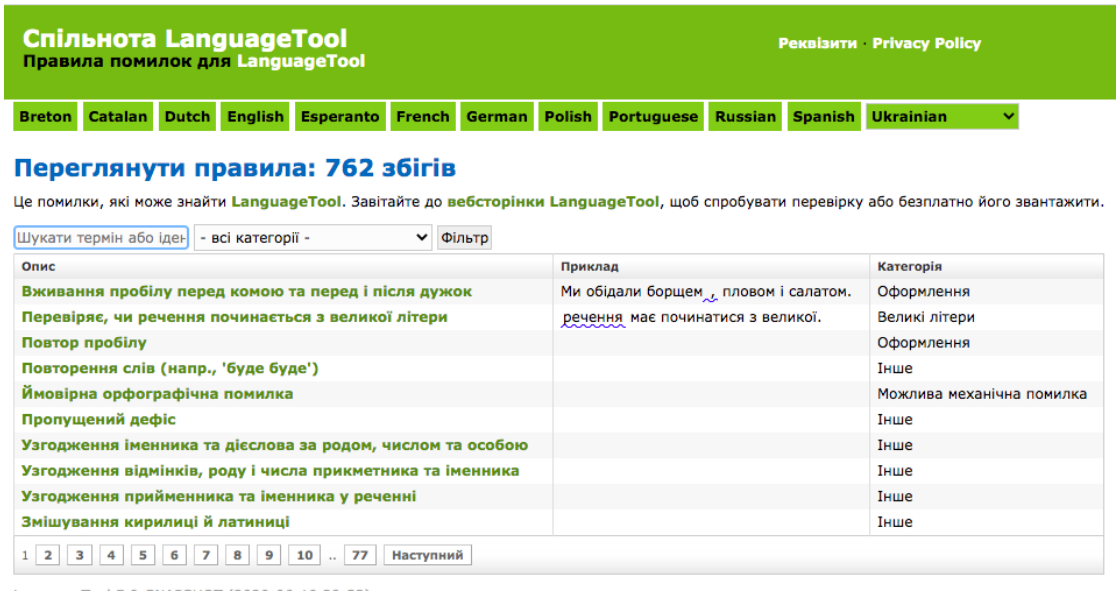
Опорний конспект

1. LanguageTool – відкритий програмний засіб перевірки граматики (див. Рис. 1):

Рисунок 1. Функціонал LanguageTool

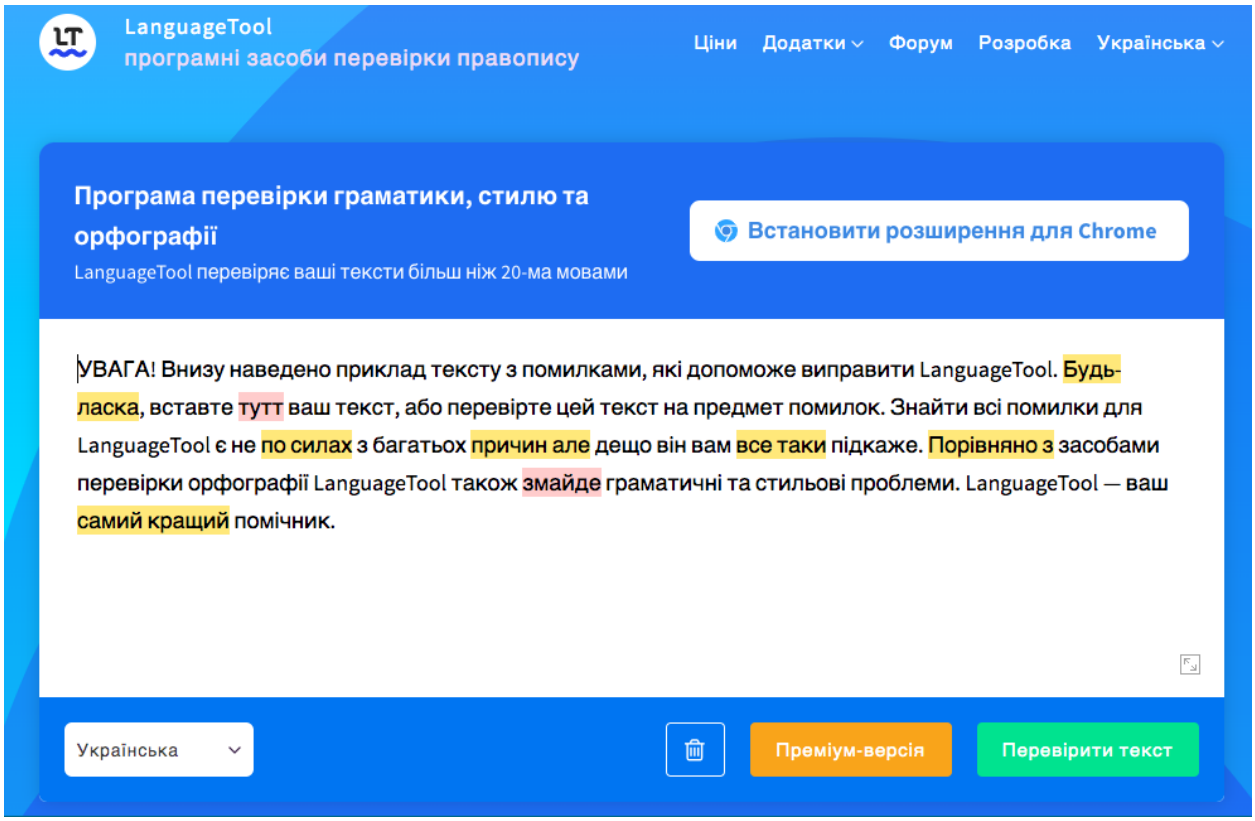
2. Аналізатор перевіряє тексти за допомогою правил. Кожне правило відповідає за ймовірну помилку в тексті (див. Рис. 2):

Рисунок 2. Правила LanguageTool



3. Результати внутрішнього аналізу LanguageTool дають змогу зрозуміти, на основі чого спрацьовують його правила (див. Рис. 3):

Рисунок 3. Інтерфейс LanguageTool



4. Редактор правил дає змогу створювати правила виявляння помилок (див. Рис. 4):

Рисунок 4. Інтерфейс для створення правил LanguageTool

Контрольні питання

1. Що таке LanguageTool?
2. Опишіть функціонал цього програмного засобу.
3. Яку роль відіграє морфологічний модуль у використанні LanguageTool?

Домашнє завдання

У редакторі граматичних правил LanguageTool (https://community.languagetool.org/ruleEditor2/index?lang=uk) запропонуйте свої варіанти правил (3-5 прикладів), пов'язаних з морфологією слова.

Лабораторна робота №1
АВТОМАТИЧНИЙ МОРФОЛОГІЧНИЙ АНАЛІЗАТОР LANGUAGETOOL

1. Проаналізуйте отриманий текст у автоматичному морфологічному аналізаторі LanguageTool https://community.languagetool.org/analysis/index?lang=uk (див. Рис. 5):

Варіант 1

У 2016 році в Сєвєродонецьку з'явився креативний простір «Пружина». Це неформальне об'єднання, яке прагне покращити простір міста з допомогою мистецьких виставок, лекцій, дискусій, фестивалів. У той же час у місті запрацював громадський майданчик «ХочуБуду». Мета цієї суспільної платформи – розвивати потенціал активних людей та сприяти втіленню громадських ініціатив. Майданчик працює у професійному, освітньому, культурному, спортивному та екологічному напрямах. А 2018 року на базі міської бібліотеки відкрили простір, обладнаний настільними іграми, аби згуртувати у такий спосіб місцеву молодь та розвивати їхні комунікативні навички. У Лисичанську на базі міського кінотеатру «Дружба» з 2017 року працює однойменний простір неформальної атлетичної культури. Це спортивний осередок місцевої молоді, де можуть тренуватися скейтери та BMX-ери. Тут проводять культурні фестивалі та діджейські вечірки. На

другому поверсі приміщення планують зробити коворкінг та простір для спільного користування, де відбуватимуться заняття із 3D-друку та за різноманітними навчальними програмами, діятимуть кіноклуб та лаунж-зона. Маріуполь, що на Приазов'ї, після 2014 року також отримав новий шанс на переосмислення і розвиток просторів. У місті з'явилося чимало переселенців, які почали активну громадську діяльність, відкриваючи різноманітні вільні простори. Одне з перших антикафе в Маріуполі – «Joy» – заснував переселенець Микола Петаєв. Цей проєкт почався зі створення квест-кімнати і надалі розширюється. Перший арт-простір нового формату «Тю» у місті заснувала переселенка з Донецька Діана Берг. «Тю» позиціонує себе як платформа культурних і соціальних ініціатив. Вільний простір «Халабуда» відкрився у Маріуполі 2016 року. Тут проводять зустрічі, навчальні та розважальні заходи. Також тут діє центр допомоги військовим, переселенцям і людям, що опинилися у скрутному становищі, працює коворкінг та бізнес-інкубатор для переселенців. У маріупольській водонапірній вежі місцевий активіст Владислав Зайцев створює проєкт «Vezha Creative Space». Тут проводять лекторії, планують відкрити IT-кластер, туристичний довідковий центр, а також оглядовий майданчик, звідки видно все місто.

Варіант 2

На початку 2017 року кінознавиця та активістка Наталка Сосницька разом з однодумцями створила у Костянтинівці вільний простір «Druzi». Це креативний простір для громадських ініціатив, навчання та відпочинку. Щотижня тут проходять різноманітні заходи: лекції, тренінги, майстер-класи, концерти, кінопокази, ігри, засідання клубів за інтересами, репетиції. Простір працює у напрямах неформальної освіти, урбаністики, культурних та соціальних проєктів, екологічних ініціатив. 2016 року соціальний підприємець Володимир Орос відкрив у Добропіллі арт-кафе «Тролейбус». Частина прибутків цього закладу йде на утримання молодіжного центру Добропілля. Концепція закладу полягає у тому, щоб люди приходили сюди не

тільки поїсти, а і з користю провести час та отримати нові знання. Заступниця директора бібліотеки-філії №1 у Добропіллі Ірина Козачук втілює на базі цього простору культурні проєкти для дітей і дорослих. Бібліотеку перетворюють на публічний простір, де є місце для проведення зустрічей і заходів, інтернет-центр, облаштований сквер тощо. Працівники бібліотеки долучаються до регіональних акцій і толок. Засновник громадської організації «Творці історії» Владислав Бурховецький допомагає облаштовувати у Добропіллі спортивні та культурні майданчики. ГО залучає молодь до їх будівництва, сприяє організації процесу і пошуку коштів. Восени 2018 року активісти облаштували у спальному районі міста публічний простір під назвою «Не пустир». На місці зарослого пустиря побудували локацію для дитячих ігор, кінопоказів, дружніх зустрічей. Наприкінці 2018 року у Добропіллі з'явився публічний еко-простір «Сквер 21». Зараз тут облаштований амфітеатр, у планах – додати зону відпочинку та сучасні вуличні меблі. Простір «Вільна Хата» в Краматорську працює у декількох напрямах: соціальному, освітньому, культурному, екологічному. Координатор простору Микола Дорохов розповідає, що до їхніх проєктів належать такі ініціативи, як «Добрий сусід» (волонтерська допомога людям, які цього потребують), неформальні освітні заходи, організація міських фестивалів, зустрічей з культурними діячами тощо.

Варіант 3

«Вільна Хата» – це платформа, де можна знайомитись з людьми та шукати підтримку для своїх ідей. Ми вже підтримали кілька десятків ініціатив у Краматорську і за містом. Також це простір для власного і суспільного розвитку. Сюди можна прийти навчитись чомусь, прийти з гуртом зіграти акустичний концерт. У нас проходять вечори поезії і вже сформувалась спільнота поетів. Микола Дорохов каже, що за чотири роки діяльності «Вільна Хата» вже вплинула на позитивні зміни у Краматорську та загалом в регіоні. Команда простору охоче ділилася досвідом із активістами із сусідніх міст, і згодом подібні громадські простори відкрилися

у Слов'янську, Бахмуті, Дружківці, Костянтинівці, Лимані. Історія ж самої «Вільної Хати» почалась у грудні 2014 року з волонтерського табору «Будуємо Україну Разом». «Тут були волонтери з Краматорська і зі всієї України. Ми ремонтували помешкання, які постраждали внаслідок бойових дій. Так сформувалась спільнота, яка хотіла далі щось робити. Відбудовувати не тільки помешкання, а й суспільство. Тому що ми вже бачили, до чого привели бездіяльність або байдужість. Не хочеться, щоб це повторювалось і повертало. Тому треба далі діяти і змінювати навколишнє середовище, тому що ми відповідальні не тільки за себе, але й за тих, хто поряд з нами. Якщо хочеш нормально жити, ну так зроби щось для цього. Ти можеш завжди кудись втікати, їздити, але все одно ти повертаєшся додому, і повертатись в якесь печальне місце не хотілося б». Протягом перших двох років простір «Вільна Хата» працював за ресурсної та інституційної підтримки Львівської освітньої фундації. На початку 2017 року команда простору вже заснувала громадську організацію. Проєкт також отримує донорську підтримку від американського фонду USAID.

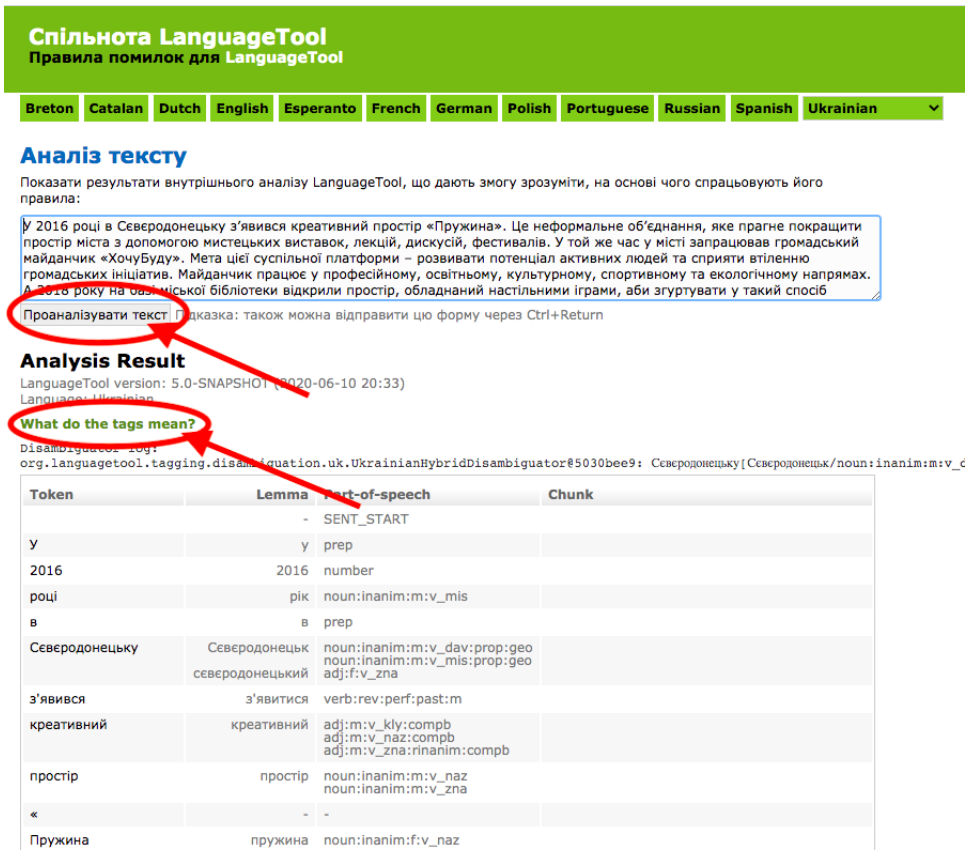


Рисунок 5. Автоматичний морфологічний аналізатор LanguageTool

2. Ознайомтеся зі значеннями використовуваних програмою тегів (див. Рис. 6):

161 lines (121 sloc) | 6.13 KB — Raw | Blame | History

```
Теги:

[КЛ] – ключ леми (тег, який розрізняє різні леми з омонімів)

noun    іменник
    [КЛ] anim       істота
    [КЛ] fname      ім'я
    [КЛ] lname      прізвище
    [КЛ] pname       по батькові
    [КЛ] inanim     неістота
    [КЛ] unanim     невизначена категорія істота/неістота (бактерія)
         prop       власна назва
         geo        топонім

verb    дієслово
    [КЛ] imperf недоконаний вид
    [КЛ] perf доконаний вид
    [КЛ] rev  зворотна форма (дієслова) (тег є неявним ключем, оскільки лема на –ся завжди відрізняється від прямого дієслова)

    inf інфінітив
    futr  майбутній час
    past  минулий час
    pres  теперішній час
    impr    наказова форма
    impers безособова форма

    1       1–а особа
    2       2–а особа
    3       3–а особа


adj     прикметник
    compb    базова форма
    compc    порівняльна форма
    comps    найвища форма
    short    короткі форми прикметників

    adjp    дієприкметник: (:&adjp – лише дієприкметник; :&&adjp – дієприкметник і прикметник)
        actv   активний
        pasv   пасивний
        imperf недоконаний вид
        perf   доконаний вид

    (past/pres є в коментарях сирців для більшості дієприкметників, але наразі не використовується)

adj/adjp:
    v_zna:rinanim   знахідний для неістот (лише ч.р.)
    v_zna:ranim     знахідний для істот (лише ч.р.)
    uncontr         нестягнені (не генеруються за уставою)

adv     прислівник
    compb    базова форма
    compc    порівняльна форма
    comps    найвища форма

advp    дієприслівник
    [КЛ] perf
    [КЛ] imperf

prep    прийменник

conj    сполучник
    subord підрядний
    coord сурядний
```

```
part    частка

intj    вигук

numr    числівник

foreign     невідмінювані запозичені слова невизначеної частини мови (Альгемайне, Юнайтед тощо)

noninfl     невідмінювані частини (най-най, брутто, екстра...)


Спільні для noun/adj/adjp:
    Відмінки:
        v_naz   називний
        v_rod   родовий
        v_dav   давальний
        v_zna   знахідний
        v_oru   орудний
        v_mis   місцевий
        v_kly   кличний
        nv    не відмінюється
        np    без множини (TODO: проставлено не всюди)
        ns    без однини (TODO: проставлено не всюди)



Спільні для noun/adj/adjp/verb
    p  множина
    s  однина

    Рід:
        m  чоловічий
        f  жіночий
        n  середній
```

```
Додаткові теги:

    abbr  абревіатура
    bad   покруч
    subst просторічна форма
    rare  рідковживане/діалектичне/застаріле
    coll  розмовне слово/розмовна форма
    slang сленг
    alt   альтернативне написання (не за чинним правописом)

    :xp[1-9] омоніми, що відрізняються парадигмою відмінювання (напр. бар - р.в. бару, бар - р.в. бара)
    # в коментарях також :xv[1-9] омоніми, що відрізняються семантично (напр. глупий (дурний, має вищий ступінь глупіший) і глу


    v-u   паралельні форми на в-/у- (для правил милозвучності, не генерується за уставою)


Додаткові теги класів слів (після &):
     &adjp - слова, що є дієприкметниками
     &&adjp - слова, що є і прикметниками і дієприкметниками
[КЛ] &pron - наразі всі займенники мають теги відповідних частин мови (noun/adj/adv), але всі мають додатковий тег &pron
        (тег &pron разом з наступним класифікатором стає ключем леми)
     &numr - слова, що є порядковими числівниками
     &&numr - слова, що є і іменниками і кількісними числівниками
     &insert - може бути вставним словом
     &predic - може бути предикативом
```

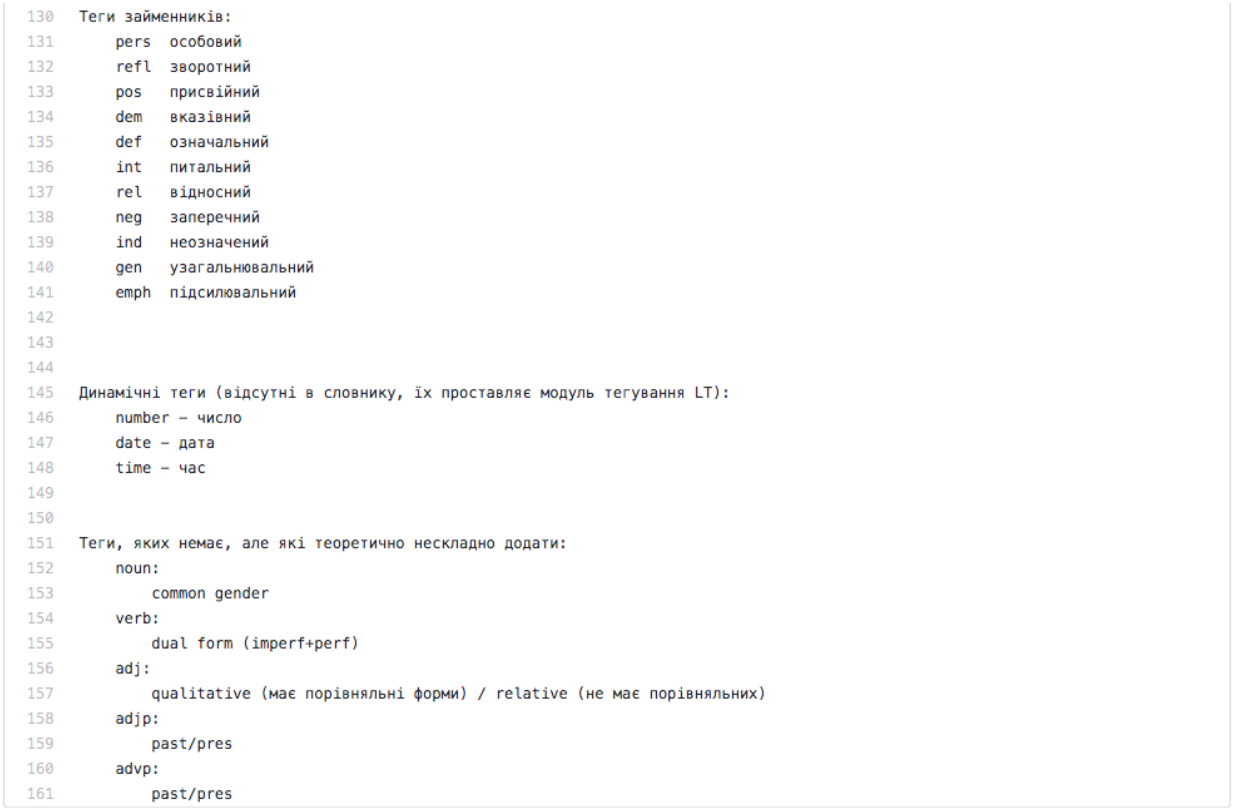


Рисунок 6. Значення тегів в автоматичному морфологічному аналізаторі LanguageTool

3. Позначте у результаті аналізу (таблиця видачі) граматичні омоніми. Як можна зняти цю омонімію? Запропонуйте свій варіант і оберіть правильну словоформу.

Література до теми:

1. Бабина О. Корпусный метод автоматического морфологического анализа флективных языков. *Вестник Южно-Уральского государственного университета. Серия «Лингвистика»*. №25(284), выпуск 15. Челябинск, 2012. С. 38-44.
2. Баранов А. Введение в прикладную лингвистику. М., 2001. 360 с.
3. Белоногов Г. Компьютерная лингвистика и перспективные информационные технологии. М., 2004. 159 с.
4. Буньо Г. Сучасні методи вирішення проблеми граматичної омонімії в тексті. *Наукові записки. Серія: Філологічна*. 2014. Вип. 49. С. 12-16.
5. Волошин В. Комп’ютерна лінгвістика. Суми, 2004. 382 с.

6. Гельбух А., Сидоров Г. К вопросу об автоматическом морфологическом анализе флективных языков. Ел. режим доступу: www.dialog-21.ru/Archive/2005.
7. Грязнухіна Т., Нікула М. Система автоматичного морфологічного аналізу українського наукового тексту. Проблеми українізації комп'ютерів. Матеріали 2-ї міжнародної конференції. Київ, 1993. С. 42-46.
8. Данилюк І. Прикладна морфологія. Донецьк, 2010. 216 с.
9. Дарчук Н. Комп'ютерна лінгвістика. К., 2008. 351 с.
10. Дарчук Н. Комп'ютерне анотування українського тексту: результати і перспективи. К., 2013. 543 с.
11. Дарчук Н. Морфологічне анотування Корпусу української мови. *Комп'ютерна лінгвістика: сучасне і майбутнє. Матеріали Міжнародної науково-практичної конференції*. К., 2012. С. 16-19.
12. Карпіловська Є. Вступ до прикладної лінгвістики: Комп'ютерна лінгвістика. Донецьк, 2006. 188 с.
13. Марчук Ю. Компьютерная лингвистика. М., 2007. 317 с.
14. Міщенко Н. Система програм морфологічного аналізу науково-технічних текстів. *Наукові записки. Серія «Філологічні науки»*. Випуск 95(2). К., 2011. С. 538-542.
15. Морфологический анализ научного текста на ЭВМ. К., 1989. 262 с.
16. Николаев И., Митренина О., Ландо Т. Прикладная и компьютерная лингвистика. М., 2016. 315 с.
17. Ножов И. Морфологическая и синтаксическая обработка текста (модели и программы). М., 2003. 140 с.
18. Партико З. Прикладна і комп'ютерна лінгвістика: Вступ до спеціальності. Львів, 2008. 224 с.
19. Романюк Ю. Прикладна морфологія. Черкаси, 2009. 117 с.
20. Соснина Е. Введение в прикладную лингвистику. Ульяновск, 2012. 110 с.

21. Сучасна українська літературна мова. Морфологія. К., 1969. 250 с.
22. Antworth E. PC-KIMMO: a two-level processor for morphological analysis. *Occasional Publications in Academic Computing*. No. 16. Dallas, 1990. 273 p.

ТЕМА 2. АВТОМАТИЧНИЙ МОРФОЛОГІЧНИЙ АНАЛІЗАТОР LIBMORPHUKR

Опорний конспект

1. Модуль морфологічного аналізу української мови libmorphukr побудований у 1997 році на основі технології, розробленої для російського морфологічного аналізатора, і містить 127107 унікальних основ слів, що покриває майже три мільйони граматично розрізнюваних форм слів. Продуктивність роботи аналізатора вимірюється сотнями тисяч слів у секунду в режимі ототожнення (лематизації), що є більш ніж достатнім для роботи в складі індексуючих і пошукових машин.

2. Модуль призначений для перевірки правопису окремих слів (правильно – неправильно), лематизації (побудови нормальних форм слів за довільною формою), вилучення граматичних описів тих форм, із якими збігся поданий рядок, морфологічного синтезу форм за нормальною формою і граматичним описом, також побудови списку можливих правильних накреслень для неправильного слова (підказка). Модуль реалізований у вигляді динамічної бібліотеки Win32, містить словники у вигляді ресурсів і експортує функції в С-стилі (тобто для їх виклику потрібна декларація extern “C”). Модуль працює з ASCIIZ-рядками, тобто рядками, де один символ кодується одним байтом, а нульовий символ позначає кінець рядка, і має на

увазі, що рядки на вхід йому подаються в кодуванні 1251 (Windows Cyrillic). Словник модуля обсягом близько 120000 слів при генерації розбивається на сторінки розміром менше 64К, що дозволяє оптимально організувати роботу з пам'яттю і використовувати коротке слово (16 біт) для адресації всередині сторінки з іншої. Алгоритм модуля гранично простий і являє собою інтерпретатор таблиць переходів кінцевого автомата, у той час як сторінки словника є цими самими таблицями. Завдяки цим рішенням словник модуля не перевищує розміром один мегабайт, стійко працює за обсягом доступної пам'яті близько 200К, а за наявності вільного мегабайта оперативної пам'яті завантажує всі сторінки і виходить на максимальну продуктивність не менше (а реально більше) 10000 слів за секунду (у режимі перевірки правопису, процесор Pentium 100). Продуктивність модуля в режимі лематизації (побудови текстів нормальних форм слова) дещо нижче, оскільки доводиться переглядати весь словник, і становить не менше 2000 слів у секунду (за тих же умов).

3. При побудові нормальних форм слова, так само як і при морфологічному синтезі, модуль морфологічного аналізу оперує частинами мови і граматичними описами форм. Частина мови і граматичний опис разом утворюють граматичну інформацію, яка описується структурою SGramInfo (див. Рис. 7):

```
typedef struct
{
  unsigned char  wInfo;
  unsigned char  iForm;
  unsigned short gInfo;
  unsigned char  other;
} SGramInfo;
```

Рисунок 7. Структура SGramInfo

У цій структурі wInfo – інформація про частину мови слова з деякими додатковими службовими полями, iForm – ідентифікатор форми слова, gInfo – розширений граматичний опис, other – додаткові ознаки.

4. Частина мови, яку видобувають із поля wInfo граматичної інформації, може бути використана, наприклад, для визначення ознаки значущості аналізованого слова або для інших цілей. Щоб визначити частину мови, слід замаскувати службові біти wInfo маскою 0x3F, після чого скористатися таблицею відповідностей (див. Табл. 1):

Таблиця 1. Таблиця відповідностей

wInfo & 0x3F	Мнемоника	Часть речи украинская
1	ч	Неодушевленное существительное мужского рода
2	чі	Одушевленное существительное мужского рода
3	ч/ж	Неодушевленное существительное общего рода
4	чі/жі	Одушевленное существительное общего рода
5	ч/с	Неодушевленное существительное мужского/среднего рода
6	чі/сі	Одушевленное существительное мужского/среднего рода
7	ж	Неодушевленное существительное женского рода
8	жі	Одушевленное существительное женского рода
9	ж/с	Неодушевленное существительное женского/среднего рода
10	с	Неодушевленное существительное среднего рода
11	сі	Одушевленное существительное среднего рода
12	мн	Неодушевленное существительное множественного числа
13	мні	Одушевленное существительное множественного числа
14	ч-ж	Неодушевленное существительное мужского рода, изменяющееся по схеме женского
15	чі-жі	Одушевленное существительное мужского рода, изменяющееся по схеме женского
16	п	Прилагательное
17	числ	Числительное
18	числ_2	Числительное “два” (имеющее род)
19	числ_п	Порядковое числительное
20	з	Личное местоимение (местоимение - существительное, например, “він” - он)
21	зп	Местоимение - прилагательное
22	ipf	Глагол несовершенного вида
24	pf	Глагол совершенного вида
26	pfipf	Двувидовой глагол
28	виг.	Междометье
29	прийм.	Предлог
30	присл.	Наречие
31	спол.	Союз
32	част.	Частица
33	незм.	Неизменяемое слово без указания части речи
34	вводн.	Вводное слово
35	аб.	Аббревиатура

5. Граматичний опис є точним двійковим ідентифікатором певної форми слова, але має властивість адитивності, тобто коротке слово (шістнадцять біт) розділене на зони, кожна з яких відповідає за певну граматичну ознаку. Нижче наведені мнемоніки з файлу прототипів зі значеннями і тлумаченнями кожного з них (див. Табл. 2):

Таблиця 2. Мнемоніки

Признак	Значение	Толкование	Встречается у
gfRetForms	0x8000	Признак возвратности	прилагательных, глаголов
gfFormMask	0x7000	Маска зоны указания падежа	глаголов, прилагательных, существительных, числительных, местоимений
gfNominative	0x0000	Именительный падеж	—"—
gfGenative	0x1000	Родительный падеж	—"—
gfDative	0x2000	Дательный падеж	—"—
gfAccusative	0x3000	Винительный падеж	—"—
gfInstrumental	0x4000	Творительный падеж	—"—
gfPrepositional	0x5000	Предложный падеж	—"—
gfCalling	0x6000	Звательный падеж	—"—
gfGendMask	0x0E00	Маска зоны рода и числа	—"—
gfMasculine	0x0200	Мужской род	—"—
gfFeminine	0x0400	Женский род	—"—
gfNewtral	0x0600	Средний род	—"—
gfMultiple	0x0800	Множественное число	—"—
gfVerbForm	0x0060	Маска зоны личностности	глаголов
vfPersonal	0x0000	Личная форма	—"—
vfActive	0x0020	Действительное причастие	—"—
vfPassiv	0x0040	Страдательное причастие	—"—
vfGerund	0x0060	Деепричастие	—"—
gfVerbFace	0x0018	Маска зоны лица	—"—
vbFirstFace	0x0008	Первое лицо	—"—
vbSecondFace	0x0010	Второе лицо	—"—
vbThirdFace	0x0018	Третье лицо	—"—
gfVerbTime	0x0007	Маска зоны времени	—"—
vtInfinitiv	0x0001	Инфинитив (неопределенная форма)	—"—
vtImperativ	0x0002	Императив (повелительное наклонение)	—"—
vtFuture	0x0003	Будущее время	—"—
vtPresent	0x0004	Настоящее время	—"—
vtPast	0x0005	Прошедшее время	—"—

6. До додаткових належать такі ознаки, як істота-неістота та ускладненість форми (див. Табл. 3):

Таблиця 3. Додаткові ознаки

Признак	Значение	Толкование
afAnimated	0x01	Данная форма прилагательного или причастия согласуется только с одушевленными существительными
afNotAlive	0x02	Данная форма прилагательного или причастия согласуется только с неодушевленными существительными
afHardForm	0x04	Затрудненная форма, отождествление которой возможно лишь в режиме распознавания затрудненных форм слов

7. Ідентифікатори форм (див. Табл. 4):

Таблиця 4. Ідентифікатори форм

Часть речи	FID	Толкование	
существительное	0	ед.	именительный падеж
	1		родительный падеж
	2		дательный падеж
	3		винительный падеж
	5		творительный падеж
	6		предложный падеж
	7		звательный падеж
	10	мн.	именительный падеж
	11		родительный падеж
	12		дательный падеж
	13		винительный падеж
	15		творительный падеж
	16		предложный падеж
	17		звательный падеж
прилагательное	0	муж.	именительный падеж
	1		родительный падеж
	2		дательный падеж
	3		винительный падеж, неодушевлённый
	4		винительный падеж, одушевлённый
	5		творительный падеж
	6		предложный падеж
	8-13	жен.	см. мужской род, формы 0 - 6
	16-22	ср.	
	24-30	мн.	

глагол	0	инфинитив	регулярная форма		
	1		возвратная форма		
	2	повелительное	невозвратный	ед.	
	3			мн.	
	4		возвратный	ед.	
	5			мн.	
	6	буд.	1-е лицо	невозвратная	ед.
	7				мн.
	8			возвратная	ед.
	9				мн.
	10		2-е лицо	невозвратная	ед.
	11				мн.
	12			возвратная	ед.
	13				мн.
	14		3-е лицо	невозвратная	ед.
	15				мн.
	16			возвратная	ед.
	17				мн.
	18	наст.	1-е лицо	невозвратная	ед.
	19				мн.
	20			возвратная	ед.
	21				мн.
	22		2-е лицо	невозвратная	ед.
	23				мн.
	24			возвратная	ед.
	25				мн.
	26		3-е лицо	невозвратная	ед.
	27				мн.
	28			возвратная	ед.
	29				мн.
	30-61		причастие	действительное	невозвратное
	62-93				возвратное
	94-125			страдательное	
	126		деепричастие	невозвратное	
	127			возвратное	
	128	прош.	невозвратные	муж.	
	129			жен.	
	130			ср.	
	131			мн.	
	132		возвратные	муж.	
	133			жен.	
	134			ср.	
	135			мн.	
	136-167		причастие	действительное	невозвратное
	168-199				возвратное
	200-231			страдательное	
	232		деепричастие	невозвратное	
	233			возвратное	

8. Коди помилок (див. Табл. 5):

Таблиця 5. Коди помилок

Код ошибки	Значение	Толкование
LEMMBUFF_FAILED	-1	При лемматизации переполнился массив нормальных форм
LIDSBUFF_FAILED	-2	При лемматизации переполнился массив идентификаторов лексем
GRAMBUFF_FAILED	-3	При лемматизации переполнился массив грамматических описаний
WORDBUFF_FAILED	-4	Слишком длинное исходное слово
PAGELOAD_FAILED	-5	Не удалась загрузка страницы словаря
PAGELOCK_FAILED	-6	

9. Налаштування (див. Табл. 6):

Таблиця 6. Налаштування

Настройка	Значение	Толкование
sfStopAfterFirst	0x0001	Достаточно одного отождествления
sfIgnoreCapitals	0x0002	Игнорировать корректность капитализации
sfHardForms	0x0004	Разрешать затрудненные словоформы

10. Перевірка правопису: функція перевіряє, чи є у словнику слово або його форма, тобто чи правильно воно написане. Функція повертає 0, якщо

слова не розпізнане, 1 – якщо слово написане правильно чи негативний код помилки (див. Рис. 8):

```
short MLMA_API EXPORT mlmaukCheckWord( const char* lpWord,
  unsigned short options );
```

Аргументы:

lpWord	– ASCIIZ-строка, украинское слово, правописание которого следует проверить;
options	– настройки морфологического анализатора.

Рисунок 8. Перевірка правопису

11. Будування списку нормальних форм: функція будує список нормальних форм поданого слова і вилучає ідентифікатори тих лексем і граматичні описи тих форм, із якими це слово збіглося. Повертає значення більше нуля – кількість побудованих лексем, якщо слово розпізнане, 0, якщо слово не розпізнане, або негативне значення – код помилки. Тексти нормальних форм відновлюються в мінімально допустимому ступені капіталізації, тобто якщо слово може бути написане усіма малими літерами, воно відновлюється усіма малими, якщо тільки з великої літери – то з великої, і т.д. Граматичні описи відновлюються у «плаваючому» форматі. Після ототожнення масив psGInfo містить кількість блоків граматичних описів, що дорівнює кількості нормальних форм. При цьому кожен блок граматичного опису на початку містить байт, що вказує кількість структур SGramInfo, що лежать відразу після нього (див. Рис. 9).

```
short MLMA_API EXPORT mlmaukLemmatize( const char* lpWord,
  unsigned short options, char* lpLemm, unsigned long* lpLIDs,
  char* lpGram, unsigned short ccLemm, unsigned short cdwLID,
  unsigned short cbGram );
```

lpWord	– ASCIIZ-строка, слово, которое следует проанализировать;
options	– настройки морфологического анализатора;
lpLemm	– указатель на массив, принимающий нормальные формы слов, или NULL, если тексты нормальных форм не требуются;
lpLIDs	– указатель на массив, принимающий идентификаторы лексем, или NULL, если идентификаторы лексем не требуются;
lpGram	– указатель на массив, принимающий грамматические описания, или NULL, если они не требуются;
ccLemm	– размерность массива lpLemm в байтах;
cdwLID	– размерность массива lpLIDs в двойных словах;
cbGram	– размерность массива lpGram в байтах.

Рисунок 9. Будування списку нормальних форм

12. Будування форми слова за граматичним описом: функція будує словоформу за її нормальною формою чи ідентифікатором лексеми та розширеним граматичним описом цієї форми. Повертає кількість побудованих

рядків у разі успіху, 0, якщо жодної форми побудувати не вдалося, або негативне значення – код помилки (див. Рис. 10):

```
short MLMA_API EXPORT mlmaukBuildFormGI( const char* lpWord,
  unsigned long dwLexID, unsigned short options,
  unsigned short grInfo, unsigned char bFlags,
  char* lpDest, unsigned short ccDest );
```

Аргументы:

lpWord	– ASCIIZ-строка, слово, которое следует проанализировать, или NULL, если задан идентификатор лексемы;
dwLexId	– идентификатор лексемы слова или 0, если построение формы идет по ключу – строке;
options	– настройки морфологического анализатора;
grInfo	– расширенное грамматическое описание требуемой формы слова;
bflags	– дополнительные грамматические флажки;
lpDest	– указатель на массив, принимающий построенные формы;
ccDest	– размерность массива lpDest в байтах.

Рисунок 10. Будування форми слова за граматичним описом

13. Перебір лексем: функція перебору всіх лексем словника (див. Рис. 11):

```
short MLMA_API EXPORT mlmaukEnumWords( TEnumWords enumproc, void *lpv );
```

Аргументы:

enumproc	– адрес функции обработки лексемы;
lpv	– пользовательский параметр, передаваемый функции enumproc.

Рисунок 11. Функція перебору всіх лексем словника

Прототип callback-функції перебору лексем: функція викликається функцією перебору лексем словника і отримує по черзі всі лексеми. Повинна повертати ненульове значення для продовження перебору або 0, якщо перебір лексем слід закінчити. Параметр lpv не використовується аналізатором і транслюється безпосередньо (див. Рис. 12):

```
typedef short (MLMA_API* TEnumWords)( unsigned long lid, void* lpv );
```

Аргументы:

lid	– идентификатор лексемы;
lpv	– параметр, переданный пользователем функции mlmaukEnumWords.

Рисунок 12. Прототип callback-функції перебору лексем

14. Ідентифікація частини мови лексеми: функція вилучення частини мови за ідентифікатором лексеми. Повертає нульове значення в разі успіху, 0,

якщо такої лексеми немає, або негативне значення – код помилки (див. Рис. 13):

```
short MLMA_API EXPORT mlmaukGetWordInfo( unsigned long dwLexID, unsigned char* wdinfo );
```

Аргументы:

dwLexId – идентификатор лексемы слова;
wdinfo – указатель на байт, получающий часть речи.

Рисунок 13. Ідентифікація частини мови лексеми

15. Абетка, використовувана в аналізаторі libmorphukr (див. Табл. 7):

Таблиця 7. Українська абетка в аналізаторі libmorphukr

	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
20								'						-		
30																
40										І						
50																
60	`									і						
70																
80																
90																
A0						Ґ					Є					Ї
B0			І	і	ґ											
C0	А	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П
D0	Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ		Ь		Ю	Я
E0	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п
F0	р	с	т	у	ф	х	ц	ч	ш	щ	ъ		ь		ю	я

Контрольні питання

1. Що таке libmorphukr?
2. Опишіть повний функціонал цього програмного засобу.

Домашнє завдання

Укладіть аналітико-порівняльну таблицю *«Функційні можливості автоматичних морфологічних аналізаторів LanguageTool та libmorphukr»*.

Лабораторна робота №2
АВТОМАТИЧНИЙ МОРФОЛОГІЧНИЙ АНАЛІЗАТОР LIBMORPHUKR

1. Проаналізуйте усі словоформи з отриманого тексту у автоматичному морфологічному аналізаторі libmorphukr http://www.keva.ru/?cat=ling-morph-dem (див. Рис. 14):

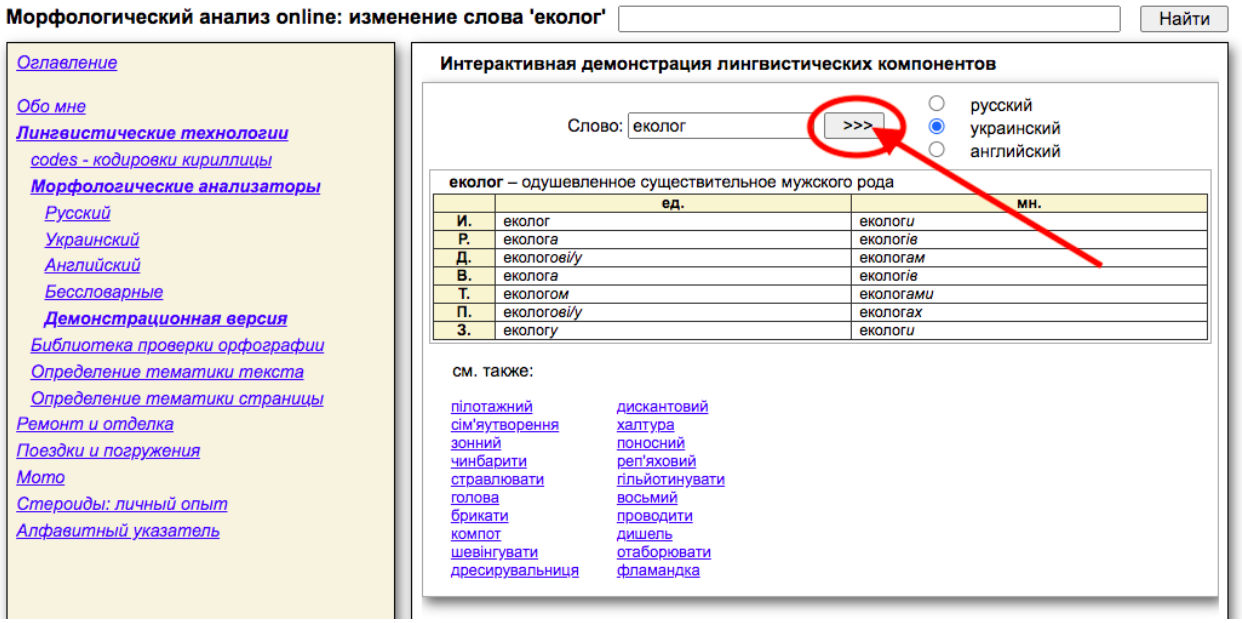


Рисунок 14. Інтерфейс автоматичного морфологічного аналізатора libmorphukr

Варіант 1

Екобудівник і практичний еколог Андрій Бобровицький зміг перетворити чотири гектари землі в селі Юшки, що на Наддніпрянщині, в освітній екоцентр. Тут зовсім відмовилися від традиційних способів обробітку ґрунту, розвивають пермакультуру. Андрій запрошує сюди всіх охочих не тільки відпочити від міського життя, а й спробувати себе в прикладній екології. Екоцентр «Юшки», започаткований Андрієм Бобровицьким, – відкритий простір, де втілюють екоініціативи й ідеї сталого розвитку. Сюди з'їжджаються люди з різних куточків країни, щоб навчитися пасічництву, землеробству, садівництву, пермакультурі й народним ремеслам. Тут можна освоїти екологічні технології будівництва, коли для зведення споруд використовують лише природні місцеві матеріали, як-от: глину, очерет, пісок чи солому. Зазвичай гості залишаються лише на вихідні, однак є й ті, які допомагають постійно. Андрій каже, що центр виник передусім через його давнє бажання жити просто й водночас добре. Він вважає, що екологія – найважливіша наука, основа всього, бо вона – про взаємозв'язки в довкіллі. За словами Андрія, раніше на екологічний стан так не зважали, думали про нього як про щось несуттєве, але зараз усе

змінюється: «Тепер нам просто це (екологія. – ред.) необхідно, щоб принаймні як вид залишитися на планеті».

Варіант 2

Серйозно цікавитися практичною екологією він почав, працюючи на будівництві. Тоді чоловік шукав способи зробити будівництво безпечнішим для довкілля. Так він перейшов на використання натуральних матеріалів. Ще у 2012 році Андрій разом із іншими ентузіастами збиралися в київському Клубі прикладної екології, де навчалися гончарству, бджільництву й плетінню з соломи. Окрім цього, учасники клубу обговорювали життя спільнот за принципами сталого розвитку, тобто раціональне й відповідальне користування всіма природними ресурсами задля їх збереження майбутнім поколінням. Через якийсь час команда відчула, що не варто вивчати сільське господарство в комфортному столичному офісі, це їх більше не влаштовує. Андрій каже, що під час кожної зустрічі клубу вони почувалися невпевнено, наче їхні знання далекі від реальності. Це підштовхнуло команду шукати земельну ділянку для екоцентру, де б вони могли практикуватися й експериментувати. Андрій їздив Україною, щоби знайти омріяне, ідеальне місце для екоцентру. Така знайшлася в селі Юшки: місцевість вразила чоловіка своїм біологічним різноманіттям і чистою водою. Річ у тім, що екоцентр оточений хвойними й акаціевими лісами, луками й пагорбами, є тут і болото. При цьому поблизу немає жодного фермерського господарства чи ремонтної бази, які б порушували тутешню гармонію. На думку екоактивіста, виняткова цінність цієї території – рослинне багатство й особливий клімат.

Варіант 3

Андрій вирішив, що форма організації його екоцентру буде такою, щоб долучитися міг кожен охочий, самостійно вирішивши, чим займатиметься. Каже, що в світі це поширена практика, коли люди, допомагаючи одне одному по господарству й обмінюючись досвідом, налагоджують довірчі стосунки. Засновник екоцентру впевнений, що безкорисливе, щире бажання

допомагати – притаманна українцям риса, тому вважає, що розвивати цю здатність взаємодіяти було би корисно всім. Андрій – один із тих небагатьох людей, які практикують пермакультуру, – стале (від англ. permanent agriculture – постійне землеробство) сільське господарство. Воно передбачає створення господарства, що не шкодить довкіллю й одночасно є економічно вигідним. Свої знання й навички він застосовує на земельній ділянці, яку називає «грядка розуму». Тут є чимало речей, здатних здивувати звичайного землероба: «Вона продумана таким чином, щоб не треба було її поливати, щоб вона сама себе зволожувала». Чоловік пояснює, що це відбувається завдяки формі грядки: вона має рівчак і два пагорби з боків. У рівчак закидають мульчу – суміш соломи й перегною. Так у рівчаку конденсується волога, тож рослини, які ростуть на пагорбах, мають її вдосталь. Мульчування захищає культури не тільки від пересихання, а й від перегрівання. А ще на цій грядці немає звичних рівних рядів якоїсь однієї культури – там можуть одночасно співіснувати різні види рослин. Річ у тім, що Андрій досліджує, як рослини співіснують й взаємодіють, самостійно підтримуючи баланс екосистеми. Він спостерігає за процесами, не втручаючись у них: не оре, не перекопує, не сапає грядку. У цьому й суть пермакультури: створення цілісної, природної інфраструктури з урахуванням всіх зв'язків між елементами. Планування дизайну зі збереженням природних форм потрібне для того, щоб екосистема балансувала, а не виснажувалася.

2. Заскріньте результати опрацювання, заархівуйте та надішліть викладачеві.

Література до теми:

1. Бабина О. Корпусный метод автоматического морфологического анализа флективных языков. *Вестник Южно-Уральского государственного университета. Серия «Лингвистика»*. №25(284), выпуск 15. Челябинск, 2012. С. 38-44.
2. Баранов А. Введение в прикладную лингвистику. М., 2001. 360 с.

3. Белоногов Г. Компьютерная лингвистика и перспективные информационные технологии. М., 2004. 159 с.
4. Буньо Г. Сучасні методи вирішення проблеми граматичної омонімії в тексті. *Наукові записки. Серія: Філологічна*. 2014. Вип. 49. С. 12-16.
5. Волошин В. Комп'ютерна лінгвістика. Суми, 2004. 382 с.
6. Гельбух А., Сидоров Г. К вопросу об автоматическом морфологическом анализе флективных языков. Ел. режим доступу: www.dialog-21.ru/Archive/2005.
7. Грязнухіна Т., Нікула М. Система автоматичного морфологічного аналізу українського наукового тексту. Проблеми українізації комп'ютерів. Матеріали 2-ї міжнародної конференції. Київ, 1993. С. 42-46.
8. Данилюк І. Прикладна морфологія. Донецьк, 2010. 216 с.
9. Дарчук Н. Комп'ютерна лінгвістика. К., 2008. 351 с.
10. Дарчук Н. Комп'ютерне анотування українського тексту: результати і перспективи. К., 2013. 543 с.
11. Дарчук Н. Морфологічне анотування Корпусу української мови. *Комп'ютерна лінгвістика: сучасне і майбутнє. Матеріали Міжнародної науково-практичної конференції*. К., 2012. С. 16-19.
12. Карпіловська Є. Вступ до прикладної лінгвістики: Комп'ютерна лінгвістика. Донецьк, 2006. 188 с.
13. Марчук Ю. Компьютерная лингвистика. М., 2007. 317 с.
14. Міщенко Н. Система програм морфологічного аналізу науково-технічних текстів. *Наукові записки. Серія «Філологічні науки»*. Випуск 95(2). К., 2011. С. 538-542.
15. Морфологический анализ научного текста на ЭВМ. К., 1989. 262 с.
16. Николаев И., Митренина О., Ландо Т. Прикладная и компьютерная лингвистика. М., 2016. 315 с.
17. Ножов И. Морфологическая и синтаксическая обработка текста (модели и программы). М., 2003. 140 с.

18. Партико З. Прикладна і комп'ютерна лінгвістика: Вступ до спеціальності. Львів, 2008. 224 с.
19. Романюк Ю. Прикладна морфологія. Черкаси, 2009. 117 с.
20. Соснина Е. Введение в прикладную лингвистику. Ульяновск, 2012. 110 с.
21. Сучасна українська літературна мова. Морфологія. К., 1969. 250 с.
22. Antworth E. PC-KIMMO: a two-level processor for morphological analysis. *Occasional Publications in Academic Computing*. No. 16. Dallas, 1990. 273 p.

ТЕМА 3. АВТОМАТИЧНИЙ МОРФОЛОГІЧНИЙ АНАЛІЗАТОР АОТ

Опорний конспект

1. Системи морфологічного аналізу і синтезу розвиваються вже не одне десятиліття, і серйозна обробка тексту вже немислима без їх допомоги. Як в Україні й Росії, так і за кордоном на ринку існує багато комерційних програм, які можуть успішно справлятися з цими завданнями, але, на жаль, вони не можуть бути використані для наукових експериментів через їх надто високу вартість і відсутність вихідного коду. З іншого боку, існують безкоштовні модулі, які, втім, часто неприйнятні через низьку швидкість обробки слів і неповноту словникових баз. Морфологічні модулі сайту www.aot.ru покликані вирішити зазначену вище проблему, забезпечивши наукові колективи і взагалі будь-яких можливих ентузіастів-експериментаторів системою морфологічного аналізу і синтезу, яка: вже володіє словниками досить великого обсягу, поповнюється добровольцями, тому не повинна в майбутньому застарівати; при пошуку в словнику використовує кінцевий

автомат, що дозволяє знаходити слово за лінійний від його довжини час (дуже швидко); написана на C++, компілюється під Linux і під Windows; володіє розвиненою системою додавання нових слів; має в розпорядженні російський, німецький та англійський лексикон; поширюється безкоштовно під ліцензією LGPL у початкових кодах. Усі зазначені вище властивості окремо можна зустріти в існуючих модулях морфологічного аналізу, однак саме це сполучення властивостей становить новизну й актуальність представленої системи.

2. Морфологічний словник, або лексикон, містить усі словоформи однієї мови, в нашому випадку: англійської, німецької або російської. Структуру словника найпростіше представити у вигляді реляційної схеми (див. Рис. 15):

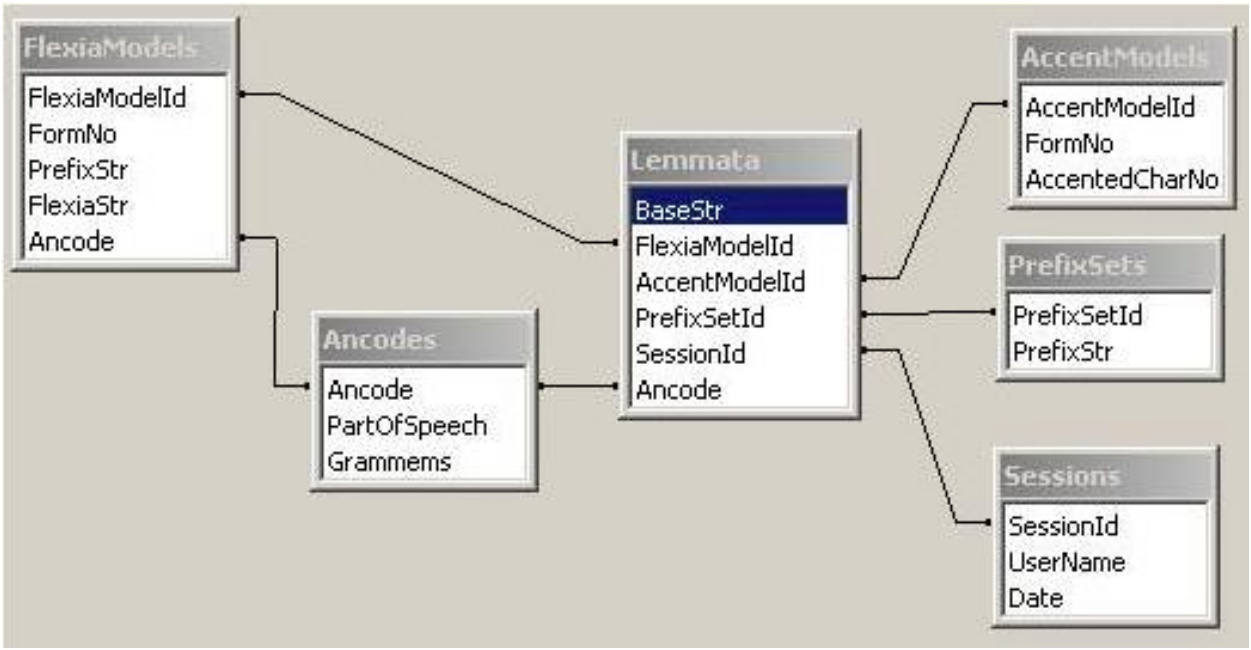


Рисунок 15. Структура схема побудови словника

Таблиця Lemmata містить перелік усіх лем даного словника, для кожної леми подані її властивості: псевдооснова слова (спільний для всіх словоформ даного слова підрядок) (поле BaseStr); посилання на набір закінчень (поле FlexiaModelId); посилання на набір наголосів (поле AccentModelId); посилання на набір префіксів (поле PrefixSetId); посилання на призначену для користувача сесію, у якій була внесена остання зміна цього запису (поле SessionId); посилання на спільні грамеми даної леми (поле Ancode) (може бути порожнім). Спільні грамеми даної леми – це ті грамеми, які повинні бути

приписані всім словоформам даної леми, наприклад, грамема «фам» (*фамилия*), або грамема «лок» (*локативность*). Часто це вже семантизовані грамеми. Набір префіксів леми – це ті префікси, з якими лема утворює повне слово. У набір префіксів може входити порожній префікс, що означає, що лема може бути використана сама по собі (без префіксів). Таблиця FlexiaModels містить перелік можливих закінчень усіх лем. Унікальним ключем тут є поля FlexiaModelId і FormNo. Поле FormNo містить порядковий номер закінчення в даному наборі закінчень, відповідно, FormNo не перевищує максимальну кількість словоформ в одній парадигмі. Далі: 1. Поле PrefixStr містить префікс даної словоформи (можливо, порожній). 2. Поле FlexiaStr містить закінчення даної словоформи (можливо, порожнє). 3. Поле Ancode містить морфологічну інтерпретацію даної словоформи. Нехай у нас є запис Q з таблиці Lemmata. Нехай P – один із її можливих префіксів, взятих за полем Q.PrefixSetId. Для того, щоб отримати i-ю словоформу даної леми, треба знайти в таблиці FlexiaModels запис R, такий, що Q.FlexiaModelId=R.FlexiaModelId і R.FormNo=i, тоді i-я словоформа буде дорівнювати: P+R.PrefixStr+Q.BaseStr+R.FlexiaStr. Таблиця AccentModels містить перелік можливих номерів наголошених голосних для словоформ. Унікальним ключем є поля AccentModelId і FormNo. Поле FormNo виконує таку ж роль, що і в таблиці FlexiaModels. Поле AccentedCharNo містить номер наголошеної голосної з кінця слова. Для кожної словоформи у словнику має бути вказаний наголос, якщо наголосу немає, тоді використовується спеціальна константа. Таблиця Ancodes містить усі можливі морфологічні інтерпретації. Ключем є поле Ancode («аношкінський код»). Поле PartOfSpeech містить частину мови (С,Г,П,...), а поле Grammems набір грамем, типу «мр,но,ед,им». Вищеописана схема показує принципові можливості та обмеження структури одного словника. Видно, що словник може зберігати інформацію про слова, можливі закінчення, можливі префікси, які можуть приєднуватися або до окремих словоформ, або до всіх словоформам даної парадигми. Словник зберігає ще інформацію про наголоси. Однак очевидно, що пропонована схема не

призначена для зберігання повного морфологічного розбору слова, пристроя компаундів і т.д. Необхідно ще відзначити, що в C++ реалізації не використовується реляційна база даних, однак на етапі редагування словника C++ структури фактично повністю повторюють вищеописану схему. Основні характеристики словників поточної версії (див. Табл. 8):

Таблиця 8. Основні характеристики словників поточної версії

Язык	Кол-во лемм	Кол-во наборов окончаний
Русский[1]	162519	2553
Немецкий[2]	212560	1171
Английский[3]	104657	442

3. Морфологічний словник у поточній версії може існувати у двох варіантах: 1. Варіант, призначений для редагування, який слідує реляційній схемі, зазначеній вище. 2. «Бінарний» варіант, призначений для обробки тексту, побудований на кінцевому автоматі. Оболонка для редагування (MorphWizard) використовує перший варіант словника. Основними функціями оболонки є: 1) пошук у словнику за лемою, словоформою, морфологічною інтерпретацією; 2) редагування однієї парадигми слова в т.зв. slf-форматі; 3) додавання нового слова, з використанням «передбачення» за «лемою»; видалення слова; 4) порівняння двох наборів закінчень, приписування набору закінчень цілій безлічі лем; 5) експорт у текстовий файл та імпорт з текстового файлу (в slf-форматі). Пошук за лемою, словоформою і морфологічною інтерпретацією здійснюється з використанням таблиці Lemmata і MorphModels. Тут, крім простого пошуку, користувачеві надана можливість використання регулярних виразів, наприклад, пошук за словоформою /^при.*ять$/ знайде всі слова, у яких є словоформи, що починаються з префікса «при» та закінчуються на «ять». Редагування однієї парадигми здійснюється у вікні текстового редактора. Парадигма представлена в т.зв. slf-форматі, тобто наступним чином (див. Рис. 16). На кожному рядку спочатку стоїть словоформа, а праворуч від словоформи стоять морфологічні характеристики:

Рисунок 16. Парадигма у slf-форматі

Словоформа в першому рядку оголошується лемою слова. Якщо у словоформи є префікс, він має бути відокремлений спеціальним символом «|», наприклад (див. Рис. 17):

Рисунок 17. Відокремлення префіксів

Наголос ставиться за допомогою апострофа. Основа парадигми – це незмінна ліва частина всіх словоформ, якщо відкинути можливі префікси словоформ. Додавання нового слова може бути здійснене принаймні трьома способами: 1) написання з нуля в вікні редагування; 2) вибір для нової леми набору закінчень за вже існуючою лемою, зазначеною користувачем; 3) використання «передбачення» за «лемою». Перший спосіб застосовується, коли треба ввести абсолютно нову парадигму слова. Другий – якщо користувач упевнений, що нова лема відмінюється так само, як інша, вже існуюча, знайома йому лема. Третій, коли користувач хоче вибрати підходящий варіант з можливих, найбільш частотних наборів закінчень. Тоді користувач вводить лему і отримує по закінченню введеної леми можливі набори закінчень, представлені існуючими лемами. Результати можна відсортувати за частотою або морфологічною інтерпретацією. Частину парадигм або весь словник можна вивести в текстовий файл, де для кожної леми даються вся інформація і сама парадигма в slf-форматі. Можливий так само імпорт з текстового файлу.

4. Словник у бінарному форматі надає наступні функції: 1) морфологічний аналіз: отримання за словоформою леми, її властивостей, унікального ID леми, морфологічних характеристик вхідної словоформи; 2) морфологічний синтез: отримання за унікальним ID леми всієї парадигми слова з усіма словоформами та їх морфологічними характеристиками. Важливо, що бінарне представлення словника оптимізоване насамперед для

проведення морфологічного аналізу. Основу цього подання становить кінцевий автомат (аксептор). Автомат детермінований і не має циклів, що дозволяє мінімізувати його у процесі побудови. Основний цикл побудови автомата виглядає так (див. Рис. 18):

```
For all  word forms W
begin
        AddStringToAutomat (W + '|'+Annot (W) );
End
```

Рисунок 18. Основний цикл побудови автомата

Символ «|» (annotation char) – спеціальний розділовий символ, якого немає в алфавіті словника, тобто він не може зустрічатися у словоформі W. Функція Annot(W) видає рядок анотації словоформи W у будь-якому текстовому вигляді, наприклад для словоформи *мама*: С жр,сд,им,од. Функція AddStringToAutomat додає вхідний рядок в автомат, зберігаючи властивість мінімальності і детермінованості автомата. Пошук словоформи в такому автоматі відбувається за лінійний від довжини вхідної словоформи час: досить просто пройти всі стани автомата, які відповідають символам вхідної словоформи, далі пройти розділовий символ і отримати обходом по графу всі анотації словоформи. Основна проблема полягає у змісті анотації. Найпростіше було б покласти в анотацію унікальний номер (ID) леми і номер словоформи в парадигмі слова. Цієї інформації достатньо, щоб обчислити всю решту необхідної інформації за константний час. Але тоді в автоматі сильно виросте кількість станів і зв'язків, однак не настільки фатально, щоб не робити цього, якщо швидкість обробки дуже важлива. Якщо, наприклад, автомат повинен видавати тільки лему, то треба включити в анотацію довжину закінчення вхідної словоформи і закінчення леми, яке треба додати справа до основи. В такому автоматі число станів буде невелике, і лему він буде видавати максимально швидко. У будь-якому випадку, зміст анотації може залежати від поставленого завдання і заданих параметрів. У поточній версії в анотації зберігаються три числа: номер набору закінчень, номер словоформи в

парадигмі, номер префікса леми. Ця інструкція дозволяє за константний час обчислити лему і морфологічні властивості вхідної словоформи, але, наприклад, для отримання інформації про наголос потрібен додатковий бінарний пошук в числовому векторі. Нижче наведені швидкісні характеристики програми, що породжує бінарне представлення (див. Табл. 9):

Таблиця 9. Швидкісні характеристики програми

Порождение автомата					
Язык	Кол-во состояний автомата	Кол-во переходов автомата	Время порождения	Размер автомата	Размер автомата и всей остальной морф. инф.
Русский	392443	815071	62 сек.	4,7 Мб	9 Мб
Немецкий	335069	598395	26 сек	3,6 Мб	9 Мб
Английский	79102	179394	5 сек	1 Мб	3 Мб

Розмір автомата і час його породження насамперед залежать від змісту анотації. Наприклад, автомат для російської мови, який може розпізнавати тільки лему, буде містити вдвічі менше число станів і в два рази швидше породжуватися. Нижче наведені швидкісні характеристики самого морфологічного аналізу (див. Табл. 10):

Таблиця 10. Швидкісні характеристики морфологічного аналізу

Скорость автомата		
Язык	Выдача леммы и морф. интерпретации словоформы	Выдача всей морф. информации
Русский	360 тыс. слов в сек.	202 тыс. слов в сек.
Немецкий	340 тыс. слов в сек.	196 тыс. слов в сек.

Перший стовпчик дає швидкість, коли вся необхідна інформація читається з анотацій, записаних у кінцевому автоматі, другий стовпчик – коли ми повинні ще використовувати анотацію для отримання додаткової інформації, наприклад, наголосів. У принципі, анотацію можна побудувати так, щоб уся інформація шукалася за константний час.

5. Морфологічне «передбачення» працює в тому випадку, якщо слово не було знайдене у словнику. Першим кроком «передбачення» є спроба знайти існуючу словоформу мови, яка максимально збігалася б справа з вхідним словом. Якщо розмір правої (невпізнаної) частини слова не перевищує певної

межі (у поточній версії це 5 символів), а розмір залишку (який співпав із якоюсь словоформою) не менший від 4 символів, тоді слово «передбачається» за знайденим правим залишком. Це повинно працювати для слів, до яких були додані продуктивні префікси, типу *квази-*, *мета-* і т.д. Пошук здійснюється послідовним відсіканням символів зліва і подачею «урізаного» слова в морфологічний аналіз. Якщо слово не можна знайти таким способом, вступає в дію «передбачення» за закінченням. Для цього був спеціально створений інший кінцевий автомат, побудований на рядках такого вигляду: ReverseSuffix(X)|Annot(X), де X якась словоформа словника, Annot(X) – анотація словоформи X, функція ReverseSuffix(X) повертає перевернуте зліва направо закінчення словоформи X певної заданої довжини (у поточній версії – 5). Крім цього, у цей автомат потрапляють тільки ті рядки, для яких частота зустрічності ReverseSuffix(X) у словнику перевищує певну межу (у поточній версії – 3). Числові параметри кінцевого автомата «передбачення» можуть бути задані в командному рядку програми генерації словника. Є ще одне важливе обмеження автомата «передбачення»: т.зв. факторизация за частиною мови. Для кожної мови вказані ті частини мови, які можуть бути продуктивними. Для російської – іменник, дієслово, прислівник і прикметник. Якщо зустрічається закінчення, для якого можливі різні інтерпретації всередині однієї продуктивної частини мови, тоді в автомат додається тільки та, що містить набір закінчень, який найбільш частотний у словнику. Таким чином, в автоматі для кожного закінчення і для кожної продуктивної частини мови міститься тільки одна морфологічна інтерпретація, причому найбільш продуктивна. Пошук у такому автоматі здійснюється наступним чином. Йдемо з кінця слова по автомату до тих пір, поки існує стан, у який можна перейти, використовуючи поточний символ слова. Далі обходом графа збираємо всі досяжні анотації. Якщо слово збіглося в повному обсязі з одним із закінчень, то можливо, що список анотацій містить кілька інтерпретацій всередині однієї частини мови, тоді доводиться знову вибирати найбільш продуктивну анотацію, використовуючи частотність набору закінчень. Якщо слово не було

передбачене як іменник, тоді у список можливих інтерпретацій додається варіант інтерпретації як незмінного іменника у всіх родах і числах (оскільки незнайдені слова найчастіше іменники). Таким чином, у кінці виходить набір з анотацій, число яких не більше числа продуктивних частин мови і який обов'язково містить варіант інтерпретації іменником. Загальна швидкість «передбачення» (обидві процедури) в два рази нижча швидкості основного пошуку словоформ в словнику, але це не настільки суттєво, оскільки число незнайдених слів у нормальних текстах рідко перевищує 5%. Швидкість основного автомата, яка була наведена в таблиці вище, була заміряна з включеним «передбаченням». Якість «передбачення» була підрахована лише для російської мови. Це було зроблено в такий спосіб. Взяті новинні тексти, навмання обрані 150 «передбачених» слів, що не повторюються. Ці слова не повинні бути абревіатурами (усі літери у верхньому регістрі). Усі слова виявилися або іменниками, або прикметниками. Для 131 слова в результатах «передбачення» був хоча б один правильний результат (одночасно лема, частина мови, рід, число і відмінок). Тобто точність передбачення – 87%. Цей результат цілком порівнюваний з результатами інших дослідників, наприклад, для англійської мови – 85%, або для французької – 88%.

Контрольні питання

1. Що таке АОТ?
2. Опишіть повний функціонал цього програмного засобу.

Домашнє завдання

Опишіть, як саме морфологічний модуль задіяний в інших програмних засобах АОТ: http://www.aot.ru/onlinedemo.html.

Лабораторна робота №3
АВТОМАТИЧНИЙ МОРФОЛОГІЧНИЙ АНАЛІЗАТОР АОТ

1. Проаналізуйте усі словоформи з поданого тексту у автоматичному морфологічному аналізаторі АОТ http://www.aot.ru/demo/morph.html

(див. Рис. 19) (попередньо ознайомтеся з дескрипторами російської морфології – див. Рис. 20):

Рабочая группа АОТ (Автоматическая обработка текста, руководитель – Алексей Сокирко) выкладывает свои разработки в свободный доступ начиная с 2002 года. Их библиотеки можно использовать бесплатно даже в коммерческих проектах. Проект включает в себя модули для проведения графематического, морфологического, синтаксического и семантического анализа. Разработчики реализовали снятие морфологической неоднозначности с использованием скрытых марковских моделей и синтаксический анализатор именных групп. Модуль морфологического анализа АОТ реализован в виде библиотеки на языке C++ и сопровождается программой-редактором словарей (MorphWiz), позволяющей в удобном интерфейсе просматривать содержимое морфологического словаря, добавлять, удалять и исправлять описание слов. Морфология в проекте АОТ включает в себя словари для русского (174 тысячи лемм), английского (104 тысячи лемм) и немецкого (218 тысяч лемм) языков. За основу русского морфологического словаря был взят грамматический словарь Зализняка. Все три словаря записаны в одинаковом формате, а поиск по ним осуществляется одним и тем же программным кодом. Демонстрационный интерфейс поиска по морфологическому словарю развернут на сайте проекта АОТ. Морфологический модуль (включающий библиотеку LemmatizerLib) обрабатывает словоформы по отдельности и не учитывает их контекст. Результатом анализа словоформы является набор морфологических гипотез, каждая из которых включает следующие данные: флаг словарности, который показывает, основана ли гипотеза на словарной лемме или сгенерирована предиктивным алгоритмом; набор неизменяемых грамматических признаков (граммем), например одушевленность существительного или вид глагола; текстовая строка, представляющая лемму (аношкинские коды); часть речи; множество наборов изменяемых грамматических признаков (граммем) для данной словоформы, по набору для

каждого варианта лемматизации, например число существительного, род прилагательного и пр.

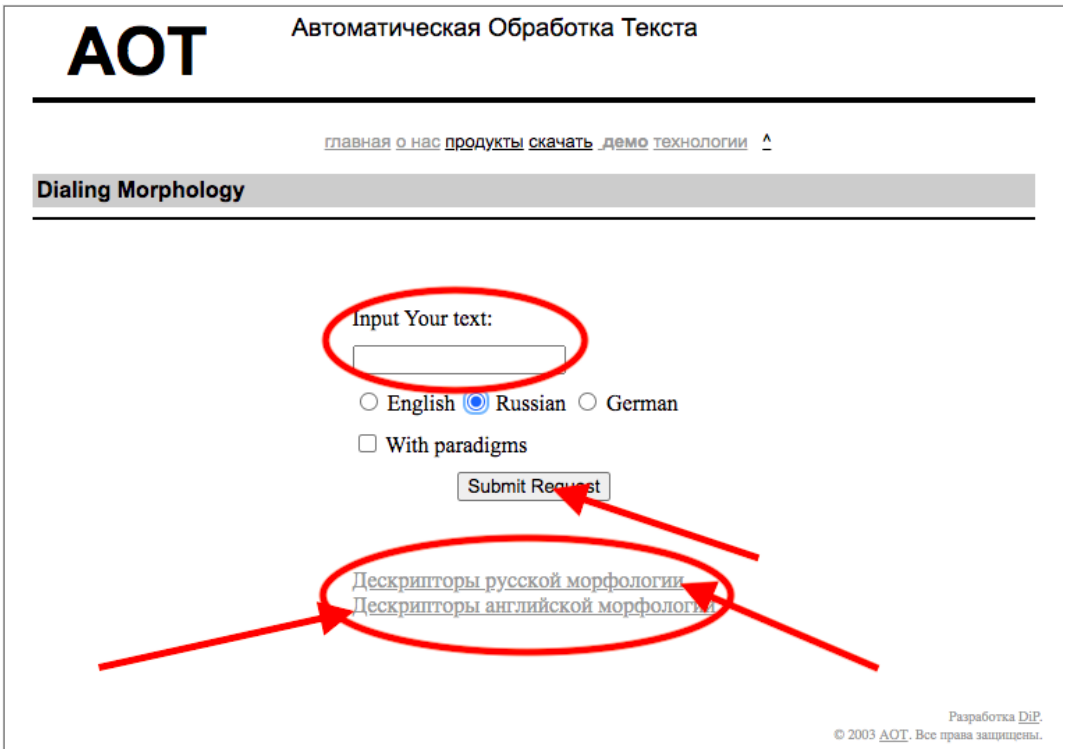
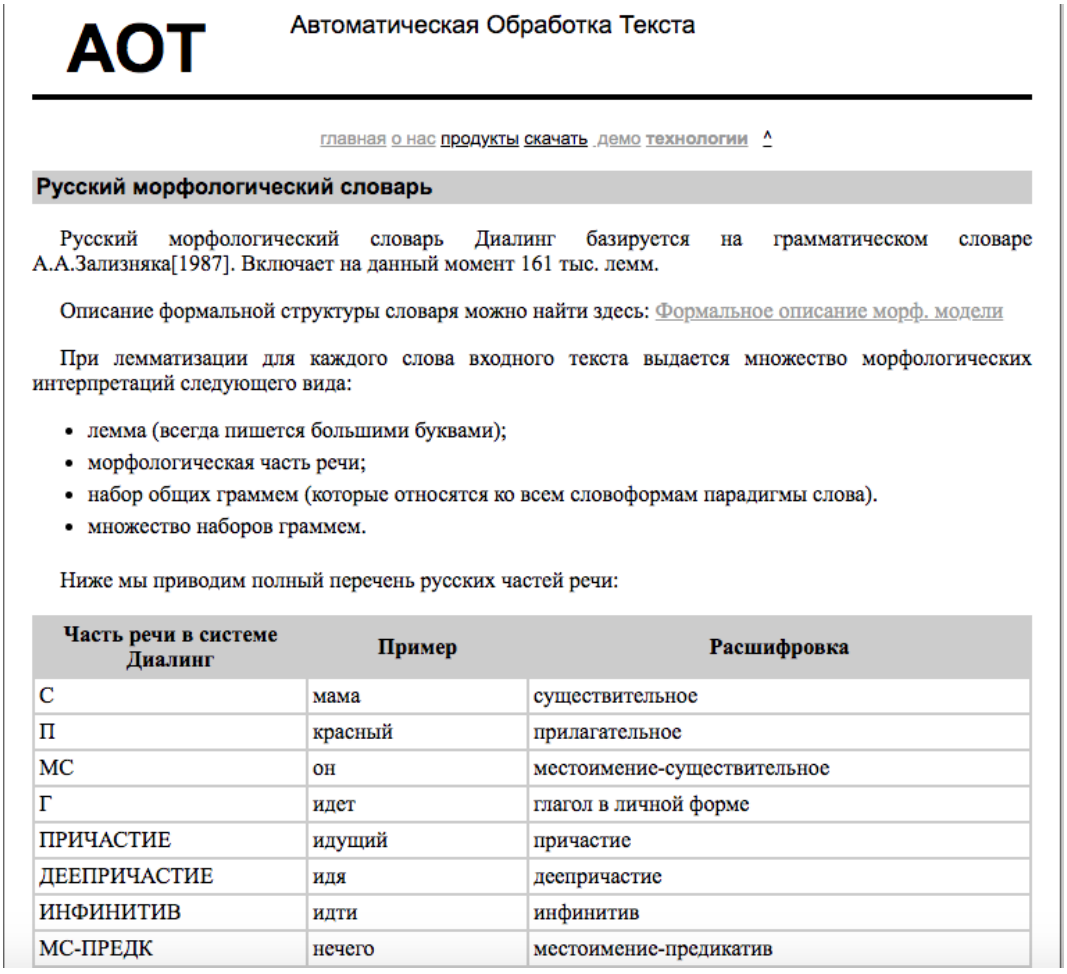


Рисунок 19. Інтерфейс автоматичного морфологічного аналізатора AOT



Часть речи в системе Диалинг	Пример	Расшифровка
С	мама	существительное
П	красный	прилагательное
МС	он	местоимение-существительное
Г	идет	глагол в личной форме
ПРИЧАСТИЕ	идущий	причастие
ДЕЕПРИЧАСТИЕ	идя	деепричастие
ИНФИНИТИВ	идти	инфинитив
МС-ПРЕДК	нечего	местоимение-предикатив

МС-П	всякий	местоименное прилагательное
ЧИСЛ	восемь	числительное (количественное)
ЧИСЛ-П	восьмой	порядковое числительное
Н	круто	наречие
ПРЕДК	интересно	предикатив
ПРЕДЛ	под	предлог
СОЮЗ	и	союз
МЕЖД	ой	междометие
ЧАСТ	же, бы	частица
ВВОДН	конечно	вводное слово
КР_ПРИЛ	красива	краткое прилагательное
КР_ПРИЧАСТИЕ	построена	краткое причастие

Граммема - это элементарный морфологический описатель, относящий словоформу к какому-то морфологическому классу, например, словоформе *стол* с леммой СТОЛ будут приписаны следующие наборы граммем: **"мр, ед, им, но", "мр, ед, вн, но"**. Таким образом, морфологический анализ выдает два варианта анализа словоформы *стол* с леммой СТОЛ внутри одной морфологической интерпретации: с винительным (**вн**) и именительным падежами (**им**).

Ниже перечислены все используемые граммемы:

мр, жр, ср - мужской, женский, средний род;

од, но - одушевленность, неодушевленность;

ед, мн - единственное, множественное число;

им, рд, дт, вн, тв, пр, зв - падежи: именительный, родительный, дательный, винительный, творительный, предложный, звательный;

2 - обозначает второй родительный или второй предложный падежи;

св, нс - совершенный, несовершенный вид;

пе, нп - переходный, непереходный глагол;

дст, стр - действительный, страдательный залог;

нст, прш, буд - настоящее, прошедшее, будущее время;

пвл - повелительная форма глагола;

1л, 2л, 3л - первое, второе, третье лицо;

0 - неизменяемое.

кр - краткость (для прилагательных и причастий).

сравн - сравнительная форма (для прилагательных).

имя, фам, отч - имя, фамилия, отчество.

лок, орг - локативность, организация.

кач - качественное прилагательное.

вопр,относ - вопросительность и относительность (для наречий).

дфст - слово обычно не имеет множественного числа.

опч - частая опечатка или ошибка.

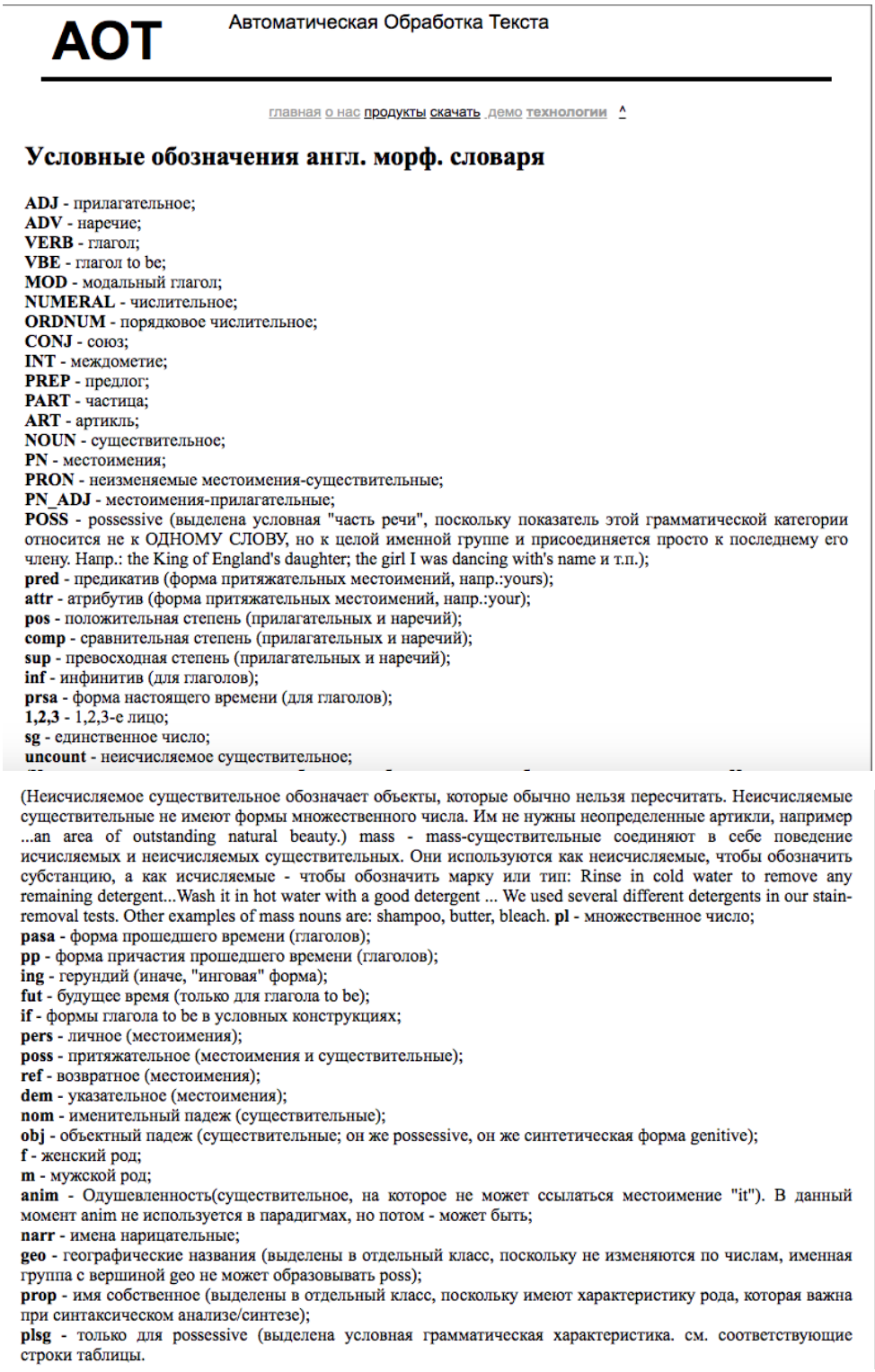
жарг, арх, проф - жаргонизм, архаизм, профессионализм.

аббр - аббревиатура.

безл - безличный глагол.

Рисунок 20. Дескриптори російської морфології

2. Перекладіть текст англійською і здійсніть той самий аналіз (попередньо ознайомтеся з дескрипторами англійської морфології – див. Рис. 21):



AOT Автоматическая Обработка Текста

главная о нас продукты скачать демо технологии ^

Условные обозначения англ. морф. словаря

ADJ - прилагательное;
ADV - наречие;
VERB - глагол;
VBE - глагол to be;
MOD - модальный глагол;
NUMERAL - числительное;
ORDNUM - порядковое числительное;
CONJ - союз;
INT - междометие;
PREP - предлог;
PART - частица;
ART - артикль;
NOUN - существительное;
PN - местоимения;
PRON - неизменяемые местоимения-существительные;
PN_ADJ - местоимения-прилагательные;
POSS - possessive (выделена условная "часть речи", поскольку показатель этой грамматической категории относится не к ОДНОМУ СЛОВУ, но к целой именной группе и присоединяется просто к последнему его члену. Напр.: the King of England's daughter; the girl I was dancing with's name и т.п.);
pred - предикатив (форма притяжательных местоимений, напр.:yours);
attr - атрибутив (форма притяжательных местоимений, напр.:your);
pos - положительная степень (прилагательных и наречий);
comp - сравнительная степень (прилагательных и наречий);
sup - превосходная степень (прилагательных и наречий);
inf - инфинитив (для глаголов);
prsa - форма настоящего времени (для глаголов);
1,2,3 - 1,2,3-е лицо;
sg - единственное число;
uncount - неисчисляемое существительное;

(Неисчисляемое существительное обозначает объекты, которые обычно нельзя пересчитать. Неисчисляемые существительные не имеют формы множественного числа. Им не нужны неопределенные артикли, например ...an area of outstanding natural beauty.) mass - mass-существительные соединяют в себе поведение исчисляемых и неисчисляемых существительных. Они используются как неисчисляемые, чтобы обозначить субстанцию, а как исчисляемые - чтобы обозначить марку или тип: Rinse in cold water to remove any remaining detergent...Wash it in hot water with a good detergent ... We used several different detergents in our stain-removal tests. Other examples of mass nouns are: shampoo, butter, bleach. **pl** - множественное число;
pasa - форма прошедшего времени (глаголов);
pp - форма причастия прошедшего времени (глаголов);
ing - герундий (иначе, "инговая" форма);
fut - будущее время (только для глагола to be);
if - формы глагола to be в условных конструкциях;
pers - личное (местоимения);
poss - притяжательное (местоимения и существительные);
ref - возвратное (местоимения);
dem - указательное (местоимения);
nom - именительный падеж (существительные);
obj - объектный падеж (существительные; он же possessive, он же синтетическая форма genitive);
f - женский род;
m - мужской род;
anim - Одушевленность(существительное, на которое не может ссылаться местоимение "it"). В данный момент anim не используется в парадигмах, но потом - может быть;
narr - имена нарицательные;
geo - географические названия (выделены в отдельный класс, поскольку не изменяются по числам, именная группа с вершиной geo не может образовывать poss);
prop - имя собственное (выделены в отдельный класс, поскольку имеют характеристику рода, которая важна при синтаксическом анализе/синтезе);
plsg - только для possessive (выделена условная грамматическая характеристика. см. соответствующие строки таблицы.

Рисунок 21. Дескриптори англійської морфології

3. Заскріньте результати опрацювання, заархівуйте та надішліть викладачеві.

Література до теми:

1. Бабина О. Корпусный метод автоматического морфологического анализа флективных языков. *Вестник Южно-Уральского государственного университета. Серия «Лингвистика»*. №25(284), выпуск 15. Челябинск, 2012. С. 38-44.
2. Баранов А. Введение в прикладную лингвистику. М., 2001. 360 с.
3. Белоногов Г. Компьютерная лингвистика и перспективные информационные технологии. М., 2004. 159 с.
4. Буньо Г. Сучасні методи вирішення проблеми граматичної омонімії в тексті. *Наукові записки. Серія: Філологічна*. 2014. Вип. 49. С. 12-16.
5. Волошин В. Комп'ютерна лінгвістика. Суми, 2004. 382 с.
6. Гельбух А., Сидоров Г. К вопросу об автоматическом морфологическом анализе флективных языков. Ел. режим доступу: www.dialog-21.ru/Archive/2005.
7. Грязнухіна Т., Нікула М. Система автоматичного морфологічного аналізу українського наукового тексту. Проблеми українізації комп'ютерів. Матеріали 2-ї міжнародної конференції. Київ, 1993. С. 42-46.
8. Данилюк І. Прикладна морфологія. Донецьк, 2010. 216 с.
9. Дарчук Н. Комп'ютерна лінгвістика. К., 2008. 351 с.
10. Дарчук Н. Комп'ютерне анотування українського тексту: результати і перспективи. К., 2013. 543 с.
11. Дарчук Н. Морфологічне анотування Корпусу української мови. *Комп'ютерна лінгвістика: сучасне і майбутнє. Матеріали Міжнародної науково-практичної конференції*. К., 2012. С. 16-19.
12. Карпіловська Є. Вступ до прикладної лінгвістики: Комп'ютерна лінгвістика. Донецьк, 2006. 188 с.
13. Марчук Ю. Компьютерная лингвистика. М., 2007. 317 с.

14. Міщенко Н. Система програм морфологічного аналізу науково-технічних текстів. *Наукові записки. Серія «Філологічні науки»*. Випуск 95(2). К., 2011. С. 538-542.
15. Морфологический анализ научного текста на ЭВМ. К., 1989. 262 с.
16. Николаев И., Митренина О., Ландо Т. Прикладная и компьютерная лингвистика. М., 2016. 315 с.
17. Ножов И. Морфологическая и синтаксическая обработка текста (модели и программы). М., 2003. 140 с.
18. Партико З. Прикладна і комп'ютерна лінгвістика: Вступ до спеціальності. Львів, 2008. 224 с.
19. Романюк Ю. Прикладна морфологія. Черкаси, 2009. 117 с.
20. Соснина Е. Введение в прикладную лингвистику. Ульяновск, 2012. 110 с.
21. Сучасна українська літературна мова. Морфологія. К., 1969. 250 с.
22. Antworth E. PC-KIMMO: a two-level processor for morphological analysis. *Occasional Publications in Academic Computing*. No. 16. Dallas, 1990. 273 p.

ТЕМА 4. АВТОМАТИЧНИЙ МОРФОЛОГІЧНИЙ АНАЛІЗАТОР STARLING

Опорний конспект

1. Автоматичний морфологічний аналізатор StarLing дає можливість ознайомитися з комп'ютерними базами даних за словниками Ожегова, Залізняка і Мюллера, а також проаналізувати будь-яке російське й англійське слово та отримати його повну акцентовану парадигму.

2. У базах даних кожне заголовкове слово має відсилання до програми автоматичного морфологічного аналізу. Цю програму можна викликати і в якості окремого вікна. В останньому випадку може бути введене будь-яке російське або англійське слово в довільній граматичній формі. Програмою аналізу видаються такі відомості: 1) для російського слова – вихідна словоформа (за Залізняком); словникова інформація, тобто морфологічний індекс російського слова і наявні коментарі з Граматичного словника Залізняка; переклад, тобто набір словникових статей зі словника Мюллера, у яких міститься відповідне російське слово, з готовими відсиланнями на відповідні словникові статті; морфологічна характеристика введеного російського слова (у разі багатозначності введеної форми виводяться всі варіанти аналізу); 2) для англійського слова – словникова стаття зі словника Мюллера (у разі багатозначності форми виводяться всі відповідні статті). Потім наводяться повні акцентовані парадигми для кожного з результатів аналізу.

Контрольні питання

1. Що таке StarLing?
2. Опишіть повний функціонал цього програмного засобу.

Домашнє завдання

Укладіть аналітико-порівняльну таблицю *«Функційні можливості автоматичних морфологічних аналізаторів АОТ та StarLing»*.

Лабораторна робота №4
АВТОМАТИЧНИЙ МОРФОЛОГІЧНИЙ АНАЛІЗАТОР STARLING

1. Проаналізуйте усі словоформи з поданих текстів у автоматичному морфологічному аналізаторі StarLing https://starling.rinet.ru/cgi-bin/morphque.cgi?encoding=win (див. Рис. 22):

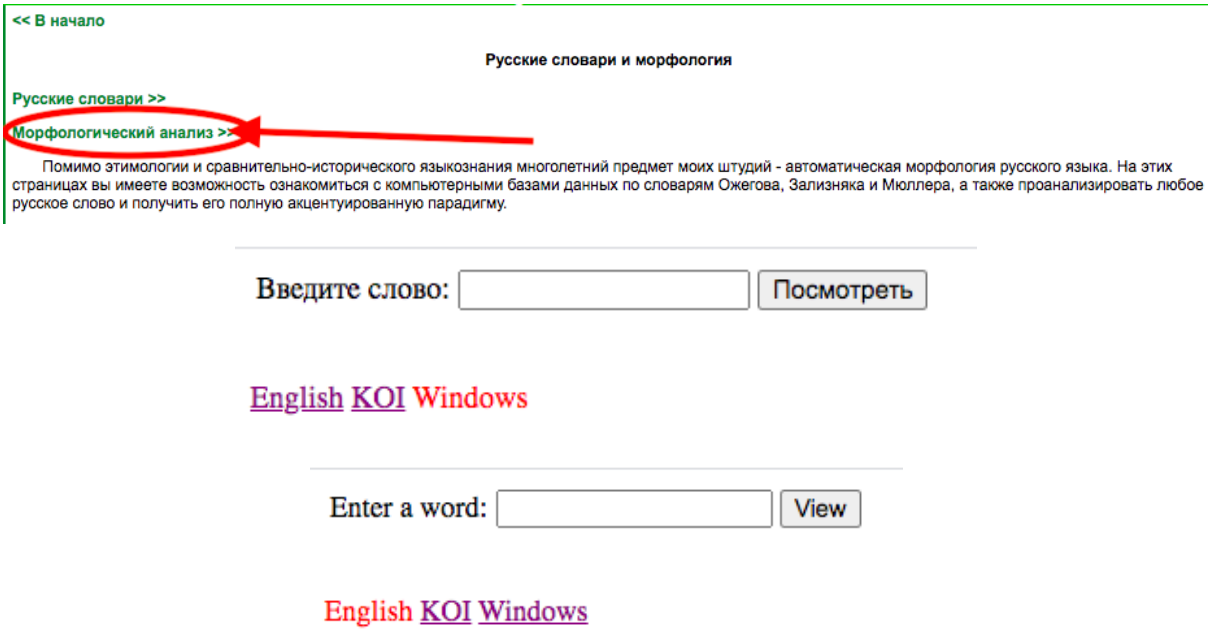


Рисунок 22. Інтерфейс автоматичного морфологічного аналізатора StarLing

В диссертации впервые в украинской прикладной лингвистике предложены алгоритмы полного цикла автоматического анализа украинского текста. Цель исследования – теоретическое и экспериментальное обоснование лингвистических и процедурных принципов интегральной модели семантико-грамматического взаимодействия знаковых единиц в тексте и создание на этой основе компьютерной грамматики украинского языка (АГАТ). Созданы частеречные лексиконы, в которых размещена необходимая грамматическая и лексическая информация для анализа морфологии, синтаксиса и семантики украинского языка; разработаны лингвистическая стратегия и правила автоматического морфологического анализа, автоматического морфного сегментирования словоформ текста; автоматического синтаксического анализа; автоматического семантического анализа в виде терминологического тезауруса информационно-поискового типа, идеографических словарей для существительного, глагола; выявлены и систематизированы грамматические явления, характерные для синтаксиса украинского языка с использованием Корпуса украинского языка. Все алгоритмы программно реализованы и протестированы на реальных текстах.

Throughout the scientific research there was proved the relation between theoretical and computational linguistics and also the role of structural linguistics as the theoretical basis for the computational linguistics. The obtained results were generalised in the grammar checking software AGAT. This software functions as a dynamic system. It comprises the grammatical meanings and the forms of its expressions, and the system of grammatical rules. They provide access to the denotative information, described by automated morphological analysis, to the relative information, described by automated syntactical analysis and also to the basic conceptual blocks which build the language thesaurus. The electronically compiled lists of words and grammar are both very closely interrelated and correlated components of the language structure. Their correlation is based on the conformity of their basic functions with the storage in computer memory of both linguistic units which are ready for use, and grammar rules according to which and with respect to a specified task the text analysis is accomplished. As part of the research following activities were performed: there were developed the lists of parts of speech, containing grammatical and lexical information for accomplishment of morphological, syntactic and semantic analyses of the Ukrainian language; there were created the linguistic approach and principles for performing of the automated morphological analysis, automated morph segmentation of word forms in texts, and automated syntactic analysis.

2. Заскріньте результати опрацювання, заархівуйте та надішліть викладачеві.

Література до теми:

1. Бабина О. Корпусный метод автоматического морфологического анализа флективных языков. *Вестник Южно-Уральского государственного университета. Серия «Лингвистика»*. №25(284), выпуск 15. Челябинск, 2012. С. 38-44.
2. Баранов А. Введение в прикладную лингвистику. М., 2001. 360 с.
3. Белоногов Г. Компьютерная лингвистика и перспективные информационные технологии. М., 2004. 159 с.

4. Буньо Г. Сучасні методи вирішення проблеми граматичної омонімії в тексті. *Наукові записки. Серія: Філологічна*. 2014. Вип. 49. С. 12-16.
5. Волошин В. Комп'ютерна лінгвістика. Суми, 2004. 382 с.
6. Гельбух А., Сидоров Г. К вопросу об автоматическом морфологическом анализе флективных языков. Ел. режим доступу: www.dialog-21.ru/Archive/2005.
7. Грязнухіна Т., Нікула М. Система автоматичного морфологічного аналізу українського наукового тексту. Проблеми українізації комп'ютерів. Матеріали 2-ї міжнародної конференції. Київ, 1993. С. 42-46.
8. Данилюк І. Прикладна морфологія. Донецьк, 2010. 216 с.
9. Дарчук Н. Комп'ютерна лінгвістика. К., 2008. 351 с.
10. Дарчук Н. Комп'ютерне анотування українського тексту: результати і перспективи. К., 2013. 543 с.
11. Дарчук Н. Морфологічне анотування Корпусу української мови. *Комп'ютерна лінгвістика: сучасне і майбутнє. Матеріали Міжнародної науково-практичної конференції*. К., 2012. С. 16-19.
12. Карпіловська Є. Вступ до прикладної лінгвістики: Комп'ютерна лінгвістика. Донецьк, 2006. 188 с.
13. Марчук Ю. Компьютерная лингвистика. М., 2007. 317 с.
14. Міщенко Н. Система програм морфологічного аналізу науково-технічних текстів. *Наукові записки. Серія «Філологічні науки»*. Випуск 95(2). К., 2011. С. 538-542.
15. Морфологический анализ научного текста на ЭВМ. К., 1989. 262 с.
16. Николаев И., Митренина О., Ландо Т. Прикладная и компьютерная лингвистика. М., 2016. 315 с.
17. Ножов И. Морфологическая и синтаксическая обработка текста (модели и программы). М., 2003. 140 с.
18. Партико З. Прикладна і комп'ютерна лінгвістика: Вступ до спеціальності. Львів, 2008. 224 с.

19. Романюк Ю. Прикладна морфологія. Черкаси, 2009. 117 с.
20. Соснина Е. Введение в прикладную лингвистику. Ульяновск, 2012. 110 с.
21. Сучасна українська літературна мова. Морфологія. К., 1969. 250 с.
22. Antworth E. PC-KIMMO: a two-level processor for morphological analysis. *Occasional Publications in Academic Computing*. No. 16. Dallas, 1990. 273 p.

ТЕМА 5. АВТОМАТИЧНИЙ МОРФОЛОГІЧНИЙ АНАЛІЗАТОР MCR DLL V2.0

Опорний конспект

Автоматичний морфологічний аналізатор MCR DLL v2.0 призначений для опрацювання російськомовних текстів (див. Рис. 23):

Описание функций для C++ и Pascal

Перейти к грамматическим характеристикам

Описание функций

Init ⟹ Load ⟹ FindId⟹ GetBid
Init ⟹ AddParadigm ⟹ Save

Init
Инициализировать словарь mcr
typedef int (*p1Func)();

LoadMcr
Загрузить словарь mcr.
typedef int (*p2Func)(const char * s);

SaveMcr
Сохранить словарь mcr
typedef int (*p3Func)(const char * s);

FindID
FindID - поиск слова s в словаре mcr, возвращает int количество найденых слов и ids (уникальный идентификтор слова = уникальный идентификатор леммы + номер слова в парадигме). Используйте полученные идентификаторы для получения грамматических характеристик слова или возврата всей парадигмы
typedef (*p4Func)(const char * s, Tids * ids);

GetByID
Для идентификатора id, функция возвращает грамматические характеристики, лемму или всю парадигмы в outdata
:: gh_only = true - возвращать только грамматические характеристики
:: gh_only = false - возвратить грамматические характеристики для id и лемму (win1251)
:: all = true - поместить в outdata всю парадигму для указанного id с грамматическими характеристиками, доступно только для словарей без пометы ReadOnly
typedef (*p5Func)(const Tid id,bool gh_only, bool all,Tinlexdata * outdata);

ReadOnly
Проверка является ли подключенный словарь - словарем только для чтения (ReadOnlyDict)
typedef bool (*p6Func)();

AddParadigma
AddPara - добавление парадигмы в словарь. См. рисунок
typedef(*p7Func)(Tinlexdata * indata);

FreeSpace

Проверка свободного места в словаре, если любой из аргументов близок к 100% то добавление парадигм будет вскоре невозможно
typedef(*p8Func)(unsigned char * ar1, unsigned char * ar2,unsigned char * ar3);

Ver

Информация о версии mcr.dll
typedefint (*p9Func)(char * s);

CSTR

Строка постоянной грамматической характеристики
Используется для перевода числа в стоку для наглядного отображения грамматических характеристик. Вы можете использовать эту функцию в программе, чтобы не обращяться к таблице кодирования гр. характиристик которая приведена ниже.
typedef char * (*p10Func)(const unsigned char cid);

VSTR

Строка переменной грамматической характеристики
typedef * (*p11Func)(const unsigned char cid, const unsigned char vid);

Рисунок 23. Функціонал автоматичного морфологічного аналізатора MCR DLL v2.0

Таблица кодирования постоянных грамматических характеристик

Имя существительное Прилагательное и др. Числительное Глагол Деепричасте Прочее

Имя существительное

Cid	Описание	Vid
1	Существительное Мужского рода (неодушевленное)	link
2	Существительное Мужского рода (одушевленное)	link
3	Существительное Женского рода (неодушевленное)	link
4	Существительное Женского рода (одушевленное)	link
5	Существительное Среднего рода (неодушевленное)	link
6	Существительное Среднего рода (одушевленное)	link
7	Существительное Мужского-Женского рода (неодушевленное)	link
8	Существительное Мужского-Женского рода (одушевленное)	link
9	Существительное Мужского-Среднего рода (неодушевленное)	link
10	Существительное Мужского-Среднего рода (одушевленное)	link
11	Существительное Женского-Среднего рода (неодушевленное)	link
12	Существительное Женского-Среднего рода (одушевленное)	link
13	Существительное только множественное число(неодушевленное)	link
14	Существительное только множественное число (одушевленное)	link
15	Существительное *	link

Имя прилагательное

Cid	Описание	Vid
20	Прилагательное	link
21	Местоимение	link
22	Местоименное прилагательное	link
23	Числительное собирательное	link
24	Числительное прилагательное	link
25	Числительное	link
26	Местоименное прилагательное(краткое)	link

Спряжение

В данном релизе, нет спецальных помет для безличных, многократных и вспомогательных глаголов.

Cid	Описание	Vid
40	Глагол НСВ (несовершенного вида) невозвратный I спряжение	link
41	Глагол НСВ невозвратн II	link
42	Глагол НСВ возвратн I	link
43	Глагол НСВ возвратн II	link
44	Глагол СВ(совершенного вида) невозвратн I спряжение	link
45	Глагол СВ невозвратн II	link
46	Глагол СВ возвратн I	link
47	Глагол СВ возвратн II	link
48	Глагол СВ-НСВ I (двувидовый глагол)	link
48	Глагол СВ-НСВ II	link
50	Глагол (СВ)-НСВ возвратный I (совершенность носит потенциальный характер)	link
51	Глагол (СВ)-НСВ возвратный II	link

Отглагольные формы

Cid	Описание	Vid
60	Причастие Настоящего времени (от НСВ I)	link
61	Причастие Настоящего времени (от НСВ II)	link
62	Причастие Настоящего времени (от НСВ I) страдательное значение(на -ся)	link
63	Причастие Настоящего времени (от НСВ II) страдательное значение	link
64	Причастие Прошедшего времени (от НСВ I)	link
65	Причастие Прошедшего времени (от НСВ II)	link
66	Причастие Прошедшего времени (от НСВ I) страдательное значение	link
67	Причастие Прошедшего времени (от НСВ II) страдательное значение	link
68	Причастие Прошедшего времени (от СВ I)	link
69	Причастие Прошедшего времени (от СВ II)	link
70	Причастие Прошедшего времени (от СВ I) страдательное значение	link
71	Причастие Прошедшего времени (от СВ II) страдательное значение	link
72	Страдательное Причастие настоящего времени (от НСВ I)	link
73	Страдательное Причастие настоящего времени (от НСВ II)	link
74	Страдательное Причастие прошедшего времени (от НСВ I)	link
75	Страдательное Причастие прошедшего времени (от НСВ II)	link
76	Страдательное Причастие прошедшего времени (от СВ I)	link
77	Страдательное Причастие прошедшего времени (от СВ II)	link
78	Деепричастие (от НСВ I)	link
79	Деепричастие (от НСВ II)	link
80	Деепричастие (от СВ I)	link
81	Деепричастие (от СВ II)	link

Остальные части речи

Cid	Описание	Vid
30	Наречие	link
31	Союз	link
32	Междометие	link
33	Частица	link
34	Предлог	link
35	Предикат	link
36	Вводное слово	link
37	Неизменяемое слово	link
200	Имя собственное *	link
201	Имя собственное мужского рода	link
202	Имя собственное женского рода	link
203	Отчество муж. род	link
204	Отчество женск. род	link
205	Фамилия	link
206	Название *	link
207	Географическое название	link
208	Географическое название мужского рода	link
209	Географическое название женского рода	link
210	Географическое название среднего рода	link
211	Географическое название только множественное число	link
212	Прилагательное образованное от геогр. названия	link
213	Аббревиатура	link
214	Аббревиатура (все прописные)	link
215	Аббревиатура (все ЗАГЛАВНЫЕ)	link
216	Сокращение *кг,сек,см. рис. и.т п*	link

Таблица кодирования переменных грамматических характеристик

Для существительного

Vid	Описание
0	Все формы одинаковы
1	Ед.ч. И.П. *(единственное число, именительный падеж)*
2	Ед.ч. Р.П.
3	Ед.ч. Д.П.
4	Ед.ч. В.П.
5	Ед.ч. Т.П.
6	Ед.ч. П.П.
7	Мн.ч. И.П. *(множественное число, именительный падеж)*
8	Мн.ч. Р.П.
9	Мн.ч. Д.П.
10	Мн.ч. В.П.
11	Мн.ч. Т.П.
12	Мн.ч. П.П.
13	только мн. ч. (все формы одинаковы)

Для прилагательных и схожих частей

Vid	Описание
1	И.П. М.р ед.ч.од/неод *(Именительный падеж, мужск. род, ед. число, одушевленное и неодушевленное)*
2	И.П. С.р ед.ч.од/неод
3	Р.П. М/С.р ед.ч.од/неод
4	Д.П. М/С.р ед.ч.од/неод
5	В.П. М.р ед.ч.неод.
6	В.П. М.р ед.ч.одуш.
7	В.П. С.р ед.ч.од/неод
8	Т.П. М/С.р ед.ч.од/неод
9	П.П. М/С ед.ч.од/неод
10	И. Ж.р ед.ч.од/неод
11	Р,Д,П и Ж ед.ч.од/неод
12	В. Ж.р ед.ч.од/неод
13	Т. Ж.р ед.ч.од/неод
14	И. Мн.ч. од/неод
15	Р.Мн.ч. од/неод
16	Д. Мн.ч. од/неод
17	В. Мн.ч. неод.
18	В. Мн.ч. од.
19	Т. Мн.ч. од/неод
20	Т. Ж.р еч.од/неод
21	Кратк.форма М.р
22	Кратк.форма Ж.р
23	Кратк.форма С.р
24	Кратк.форма Мн. всех родов
25	Сравнительная степень
26	Сравнительная степень (параллельный вариант ее/ей)

Числительное

Vid	Описание
0	все формы одинаковы
1	И.П.
2	Р.П.
3	Д.П.
4	В.П.
5	В.П. одушевленное
6	Т.П.
7	П.П.
8	Т.П. (параллельн)
9	М/С род И.П.
10	М/С род Р.П.
11	М/С род Д.П.
12	М/С род В.П.
13	М/С род В.П. одушевл
14	М/С род Т.П.
15	М/С род П.П.
16	Ж род И.П.
17	Ж род Р.П.
18	Ж род Д.П.
19	Ж род В.П.
20	Ж род В.П. одушевл
21	Ж род Т.П.
22	Ж род П.П.

Глагол

Vid	Описание
1	ИнФинитив
2	Н.вр Ед.ч 1 лицо
3	Н.вр Ед.ч 2 лицо.
4	Н.вр Ед.ч 3 лицо
5	Н.вр Мн.ч 1 лицо
6	Н.вр Мн.ч 2 лицо
7	Н.вр Мн.ч 3 лицо
8	Пр.вр Ед.всех лиц М род
9	Пр.вр Ед.всех лиц Ж род
10	Пр.вр Ед.всех лиц С род
11	Пр.вр Мн.всех лиц родов
12	Повел. 2 лицо Ед.
13	Повел. 2 лицо Мн.
14	Повел. 1 лицо Мн.(к одному)
15	Повел. 1 лицо Мн.(ко многим)
16	Буд.вр Ед.ч 1 лицо
17	Буд.вр Ед.ч 2 лицо
18	Буд.вр Ед.ч 3 лицо
19	Буд.вр Мн.ч 1 лицо
20	Буд.вр Мн.ч 2 лицо
21	Буд.вр Мн.ч 3 лицо
25	Н/Буд. вр Ед.ч 1 лицо
26	Н/Буд. вр Ед.ч 2 лицо
27	Н/Буд. вр Ед.ч 3 лицо
28	Н/Буд. вр Мн.ч 1 лицо
29	Н/Буд. вр Мн.ч 2 лицо
30	Н/Буд. вр Мн.ч 3 лицо

Деепричастие

Лингвисты как правило не различают времени у прилагательного, но раз такая инфрмация была введена в словаре Зализнякак, то 2 характеристики имеют место

Vid	Описание
1	Настоящего времени
2	Прошедшего времен

Прочее

Vid	Описание
0	NULL

Контрольні питання

1. Що таке MCR DLL v2.0?
2. Опишіть повний функціонал цього програмного засобу.

Домашнє завдання

Порівняйте програмні можливості MCR DLL v2.0 з потенціалом інших автоматичних морфологічних аналізаторів.

Лабораторна робота №5

АВТОМАТИЧНИЙ МОРФОЛОГІЧНИЙ АНАЛІЗАТОР MCR DLL V2.0

1. Скачайте програмний засіб MCR DLL v2.0 http://macrocosm.narod.ru/manual.html, попередньо ознайомившись із інструкцією (див. Рис. 24):

C#

Класс для использования морфологического модуля mcr.dll 32-бит на C# можно скачать здесь

Javadoc

Описание работы морфологии с Java вы найдете здесь

Рисунок 24. Інструкція до програмного засобу MCR DLL v2.0

2. Перевірте функційні можливості зазначеного автоматичного морфологічного аналізатора на прикладі невеликого тексту публіцистичного стилю (довжиною 1500 знаків без пробілів – знайдіть самостійно).

3. Заскріньте результати опрацювання, заархівуйте та надішліть викладачеві.

Література до теми:

1. Бабина О. Корпусный метод автоматического морфологического анализа флективных языков. *Вестник Южно-Уральского государственного университета. Серия «Лингвистика»*. №25(284), выпуск 15. Челябинск, 2012. С. 38-44.
2. Баранов А. Введение в прикладную лингвистику. М., 2001. 360 с.
3. Белоногов Г. Компьютерная лингвистика и перспективные информационные технологии. М., 2004. 159 с.
4. Буньо Г. Сучасні методи вирішення проблеми граматичної омонімії в тексті. *Наукові записки. Серія: Філологічна*. 2014. Вип. 49. С. 12-16.
5. Волошин В. Комп'ютерна лінгвістика. Суми, 2004. 382 с.
6. Гельбух А., Сидоров Г. К вопросу об автоматическом морфологическом анализе флективных языков. Ел. режим доступу: www.dialog-21.ru/Archive/2005.

7. Грязнухіна Т., Нікула М. Система автоматичного морфологічного аналізу українського наукового тексту. Проблеми українізації комп'ютерів. Матеріали 2-ї міжнародної конференції. Київ, 1993. С. 42-46.
8. Данилюк І. Прикладна морфологія. Донецьк, 2010. 216 с.
9. Дарчук Н. Комп'ютерна лінгвістика. К., 2008. 351 с.
10. Дарчук Н. Комп'ютерне анотування українського тексту: результати і перспективи. К., 2013. 543 с.
11. Дарчук Н. Морфологічне анотування Корпусу української мови. *Комп'ютерна лінгвістика: сучасне і майбутнє. Матеріали Міжнародної науково-практичної конференції*. К., 2012. С. 16-19.
12. Карпіловська Є. Вступ до прикладної лінгвістики: Комп'ютерна лінгвістика. Донецьк, 2006. 188 с.
13. Марчук Ю. Компьютерная лингвистика. М., 2007. 317 с.
14. Міщенко Н. Система програм морфологічного аналізу науково-технічних текстів. *Наукові записки. Серія «Філологічні науки»*. Випуск 95(2). К., 2011. С. 538-542.
15. Морфологический анализ научного текста на ЭВМ. К., 1989. 262 с.
16. Николаев И., Митренина О., Ландо Т. Прикладная и компьютерная лингвистика. М., 2016. 315 с.
17. Ножов И. Морфологическая и синтаксическая обработка текста (модели и программы). М., 2003. 140 с.
18. Партико З. Прикладна і комп'ютерна лінгвістика: Вступ до спеціальності. Львів, 2008. 224 с.
19. Романюк Ю. Прикладна морфологія. Черкаси, 2009. 117 с.
20. Соснина Е. Введение в прикладную лингвистику. Ульяновск, 2012. 110 с.
21. Сучасна українська літературна мова. Морфологія. К., 1969. 250 с.

Додаток 1

МОДУЛЬНА КОНТРОЛЬНА РОБОТА

Варіант 1

Простий рівень

1. Автоматичний морфологічний аналіз (АМА) є частиною …:

1) процесу автоматичного опрацювання текстів (АОТ);

2) прикладної граматики;

3) прикладного синтаксису;

4) прикладної морфонології.

2. Способи вираження граматичного значення – це …:

1) синтетичний, аналітичний;

2) аналітико-синтетичний, суплетивний;

3) синтетичний, аналітичний, суплетивний;

4) синтетичний, аналітичний, аналітико-синтетичний, суплетивний.

3. Перші системи автоматичного морфологічного аналізу (АМА) було створено у …:

1) 50-70-х рр. ХХ ст.;

2) 50-60-х рр. ХХІ ст.;

3) 50-60-х рр. ХХ ст.;

4) 40-50-х рр. ХХ ст.

4. Машинне слово – це …:

1) одиниця писемної мови;

2) ланцюжок графем від пробілу до пробілу, у тому числі пунктуаційні знаки;

3) найменша (неподільна) структурно-семантична одиниця, що здатна виконувати функції у мовленні;

4) найменша самостійна й вільно відтворювана в мовленні відокремлено оформлена значуща одиниця мови.

5. Списки кінцевих буквосполучень служать для створення словників …:

1) основ і флексій;

2) тлумачних;

3) орфографічних;

4) перекладних.

6. Для укладання списку квазіфлексій створюється …:

1) гілка;

2) гніздо;

3) кущ;

4) дерево.

7. У чому суть автоматичного морфологічного аналізу (АМА) на базі словника словоформ:

1) аналіз графічної будови словоформи;

2) пошук словоформи у словнику й вибір відповідної граматичної інформації;

3) поділ словоформи на морфеми;

4) поділ словоформи на основу й флексію.

8. На скільки типів поділяються усі основи в АМА?

1) на 4;

2) на 2;

3) на 6;

4) на 8.

9. Що вважається елементарною одиницею морфологічного аналізу у процесі АМА методом логічного множення?

1) виділена основа;

2) виділений корінь;

3) виділена флексія;

4) кожна відкинута графема.

10. В автоматичному морфологічному аналізаторі АОТ можна здійснювати аналіз словоформ …:

1) української, російської, англійської мов;

2) російської, англійської, німецької мов;

3) української, німецької, англійської мов;

4) української, російської, французької мов.

Середній рівень

Дайте відповіді на питання:

* *Дайте визначення автоматичному морфологічному аналізу. Назвіть його завдання.*
* *Чим лексичне значення відрізняється від граматичного? Наведіть приклади.*
* *Дайте визначення морфологічній граматичній категорії. Наведіть приклади для однієї із самостійних частин мови.*
* *Які функції може виконувати крапка в тексті? Наведіть приклади.*
* *Дайте визначення квазіфлексії. Наведіть приклад.*
* *Дайте визначення дереву квазіфлексій.*
* *Що є недоліками автоматичного морфологічного аналізу на основі флективного аналізу?*

Високий рівень

1. Здійсніть автоматичний морфологічний аналіз поданого нижче тексту, використовуючи один із методів: на основі графемного аналізу / на основі флективного аналізу / на основі словника словоформ / на основі словника основ / на основі операції логічного множення:

Заповідні зони України станом на 2020 рік складають 6,6% її території. Це зовсім небагато, якщо порівнювати з площею заповідних зон країн світу та Європи зокрема: у Польщі, наприклад, ця частка складає 17%, а у Німеччині та Австрії – понад 30 %. Тимчасом як Україна є найбільшою європейською країною, площа її заповідних зон – найменша. В Україні найбільш поширеними типами заповідних зон є природні заповідники та

національні природні парки. Особливість природних парків – саме у їхній подвійній ролі: зберегти природу і поділитися її красою та ресурсами, залучаючи до парку туристів і даючи частковий дозвіл на господарську діяльність на території парку. І тут важливо дотримуватися балансу, аби діяльність людини не шкодила природі. Світові національні парки ретельно дотримуються цього балансу. У деяких парках можуть контрольовано проживати люди, в деяких – ні. Наприклад, у Гренландському національному парку – найбільшому парку світу площею 972 тисячі квадратних метрів – люди не проживають, проте його можуть відвідувати туристи. Єллоустоунський парк у США, який вважають одним із перших національних парків світу, пройшов довгий шлях до віднайдення балансу і нині є одним зі зразкових. В Україні ж довгі роки, починаючи з радянських часів, культивувалася мода на мисливство і значно поширилася його протизаконна форма – браконьєрство. Нераціональне використання природних ресурсів поступово стало нормою: полювання на червонокнижні види, понаднормова вирубка лісів, незаконна господарська діяльність у заповідних зонах тощо. З наслідками такого ставлення до природи доводиться боротися й досі.

2. Здійсніть автоматичний морфологічний аналіз поданого нижче тексту, використовуючи один із ресурсів: LanguageTool / libmorphukr / AOT / StarLing / MCR DLL v2.0:

Перед нами стояло завдання розробити комп'ютерну граматику для автоматичного аналізу українських текстів (АГАТ). АГАТ-граматика належить до комп'ютерних граматик за правилами і методами. Засадничими принципами комп'ютерної граматики є класичні аксіоми рівневості та відкритості опису, за якими можна за потреби розширювати і поглиблювати лінгвістичний базис АОТ, ускладнювати словникове й модифікувати програмне забезпечення без перебудови всієї системи. Автоматичне визначення частиномовних і словозмінних характеристик словоформ, автоматична ідентифікація словоформ однієї лексеми як основні завдання в морфологічній частині АГАТ-граматики можуть вирішуватися

формалізованими методиками, розробленими на основі традиційної і комп'ютерної граматик. Автоматичне визначення морфологічної інформації текстових одиниць, на основі якого в АМА здійснюється ідентифікація словоформ тієї самої лексеми, є обов'язковим складником лінгвістичного забезпечення систем АОТ флективних мов, що створює умови для конструювання й ефективної роботи таких модулів: синтаксичного, морфемного, а також семантичного. Для роботи АМА потрібне адекватне лінгвістичне забезпечення, яке базується на теоретичних засадах сучасної фундаментальної морфології. Це підвищує ефективність комп'ютерного аналізу, який зорієнтовано на максимальну формалізацію найважливіших понять морфології – частин мови і граматичних морфологічних категорій. Для АГАТ-морфології необхідно було вибрати частиномовну класифікацію з пропонованих традиційними граматиками, зважаючи на ступінь формальної обґрунтованості класифікаційних ознак. Між комп'ютерною і теоретичною морфологією існує глибинний зв'язок, експлікація якого є одним із важливих завдань створення АМА. Поясненням цього є різні «адресати» теорій – машина і людина. Сприймання інформації та оперування нею в комп'ютера і людини різні: у комп'ютері воно здійснюється тільки формальним або формалізованим способом; для людини це насамперед інтуїтивне сприйняття, посилене набутими нею спеціальними знаннями, певною мірою формалізованими.

Варіант 2

Простий рівень

1. Завданнями автоматичного морфологічного аналізу (АМА) є …:

1) перетворення фрагментів тексту без аналізу його змісту;

2) визначення частиномовної приналежності текстових одиниць, ідентифікація словоформ однієї лексеми;

3) проведення морфемного, синтаксичного й семантичного аналізів;

4) перетворення фрагментів тексту з аналізом його змісту, встановленням логіко-семантичних відношень між його компонентами.

2. Якими факторами обумовлений вибір принципів автоматичного морфологічного аналізу (АМА):

1) система мови, система письма і друку, тематика тексту;

2) граматика мови;

3) система мови, система мовлення;

4) система мовлення, система друку.

3. Яка із перерахованих систем автоматичного морфологічного аналізу (АМА) є експериментальною:

1) ПЛАЙ;

2) РЕФЕРАТ;

3) SMART;

4) ECAIT.

4. Початковим етапом автоматичного морфологічного аналізу (АМА) є …:

1) семантичний аналіз;

2) синтаксичний аналіз;

3) доморфологічний аналіз;

4) текстовий аналіз.

5. Інформація про граматичні значення в словах української мови зосереджена зазвичай …:

1) в середині слова;

2) в кінці слова;

3) на початку слова;

4) в корені слова.

6. Аналіз графемної структури словоформи може бути використаний …:

1) як інструмент ідентифікації лексико-граматичних класів у тексті та при визначенні граматичних підкласів (у межах класу);

2) як інструмент визначення флексії;

3) як інструмент визначення основи;

4) усі перераховані варіанти правильні.

7. Які проблеми лишаються при здійсненні автоматичного морфологічного аналізу (АМА) на базі словника словоформ:

1) аналіз не знайдених у словнику словоформ;

2) ототожнення різних словоформ одного й того самого слова;

3) потреба постійного поповнення й оновлення створеного словника словоформ;

4) усі перераховані варіанти.

8. Структура морфологічної таблиці передбачає …?

1) перевірку правильності членування слова на основу й закінчення;

2) перевірку правильності виділення кореня слова;

3) перевірку правильності виділення усіх афіксів слова;

4) перевірку правильності написання слова.

9. LanguageTool – відкритий програмний засіб перевірки граматики, що використовує морфологічний модуль і працює на основі …?

1) правил;

2) закономірностей;

3) аналогії;

4) протиставлення.

10. Автоматичний морфологічний аналізатор StarLing дає можливість ознайомитися з комп'ютерними базами даних за словниками …:

1) Ожегова, Ужченка і Мюллера;

2) Ожегова, Залізняка і Білоноженка;

3) Ожегова, Залізняка і Мюллера;

4) Ожегова, Івченка і Мюллера.

Середній рівень

Дайте відповіді на питання:

* *Дайте визначення морфології. Назвіть її завдання.*

* *Дайте визначення морфологічному слову. Наведіть приклади.*
* *Які етапи включали перші експериментальні системи автоматичного морфологічного аналізу, створювані в 50-60 рр. XX ст.?*
* *Від чого залежать правильність і повнота здійсненого автоматичного морфологічного аналізу?*
* *Що таке принцип навчальної вибірки при укладанні списку квазіфлексій?*
* *Що є завданням аналізу графемної структури словоформи?*
* *У чому полягає процедура автоматичного морфологічного аналізу на основі словника словоформ?*

Високий рівень

1. Здійсніть автоматичний морфологічний аналіз поданого нижче тексту, використовуючи один із методів: на основі графемного аналізу / на основі флективного аналізу / на основі словника словоформ / на основі словника основ / на основі операції логічного множення:

Одеські інженери-ентузіасти з компанії EcoFactor працюють над тим, щоб популяризувати електромобілі та створити кращі умови для переходу на екологічний транспорт. EcoFactor проєктує електрокари та зарядні пристрої для них. Засновник компанії Сергій Вельчев пояснює: в Одесі використовується багато вантажних авто, що забруднюють повітря, і щоб розв'язати цю проблему, треба вивести весь габаритний транспорт за межі міста і заборонити автівкам із двигунами внутрішнього згоряння заїжджати в центр. Ми хочемо якомога більше людей перевести на «світлу сторону», щоб вони пересувалися за допомогою електричної тяги. Для поступових змін мають бути підготовлені інфраструктура та експерти. Наше завдання – щоб багато людей зрозуміло, що таке електричний транспорт, як ним користуватися та які в цьому переваги. Команда EcoFactor викликала справжній фурор на ралі Київ – Монте-Карло 2015 року, обігнавши на модифікованому ЗАЗ-966 електрокар фірми Tesla на одній з ділянок марафону. Модернізований ZAZ Electro проїхав три з половиною тисячі кілометрів, при цьому він міг проїхати близько 500 кілометрів на одній

зарядці. Екіпаж подолав шлях через 10 країн, посівши п'яте місце в загальному рейтингу та друге – серед переобладнаних автомобілів. «Основною задачею переробленого «запорожця» було привернути увагу [до електротранспорту] і показати, що це – вже теперішнє. Чому вибрали саме «запорожець»? Щоб показати, що Україна може бути інноваційною. Щось зі старою зовнішністю може бути наповнене новим вмістом. Зі своєю задачею автомобіль впорався на 100%».

2. Здійсніть автоматичний морфологічний аналіз поданого нижче тексту, використовуючи один із ресурсів: LanguageTool / libmorphukr / AOT / StarLing / MCR DLL v2.0:

При розробленні АГАТ-морфології необхідно було вирішити лінгвістичні завдання, спрямовані на якомога глибшу формалізацію частиномовної семантики слів, а саме: 1. Аналіз і кодування частин мови з їх класифікаційними ознаками для використання в комп'ютерній морфології. В АМА визначається десять частин мови з відповідними кодами: іменники (для загальних назв – коди Й, К, Л, И; для власних – й, к, л, и); дієслова (Г); ад'єктивний клас (власне прикметники, дієприкметники, порядкові прикметники – А); займенники (прикметникові – О; іменникові – М); числівники (Ч); прислівники (Н, предикативні прислівники – @); прийменники (П); сполучники (С); частки (Ь), вигуки (В). 2. Обґрунтування і кодування морфологічних значень як компонентів словозмінних і несловозмінних категорій: – несловозмінна родова диференціація іменників оформляється спеціальними кодами: іменники чол. р. – Й; жіночого роду – К, середнього роду – Л; числово-відмінкова парадигма представлена 13-ма формами; іменники pluralia tantum (И) – шістьома формами, singularia tantum – також шістьома формами (без спеціального коду); – парадигма дієслова представлена часово-особово-числовими формами (для неминулих часів) та часово-особово-родово-числовими формами (для минулого часу) від 19 форм (для доконаного виду) до 38 (для недоконаного виду). Членами парадигми є також форми синтетичного наказового способу; аналітичні наказовий та

умовний способи встановлюються тільки на наступному, синтаксичному етапі аналізу; категорія виду є класифікаційною, тому кожний член дієслівної видової пари має свою словозмінну парадигму, як і кожний із членів станової пари; з урахуванням морфолого-синтаксичної міжрівневості категорії перехідності/неперехідності кожному дієслову присвоюється відповідна мітка, за якою на синтаксичному рівні відбуватиметься звертання до неї при встановленні синтаксичного зв'язку; – для ад'єктивного класу словозмінна родово-відмінково-числова парадигма представлена 24-ма формами; – числівник має словозмінну відмінкову шестиформну парадигму. 3. Опис формальних засобів, характерних для відповідних частин мови з їх морфологічними значеннями (списки квазіоснов і квазіфлексій). За ступенем формалізації граматичних категорій при морфологічному аналізі АГАТ-морфологію можна вважати автоматичною формально-морфологічною системою з елементами морфолого-синтаксичного аналізу. За участю в АМА лінгвіста обрано автоматичний режим індексування.

Варіант 3

Простий рівень

1. Словоформи – це …:

1) найменші самостійні й вільно відтворювані в мовленні відокремлено оформлені значущі одиниці мови;

2) слова, які однаково звучать та пишуться, але мають різне значення;

3) граматичні форми одного слова, тотожні лексично (спільне лексичне значення), але протиставлені граматичним значенням;

4) слова, досить близькі за звуковим складом і звучанням, але різні за значенням.

2. Система української мови є …:

1) аналітико-синтетичною;

2) аналітичною;

3) синтетичною;

4) суплетивною.

3. Промислові системи автоматичного морфологічного аналізу (АМА) почали створювати в …:

1) 70-80-ті рр. XX ст.;

2) 70-80-ті рр. XXI ст.;

3) 60-80-ті рр. XX ст.;

4) 50-80-ті рр. XX ст.

4. Хто є автором однієї з найґрунтовніших робіт з автоматичного морфологічного аналізу (АМА) на основі графемного принципу:

1) Іван Ковалик;

2) Катерина Городенська;

3) Іван Вихованець;

4) Герольд Бєлоногов.

5. Основним інструментом автоматичного морфологічного аналізу (АМА) на основі флективного аналізу як засобу ідентифікації граматичної інформації є …:

1) основа словоформи;

2) корінь словоформи;

3) список квазіфлексій;

4) морфемна будова словоформи.

6. На якому з етапів аналізуються диз'юнктивні коди класів і підкласів …:

1) на контекстному;

2) на доморфологічному;

3) на семантичному;

4) на морфологічному.

7. Який вид АМА є найбільш розповсюдженим?

1) на основі графемного аналізу;

2) на основі флективного аналізу;

3) на базі словника основ;

4) методом логічного множення.

8. Комбінований метод, розроблений для АМА української і російської мов, поєднує …:

1) таблицю основ і додаткову таблицю коренів;

2) таблицю основ і додаткову таблицю флексій;

3) таблицю коренів і додаткову таблицю флексій;

4) таблицю коренів і додаткову таблицю суфіксів.

9. Автоматичний морфологічний аналізатор libmorphukr призначений для …:

1) перевірки правопису окремих слів;

2) лематизації;

3) морфологічного синтезу форм за нормальною формою і граматичним описом;

4) усі перераховані варіанти правильні.

10. Автоматичний морфологічний аналізатор MCR DLL v2.0 призначений для опрацювання …:

1) російськомовних текстів;

2) англійськомовних текстів;

3) українськомовних текстів;

4) усі перераховані варіанти правильні.

Середній рівень

Дайте відповіді на питання:

* *Що є основною одиницею, предметом і об'єктом вивчення морфології?*
* *Що може виступати засобом вираження граматичних значень? Наведіть приклади.*
* *Опишіть принцип роботи однієї промислової системи автоматичного морфологічного аналізу – на вибір.*
* *Чому методика графемного аналізу виявилася доцільною для автоматичного морфологічного аналізу текстів вторинних документів (рефератів, патентів тощо)?*

* *Що таке робочий принцип укладання списку квазіфлексій?*
* *Чи відрізняються за граматичною природою підкласи різних граматичних класів? Як саме?*
* *Як у процесі автоматичного морфологічного аналізу вирішується проблема граматичної омонімії?*

Високий рівень

1. Здійсніть автоматичний морфологічний аналіз поданого нижче тексту, використовуючи один із методів: на основі графемного аналізу / на основі флективного аналізу / на основі словника словоформ / на основі словника основ / на основі операції логічного множення:

Розвиток соціальних мереж скорочує відстань між людьми. Ініціатива, що виникла в одній частині планети, може швидко поширитися світом, оминаючи кордони та бар'єри, об'єднуючи людей у віртуальні спільноти, що створюють цінності за межами інтернету. Таким «вірусним» чином ідеї та течії, знаходячи підтримку широкої аудиторії, змінюють світ. Новий тренд, що поєднав догляд за фізичною формою та піклування про чистоту планети, шириться світом. Адепти здорового способу життя та прихильники бігу почали збирати сміття під час своїх щоденних занять бігом. Така активність отримала назву «плогінг», яка утворилася поєднанням двох слів: шведського plocka upp, що означає «піднімати», та англійського jogging, тобто «повільний біг». Рух стартував у Швеції і швидко набув популярності серед прихильників здорового способу життя та небайдужих до чистоти довкілля. Як вид спорту, заняття передбачає згинання, присідання та розтягування на додачу до основної дії – бігу, походів або ходьби. Метою такого процесу є привернення уваги до проблеми забруднення планети. Ерік Альстрьом започаткував цей рух у 2016 році. Після переїзду з малого міста у Стокгольм чоловік здивувався кількості сміття довкола. Звичні пробіжки він почав поєднувати зі збиранням сміття, яке траплялося на шляху. Завдяки активності Еріка в соціальних мережах ідея швидко знайшла підтримку. Люди почали об'єднуватись у групи і займатися плогінгом разом.

2. Здійсніть автоматичний морфологічний аналіз поданого нижче тексту, використовуючи один із ресурсів: LanguageTool / libmorphukr / AOT / StarLing / MCR DLL v2.0:

АГАТ-морфологія зорієнтована на «комп'ютерний мозок», тому теоретична морфологія була скоригована стосовно можливостей автомата. По суті, це процес побудови комп'ютерної морфології української мови на алгоритмічних засадах, здійснюваний у три етапи: 1) доморфологічний (у версії 2012 р. використовується частково, є технічним, спрямованим на підготовку тексту до граматичного аналізу); 2) формально-морфологічний, або флективний; 3) контекстний. Ці етапи базуються винятково на аналітичних процесах, характерних для морфології української мови, і уможливлюють мінімізування ролі інтуїції, властивої традиційному морфологічному аналізу. Формально-морфологічний, або флективний, етап базується на поєднанні двох типів мовної інформації, систематизованої в таблицях: таблиці квазіоснов (незмінної частини словоформи або її змінної частини без флексії) і допоміжної таблиці квазіфлексій (змінної частини словоформи із флексіями). Квазіосновам приписана частиномовна і категорійна характеристика (рід, число, відмінок, особа, час), які разом становлять двочленний граматичний код. Кожній лексемі, що має словозміну, приписувався номер парадигматичного класу, який у відповідній таблиці пов'язаний із формами словозміни. Словник квазіоснов містить 210 тис. одиниць, і, відповідно, словник словоформ, породжених поєднанням інформації, взятої з таблиці основ і допоміжної таблиці, становить близько 3,2 млн. слововживань, що забезпечує морфологічну ідентифікацію словоформ аналізованого тексту практично на 97% (3% – це оказіоналізми чи форми, не унормовані граматикою української мови, або неукраїнські слова). Флективний аналіз базується на граматично унормованих формах, представлених в орфографічних словниках та лексичному фонді української мови, репрезентованому тлумачним 11-томним словником, словником іншомовних слів, частотними словниками сучасної художньої прози,

поетичної мови, наукового стилю. Усі нестандартні морфологічні форми та оказіоналізми, неологізми тощо автоматично відсортовуються й у разі доцільності вводяться лінгвістом у додатковий список, що містить форми, виведені зі складу парадигми слова як неунормовані (напр., він чита; читать – інф. тощо). Цей список підключено до АМА, тому що будь-якій словоформі тексту в подальшому повинна бути приписана граматична інформація. По суті, цей етап є прикладом взаємодії традиційної граматики і комп'ютерної граматики, чому сприяє загальна сфера морфології, предметом опису якої є форми слів.

КЛЮЧІ ДО ТЕСТІВ			
№	**Варіант 1**	**Варіант 2**	**Варіант 3**
1.	1)	2)	3)
2.	4)	1)	3)
3.	3)	4)	1)
4.	2)	3)	4)
5.	1)	2)	3)
6.	4)	4)	1)
7.	2)	4)	3)
8.	1)	1)	2)
9.	4)	1)	4)
10.	2)	3)	1)

Додаток 2

ІНДИВІДУАЛЬНА РОБОТА

На основі поданого нижче теоретичного матеріалу підготуйте реферат / презентацію / наочність.

Варіант 1

(на матеріалі статті М. Лангенбах «Стратегії й методи вдосконалення автоматичного морфологічного анотування Корпусу української мови» (http://ena.lp.edu.ua:8080/bitstream/ntb/40813/2/2017_Lanhenbakh_M-Stratehii_y_metody_vdoskonalennia_37-42.pdf))

Використання автоматичних морфологічних аналізаторів природних мов вже досить поширене у світовій практиці, їх розробка спирається на серйозне теоретичне і практичне підґрунтя, проте для жодної мови світу досі не вдалося укласти цілком досконалу систему граматичного кодування тексту. Поліпшення якості машинного морфаналізу сьогодні лишається актуальним питанням у галузі комп'ютерної лінгвістики. Увага до цієї теми зумовлена необхідністю розвитку комп'ютерного інструментарію для опрацювання текстів української мови. Збільшення обсягу україномовного матеріалу в мережі Інтернет, поступова переорієнтація сучасної науки (зокрема й мовознавства) на роботу з великими масивами даних, а також екстралінгвістичні (суспільно-політичні) чинники зумовлюють потребу в сучасних ефективних програмах для текстового аналізу. Проблема якості автоматичного морфологічного аналізу у теоретичній літературі зазвичай висвітлюється в аспектах загального огляду та оцінки ефективності різних стратегій або аналізу помилкової розмітки текстів, пов'язаної, зокрема, з мовною омонімією. Натомість до розгляду переважно не беруться випадки ігнорування аналізаторами певних лексем.

Актуальним завданням є розглянути способи підвищення ефективності роботи автоматичного морфологічного аналізатора АГАТ через зменшення

кількості неопрацьованих лексем. Розв'язання проблеми передбачало, перш за все, окреслення її меж та пошук відповідей на такі питання: 1. Наскільки ефективний на сьогодні морфологічний аналізатор? Чи залежить його ефективність від характеру аналізованих текстів (функціональний стиль, жанр, тематика)? 2. Якщо для різних текстів система демонструє різну якість опрацювання матеріалу, то якими чинниками це пояснюється? 3. Які явища є типово проблемними для автоматичного морфологічного опрацювання? 4. Які способи вирішення виявлених проблем?

Помилкам кодування омонімічних форм у науковій літературі приділено чимало уваги, проте існує ще один аспект – певний відсоток слів у ході опрацювання морфологічним модулем не отримують граматичного коду взагалі. Стандартно проблемними об'єктами для морфологічних аналізаторів є нехарактерні для мови символи, зокрема слова, написані іншою абеткою (Curiosity зробив нові приголомшливі знімки). Крім того, морфологічні аналізатори словникового типу можуть не опрацьовувати такі групи лексики: 1) рідковживані лексеми, зокрема галузеві терміни, діалектизми та застарілі слова, власні назви (фітофаг, віргінільний, скарамангія, Вальденфельс, Ерзурум) тощо; 2) неологізми (у т. ч. авторські) та іншомовні слова, не зафіксовані у словнику (знепримітниться, твітер, флешмоб); 3) словоформи в нестандартному записі: а) скорочення (гром., ред., ст., МАКтє, ЛСП); б) друкарські помилки; в) цифровий запис фрагментів словоформ (12-ох, 11-томний, 5-кратний); г) складні слова, написані через дефіс (дворянсько-поміщицькі, два-три, бак-акумулятор); д) приклади т. зв. «авторської орфографії» (переважно характерні для художніх текстів: до-о-овгий, ссстій, багатиє, везють). Неповне або некоректне морфологічне опрацювання тексту призводить до помилок на подальших етапах роботи автоматичної граматики. Так, через відсутність граматичних позначок для деяких словоформ унеможливлюється контекстне коригування омонімічних кодів у зв'язаних із ними текстових одиниць. Прогалини у морфологічній розмітці зумовлюють некоректну роботу синтаксичного аналізатора: через пропуск зв'язків у

словосполученнях і реченнях будуються неповні синтаксичні схеми. З наведеної інформації очевидно, що пропуск елементів під час морфологічного кодування тексту є серйозною проблемою, яка помітно впливає на якість граматичного аналізу в цілому. Вирішувати це завдання можна рухаючись у кількох напрямках. Зокрема, однією зі стратегій є поліпшення доморфологічного аналізу. На цьому етапі роботи машина повинна, по-перше, коректно визначити межі слів (якими типово є пробіли, проте є і чимало винятків з цього правила); по-друге, виявити в тексті елементи, потенційно складні для подальших модулів опрацювання тексту (морфологічного, синтаксичного, семантичного тощо). Інша стратегія передбачає подолання словникової обмеженості аналізатора. Використання словників дає змогу досягти більшої ефективності опрацювання порівняно з несловниковими підходами, проте приводить до того, що слова, не зафіксовані у словнику, ігноруються системою. Для того щоб зменшити кількість неопрацьованих лексем, необхідно навчити аналізатор певним чином реагувати на позасловникові елементи. Для аналізу основних проблем у роботі морфологічного модуля системи АГАТ було сформовано тестову вибірку розміром 1,2 млн слововживань (на матеріалі Корпусу української мови). Вибірка складалася з трьох частин, що репрезентували різні функціональні стилі (художня література, публіцистика та наукові тексти) і містили по 4000 текстових фрагментів, кожен обсягом 1000 слововживань. Після застосування до вибірки процедури автоматичного морфологічного аналізу було укладено реєстр неопрацьованих одиниць. Його обсяг становив 39841 слововживання (3,32% від загального розміру вибірки). Кількість та склад неопрацьованих лексем для підвибірок дещо відрізнялися. Сумарно найменшу кількість проігнорованих програмою лексем було зафіксовано в публіцистичній підвибірці, що можна пояснити порівняно нейтральним характером лексики, яку добирають для текстів цього стилю, а також дотриманням стандартів орфографії. Найбільша кількість неопрацьованих одиниць виявилася в наукових текстах. Кількісний розподіл матеріалу за

виділеними типами дає можливість детальніше проаналізувати, які саме одиниці спричиняють помилки в роботі програми з текстами певних стилів. Так, у наукових і публіцистичних текстах зафіксовано велику кількість скорочень, натомість у художніх текстах їх значно менше. Водночас у художній прозі набагато вищим є відсоток слів з нестандартною орфографією, тоді як для публіцистики такого типу помилки є найменш характерними. Це підтверджує нашу гіпотезу про те, що стандартизованість написання слів у публіцистичних текстах є причиною вищої якості морфологічного аналізу на матеріалах цього стилю. На наступному етапі роботи необхідно було виявити типові ознаки, що дозволили б охарактеризувати кожну з наведених категорій помилок та розробити стратегію їхнього опрацювання. Ручне поповнення словника для більшості типів нерозпізнаних лексем не дало б вагомого результату, оскільки, за законом переваги співвідношення асортименту подібних одиниць у мові із частотою їхнього вживання зумовлює неефективність такого методу (це призвело б до збільшення обсягу бази за рахунок одиниць, потреба в яких незначна, а потенційна кількість таких елементів майже безмежна). Ілюстрацією цієї тези може бути список абревіатур, отриманих із нашої експериментальної вибірки: з 609 унікальних одиниць 142 одиниці не зафіксовані навіть у найсучаснішому і найдинамічнішому джерелі з цієї тематики – інтернет-словнику абревіатур www.ukrskor.info (МЕТ – маркетинг елітних товарів, МІКУ – музей історії коштовностей України, МКЗН – міжнародний кодекс зоологічної номенклатури, НКРЕКП – Національна комісія з регулювання у сферах енергетики і комунальних послуг, ЗНС – загони народної самооборони тощо). До того ж, словниковий підхід актуалізував би питання мовного статусу ненормативної орфографії (друкарських помилок і авторського правопису). Тому було вирішено звернутися до інших способів визначення морфологічних характеристик словоформ, зокрема графічного та графемного. Графемний аналіз дає можливість ідентифікувати класи проблемних словоформ, що містять нехарактерні для української мови літерні послідовності. Так, технічно

нескладним завданням є відсіювання словоформ, що включають неукраїнські символи. Порівняння графемного складу слова із символьним набором української мови уможливлює виявлення таких одиниць зі 100% точністю (проте такий підхід не дозволяє виявити іншомовні слова, що містять лише спільні з українською мовою символи). Також цілком очевидні ознаки мають деякі типи рідковживаної лексики, зокрема абревіатури: збіг трьох і більше голосних, чотирьох і більше приголосних, довжина словоформи у скороченнях типу ст., нм. За графічними особливостями можна ідентифікувати такі типи неопрацьованих одиниць: власні назви (написання з великої літери); ініціальні абревіатури (написання усього слова великими літерами); складні слова та словоформи із буквено-цифровим записом (написання через дефіс). Для подібних випадків було укладено формалізовані моделі та розроблено правила їх застосування (зокрема, на базі регулярних виразів). Проте підхід, що базується на правилах, хоча й долає до певної міри обмеження словникового аналізу, має меншу точність. Так, наприклад, умови «збіг 4-х і більше приголосних» та «довжина не більше ніж 6 символів» задовольняє, скажімо, слово вогнлх ‘вогнях’, яке є друкарською помилкою. Також розроблені шаблони не завжди дозволяють надати точну й повну морфологічну інформацію про слово. Зокрема, слова, кваліфіковані як власні назви, та іншомовні слова (за винятком римських цифр) умовно позначаються кодом іменника (що є статистично обґрунтованим, але, тим не менш, продукує певний відсоток помилок), проте визначити їхню відмінково-часово-родову форму без застосування додаткових процедур (контекстного або імовірнісного аналізу, графемного аналізу за методом квазіфлексій тощо) неможливо. У деяких випадках правила не забезпечують і точної ідентифікації частиномовної належності. Наприклад, словам із цифровим записом надається омонімічний код «іменник–прикметник–числівник»; словоформи, записані через дефіс, програма у процесі аналізу розбиває на окремі частини і на виході приписує слову ланцюжок кодів усіх їхніх складників. Уточнення цієї морфологічної інформації має здійснюватися на наступному етапі роботи

модуля – в ході контекстного аналізу. Окремої уваги в цьому класі одиниць потребують складні слова типу науково-технічний, шапка-вушанка, пекучо-беззахисне, оскільки проблемним є загальне питання «розуміння» їх машиною – як двох різних слів чи як однієї лексеми, що суттєво впливає на побудову морфологічної парадигми таких слів та використання їх у подальшій роботі – на наступних етапах опрацювання тексту, під час пошуку за морфологічно розміченим корпусом і т.п. Крім того, вагомим чинником, що впливає на якість опрацювання проблемних одиниць, є порядок застосування розроблених правил. Так, діагностика словоформ із буквено-цифровим записом і неукраїнськими символами доцільна на етапі доморфологічного аналізу, тоді як опрацювання власних назв і абревіатур через «шумність» алгоритмів вимагає попередньо відсіяти всі словникові одиниці, що можуть бути помилково інтерпретовані. Це, з одного боку, пришвидшить процес роботи допоміжних алгоритмів (оскільки кількість аналізованих одиниць суттєво зменшиться після застосування АМА), з іншого ж, покращить якість опрацювання:

Схема роботи морфологічного аналізатора із застосуванням допоміжних процедур

Результати опрацювання текстового матеріалу (за підбірками)

	Публіцистика			Художні			Наукові		
	опра-цьова-ні	неоп-рацьо-вані	к-ть помилок (%)	опра-цьова-ні	неоп-рацьо-вані	к-ть помилок (%)	опра-цьова-ні	неоп-рацьо-вані	к-ть помилок (%)
буквено-цифровий запис	381	0	25 *(6,56)*	36	0	2 *(5,56)*	280	0	30 *(10,71)*
скорочення	2374	52	145 *(6,11)*	119	3	18 *(15,13)*	1135	7	23 *(2,03)*
власні назви	2263		108 *(4,77)*	4416		872 *(19,75)*	3007		433 *(14,40)*
неукраїнські символи	592	0	0	779	353	0	3711		0
дефісний запис	924	46	8 *(0,87)*	718	130	33 *(4,60)*	1123	129	75 *(6,68)*

Як видно, ефективність роботи допоміжних модулів коливається залежно від типу неопрацьованих словоформ та функціонального стилю текстів. Найкращих результатів вдалося досягти у виявленні одиниць із цифровими фрагментами та слів, написаних неукраїнською абеткою, також досить високі показники продуктивності продемонстрували правила діагностики абревіатур та власних назв. Натомість поки що не вироблено успішної стратегії аналізу випадків нетипової орфографії, рідковживаних слів та неологізмів, які не є власними назвами. Опрацювання таких класів словоформ потребує проведення подальших досліджень та використання складніших алгоритмів аналізу. Також проблемним лишається питання автоматичного кодування складних слів, записаних через дефіс, та друкарських помилок (зокрема помилок розпізнавання тексту). Різне кількісне співвідношення типів неопрацьованої лексики в межах різних функціональних стилів зумовлює неоднаковий результат роботи допоміжних процедур на відповідних текстових зразках. Так, значна частка скорочень, слів із неукраїнськими символами і буквено-цифровим записом у публіцистичному та науковому підкорпусах забезпечують вищу ефективність роботи на текстах цих стилів. На противагу їм, художній стиль містить велику кількість рідковживаної лексики та зразків авторської орфографії, які важко піддаються опрацюванню.

Проведений експеримент засвідчив, що, незважаючи на якісні переваги використання словникового морфологічного аналізу, є певні групи лексики і певні категорії текстового матеріалу, що вимагають застосування інших, несловникових методів. Такими, зокрема, є тексти, що містять значний відсоток лексем у нестандартному записі та рідковживаних слів. Відповідно, оптимізація роботи морфологічного аналізатора передбачає часткову відмову від уніфікованого підходу. Шляхом комбінування різнопланових методик видається можливим підвищити ступінь покриття матеріалу аналізатором в середньому на 1,5%. Ефективність застосованої стратегії виявилася неоднаковою для текстового матеріалу різних функціональних стилів. Найкращих результатів вдалося досягти для публіцистики, дещо менших – для наукових текстів; найгірше піддаються опрацюванню тексти художньої літератури. Крім того, експеримент засвідчив, що кожен клас лексики, «проблемної» для системи, вимагає специфічної стратегії опрацювання. Так, власні назви, абревіатури й лексеми, що містять у своєму складі цифри, можна ідентифікувати за графічними ознаками. Слова з нетиповим для української мови графемним складом (збігом великої кількості голосних або приголосних) досить вдало опрацьовуються методом графемного аналізу. Однак запропоновані методи не завжди дозволяють отримати точний результат. Загалом же, попри очевидно позитивний результат роботи варто визнати, що робота над деякими типами помилок потребує детальнішого розгляду та залучення складніших методик опрацювання текстового матеріалу.

Варіант 2

(на матеріалі статті О. Шипнівської «Розробка процедури лематизації словоформ сучасної української мови» (https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxwbGRvbm51fGd4OjM1NDY4N2JkYmQ1NzYyMzE))

Невід’ємною складовою будь-якої системи автоматичної обробки мови – системи автоматичного індексування, системи машинного перекладу –

є програма лематизації словоформ, що входить до модуля автоматичного морфологічного аналізу. Цей компонент лінгвістичного процесора передбачає зведення текстової словоформи мови до її початкової, словникової форми. Необхідність підвищення ефективності автоматичного аналізу природномовного тексту, нові прикладні завдання змушують дослідників повсякчас шукати кращі підходи та можливості їх реалізації й щодо цього, здавалося б, давно вирішеного питання, на що вказують розвідки останніх років. У своїй праці пропонуємо процедуру лематизації української словоформи на основі словника квазізакінчень. Правила побудовані відповідно до розроблюваної знання-орієнтованої системи машинного перекладу. Зазвичай метод лематизації текстових словоформ залежить від обраної дослідниками моделі комп'ютерної морфології та від конкретних прикладних завдань. Найпоширенішим у системах машинного перекладу для мов флективного типу є алгоритмічний метод автоматичного морфологічного аналізу, за яким текстова словоформа за відповідними таблицями основ та закінчень отримує відповідну морфологічну інформацію. У разі зворотної процедури текстова словоформа, синтезована із двох таблиць, дає змогу за відповідним кодом вивести вихідну форму. Так, запропонований для української та російської мов алгоритм лематизації передбачає виділення за певними умовами основи та закінчення словоформи та визначення початкової форми в межах граматичного класу (на текстах, де словоформи мають заздалегідь визначені граматичні класи). Зведення парадигм, на думку В. Перебийніс, включає в себе кілька часткових завдань: виявлення флексії й відокремлення її від основи слова; встановлення варіантів основи (якщо такі є); об'єднання всіх форм слова в одну групу (парадигму); виділення або реконструкція словникової форми слова. В алгоритмі відповідно до кількості парадигматичних класів виділяємо блоки лематизації різних частин мови. Розроблювана концепція знання-орієнтованої системи машинного перекладу передбачає створення гнучкої та відкритої щодо різних предметних галузей комп'ютерної морфології. Для ефективності нами був обраний метод

морфологічного аналізу на основі словника квазізакінчень, який і слугує базою процедури лематизації українських словоформ. Словник квазізакінчень було побудовано на основі словника словоформ, що нараховує понад 32 тис. одиниць. Джерельною базою словника слугували тексти військово-спеціальних та військово-гуманітарних наук різних жанрів та матеріали Internet. Словникова стаття словника квазізакінчень має такий формат:

$$<\{F\}_n^1 \quad K_l * \{K_g\}>,$$

де $\{F\}_n^1$ – ланцюг літер вхідного слова, розташований в інверсному порядку, n – остання літера в слові, довжина ланцюга в загальному випадку може дорівнюватися довжині вхідної словоформи. K_l – код лексико-граматичного класу вхідної словоформи, K_g – множина кодів значень граматичних категорій, які визначаються для K_l.

Основою пропонованого принципу автоматичного морфологічного аналізу є розроблений метод, який спирається на позиційно-цифрове кодування граматичної інформації в словниковій статті. У такий спосіб кожна аналізована словоформа отримує свій код, який містить інформацію про її частиномовну приналежність та конкретне граматичне значення. Саме словник квазізакінчень і став основою лематизаційного словника. Словникова стаття цієї лексикографічної системи містить в інверсному порядку квазізакінчення вхідної словоформи, множину кодів лексико-граматичних класів, словоформи яких можуть мати таке квазізакінчення, та множину квазізакінчень вихідних форм. Для словоформи визначається така кінцівка графем (квазізакінчення), за якою їй можна приписати граматичну інформацію в межах певного лексико-граматичного класу. Максимальна довжина квазізакінчення може відповідати довжині слова. Процедура будується як для однозначних, так і для омонімічних класів. Найчастіше в словнику зустрічаємо граматично неоднозначні квазізакінчення.

Фрагмент лематизаційного словника

від 1*12/д/1*22/дова/1*42/дя/16*11/дів/дова/дове/
відор 1*12/рід/
відора 1*12/арод/
відох 1*12/хід/
віж 1*12/ж/
віже 1*12/іж/
віз 1*12/з/16*11/зів/
вік 1*12/к/
вікв 1*12/вок/
віквов 1*12/вовк/
вікд 1*12/док/1*22/дкова/
вікж 1*12/жок/
вікз 1*12/зок/
вікй 1*12/йок/
вікл 1*12/лок/
вікм 1*12/мок/

Алгоритм лематизації працює так: 1. В аналізованому слові виділяється квазізакінчення. 2. Визначене квазізакінчення відсікається. 3. Код граматичної інформації квазізакінчення зіставляється із лематизаційним словником. 4. У випадку знаходження тотожного запису до аналізованого слова (отриманої основи) додається квазізакінчення початкової форми, розміщеному в словнику. Так, відповідно до закодованої лексико-граматичної інформації, за словниковою статтею

*від 1*12/д/1*22/дова/1*42/дя/16*11/дів/дова/дове/*

виводиться словникова форма для іменника *приладів* (*1*12/д/*) – *прилад, вдів* (1*22/дова/) – *вдова, суддів* (/1*42/дя/) – *суддя,* присвійного прикметника *дідів* (16*11/дів) – *дідів.*

У таблиці подано фрагмент словника квазізакінчень з кінцевою графемою в. Загалом, у словнику зафіксовано 108 словникових статей з цією кінцевою графемою. Наприклад, за словниковими статтями

*вікй 1*12/йок/,*
*вікл 1*12/лок/,*
*вікм 1*12/мок/*

можна однозначно синтезувати вихідні лексеми словоформ *буйків, білків, димків – буйок, білок, димок*. Часто для генерації канонічної форми до словникової статті додається вся словоформа:

*вілотс 1*12/стіл/*

*вілс 1*32/слово/.*

Інколи словникова стаття подає омонімічну словоформу щодо різних лексем:

*мокнат 1*11/танок/танк/* або ж *вікотив 1*12/витік/виток/.*

У розроблюваній нами процедурі початкові форми визначаємо так. Для іменників вихідною формою є називний відмінок однини, для множинних іменників – називний відмінок множини. Для всіх форм дієслова (а в системі окремо визначаються дієслова неминулого часу та минулого часу, дієслова наказового способу та інфінітив) лемою виступає неозначена форма. Початковою формою числівників є називний відмінок однини. Для займенників початковою формою є називний відмінок чоловічого роду однини. Для прикметників та дієприкметників з урахуванням потреб машинного перекладу початковою формою виступає граматична форма називного відмінка однини чоловічого, жіночого та середнього родів; крім того, для дієприкметника як особливої форми дієслова ще й додається початкова форма відповідного йому дієслова. Так, для дієприкметникової словоформи куплена за словниковою статтею

*анел/12*21/лена/лене/ленй/ити*

буде синтезовано леми – купити, куплений, куплена, куплене.

Простота та, як наслідок, висока швидкість аналізу запропонованої процедури забезпечено мінімальною інформацією, що подається в словниковій статті, та завдяки обраному підходу на основі словника квазізакінчень. Переваги запропонованої нами процедури визначення початкової форми вбачаємо у швидкості та прозорості алгоритму та в можливості опрацювання нових, невідомих для системи слів, що робить її відкритою стосовно різних предметних галузей.

Варіант 3

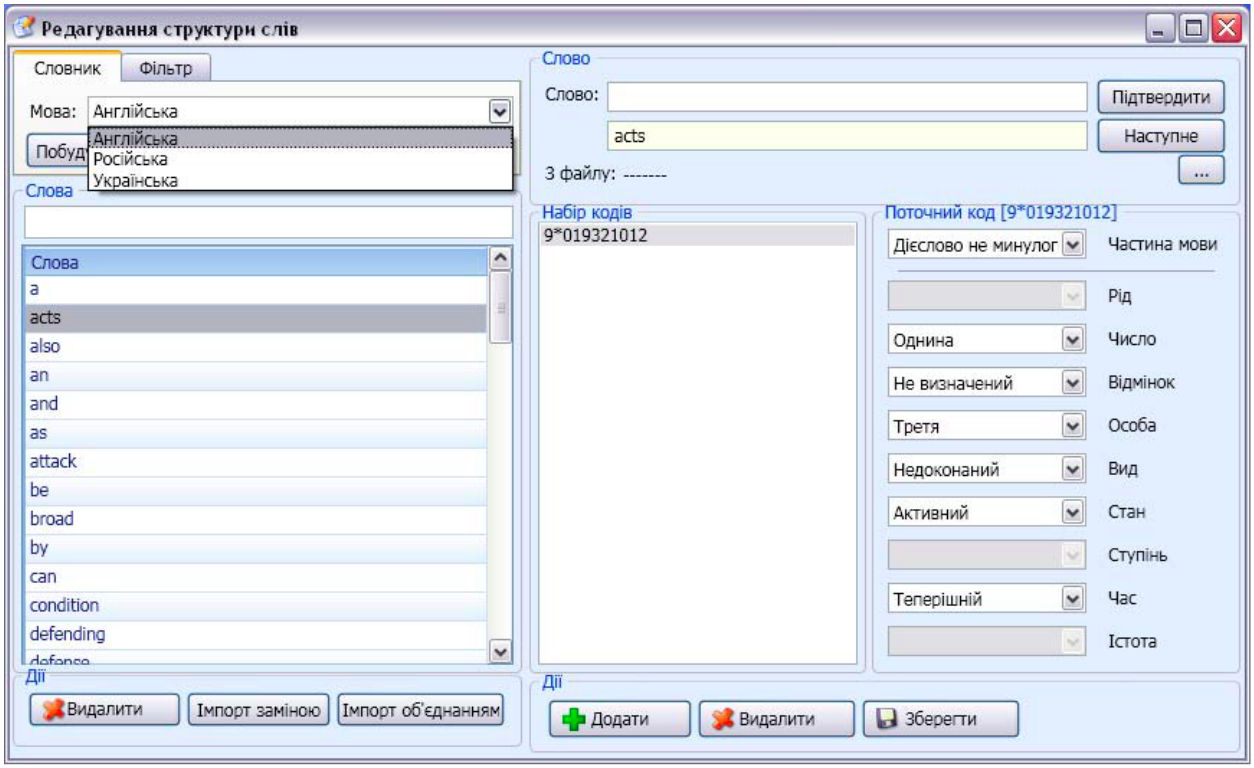
(на матеріалі статті О. Ніколаєвського «Автоматизація укладання компонентів лінгвістичного забезпечення модуля автоматичного морфологічного аналізу різномовних текстів» (http://www.irbis-nbuv.gov.ua/cgi-bin/irbis_nbuv/cgiirbis_64.exe?I21DBN=LINK&P21DBN=UJRN&Z21ID=&S21REF=10&S21CNR=20&S21STN=1&S21FMT=ASP_meta&C21COM=S&2_S21P03=FILA=&2_S21STR=VKNU_vsn_2012_28_10))

Відповідні словники, що забезпечують опрацювання вхідного тексту на морфологічному рівні мовної системи, входять до складу лінгвістичної бази даних і потребують спеціалізованого інструментарію для їх розробки відповідно до сучасного рівня лінгвістичних знань та інформаційних технологій. Аналітичні граматичні словники (АГС) є невід'ємним компонентом лінгвістичного забезпечення системи машинного перекладу (СМП), які призначені для автоматичного морфологічного аналізу вхідного тексту. Словникова стаття АГС містить в собі інформацію щодо частини мови та відповідних цій частині мові граматичних категорій (відмінок, рід, число, час, особа тощо). Структура словникової статті залежить від підходу до автоматизації морфологічного аналізу зокрема в СМП. Аналіз сучасних систем машинного перекладу показав три принципово різних підходи щодо організації аналітичних граматичних словників. Лінгвістичне забезпечення за одним із підходів може формуватися на основі словника словоформ, але такий словник повинен мати не менше 0,5 млн. слів для флективних мов (російська, українська) і близько 150-200 тис. слів для аналітичних мов (англійська). Формування такого словника є дуже трудомістким процесом, який потребує багато часу та людських ресурсів, хоча для англійської мови він вважається прийнятним. Крім того, при цьому підході СМП не зможе аналізувати нові слова. Другий підхід – формування АГС на основі словника квазіоснов та таблиці закінчень. Обсяг такого словника складає близько 100 тис. словникових одиниць, що теж є достатньо великою кількістю та потребує

значного часу на його складання. За структурою словникова стаття такого АГС містить квазіоснову+відповідний словозмінний клас. Недоліком цього підходу також є неможливість автоматичного морфологічного аналізу нових слів, тобто слів які не належать до однієї парадигми. Третій підхід визначає на основі словника квазізакінчень всі закономірності словозмінення для певної вхідної даної мови. Даний підхід дозволяє при обсязі АГС близько 5 тис. словникових одиниць аналізувати всі слова певної мови, включаючи і нові слова, що не можливо при інших підходах. Отже, доцільно будувати АГС на основі словника квазізакінчень в якості компоненти лінгвістичного забезпечення для автоматизації морфологічного аналізу в СМП. Мета даної роботи – висвітлити досвід автоматизації формування компонентів лінгвістичного забезпечення для автоматичного морфологічного аналізу.

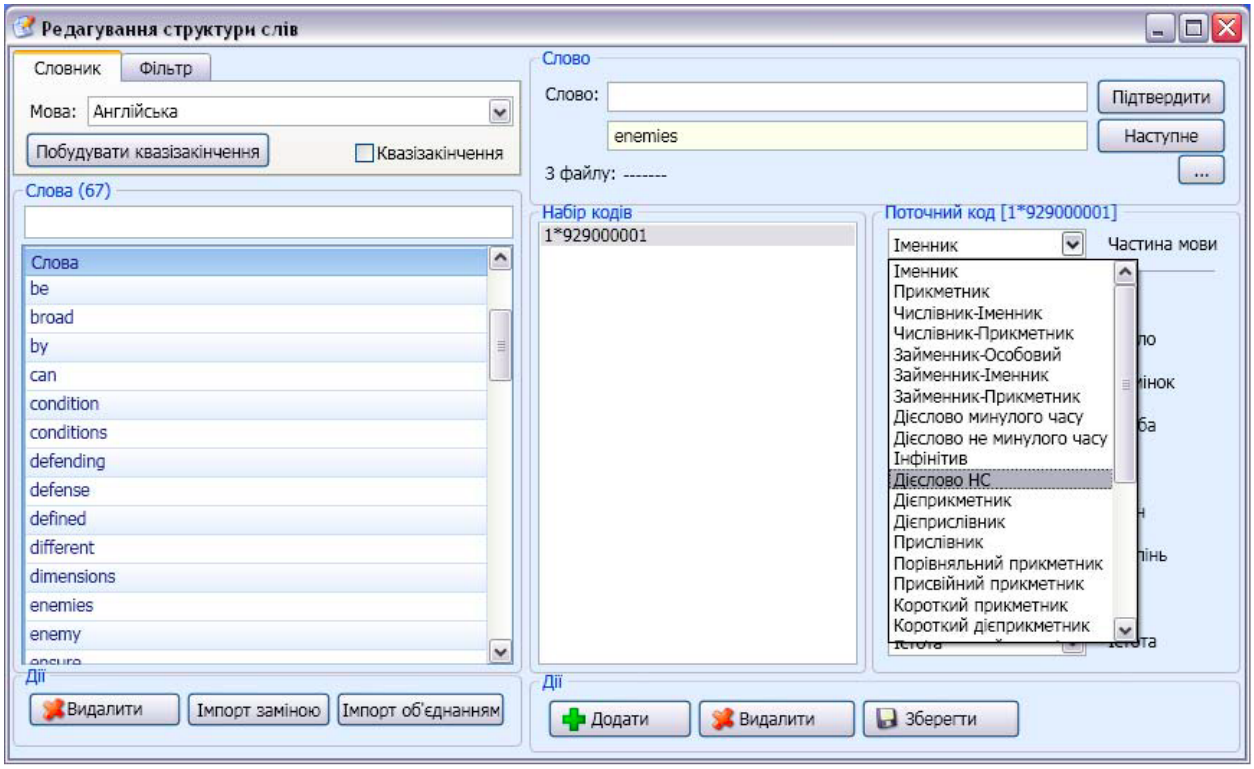
Автоматизація формування АГС значно може скоротити час підготовки компонентів лінгвістичного забезпечення, але для цього потрібно розробити спеціальні інструментальні засоби. Таким спеціальним засобом ми пропонуємо автоматизоване робоче місце лінгвіста (АРМ «ПАРАДИГМА»), яке призначено для формування аналітичних словників розпізнавання текстових одиниць на морфологічному рівні. Така віртуальна лінгвістична лабораторія розрахована на фахівця-мовознавця й надає необхідні засоби для швидкого й високоякісного створення аналітичних словників. У розроблюваній СМП АРМ «ПАРАДИГМА» виступає як окрема система, яка дає змогу фахівцю-лінгвісту оптимізувати розроблення компонентів лінгвістичного забезпечення, забезпечити його функціональну повноту. Результатом роботи АРМ «ПАРАДИГМА» є: 1) словник службових слів для відповідної вхідної мови; 2) словник квазізакінчень для відповідної вхідної мови. АРМ лінгвіста підтримує формування словника словоформ для 3-х мов: української, російської, англійської. З цією метою розроблено єдину систему кодування для зазначених мов.

Інтерфейс системи АРМ «ПАРАДИГМА»



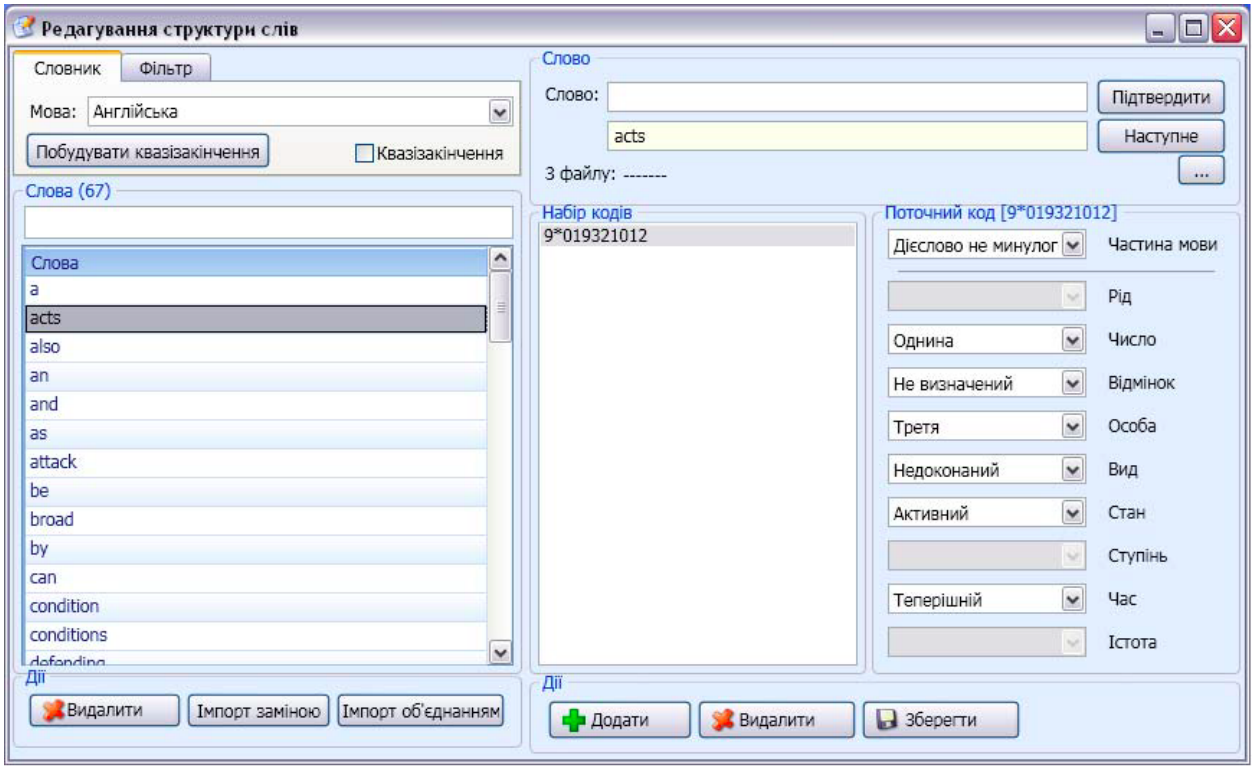
Класифікація, покладена в основу автоматичного морфологічного аналізу (АМА), зорієнтована на те, що результати слугують вихідними даними для автоматичного синтаксичного, лексичного та семантичного аналізів декількох мов. Список граматичних класів прийнятої в роботі класифікації складається із 27 виділених груп (лексико-граматичних класів). Звісно, виокремлюємо традиційні частини мови: іменник, дієслово, прикметник, числівник, прийменник, сполучник та ін. Визначаємо: артикль, герундій. Уведені зміни до АРМ «ПАРАДИГМА» стосуються класів: дієслова – в окрему групу виділяємо дієслова минулого часу, неминулого часу, інфінітив, вказуючи на особливості керування, дієслова наказового способу; в окремі групи виділяються дієслівні форми – дієприкметник та дієприслівник; числівника (виокремлено числівник-іменник, числівник-прикметник); для займенників ураховується як характер їх значень, так і специфіка словозміни та функціонування в мові. Відповідно, розглядаються займенники, що мають прикметниковий тип відмінювання (деякий, кожний, всякий), займенники-іменники (вона, він), особові займенники; для прийменник додається лексико-граматична категорія керування. Щоб сформувати словник квазізакінчень,

який би забезпечував апріорну повноту, необхідно набрати достатню вибірку словоформ з текстів певної вхідної мови. З цією метою до АРМ ведено функції: формування словника словоформ за текстом, при цьому якщо дана словоформа вже є словнику АРМ, то вона не завантажується до вікна користувача; формування словника словоформ безпосередньо набором із вікна користувача. Це зручно, коли потрібно ввести до базового словника словоформи, що виключенням з певного правила. Отриманий словник словоформ формується за текстами різної жанрово-тематичної спрямованості. У разі необхідності залучались дані з граматик відповідних мов. Основою пропонованого принципу є розроблений метод, який спирається на позиційно-цифрове кодування граматичної інформації в словниковій статті. У такий спосіб кожна аналізована словоформа отримує свій код, що містить інформацію про її частиномовну приналежність та конкретне граматичне значення. При введенні нового слова лінгвіст визначає йому лексико-граматичний клас:



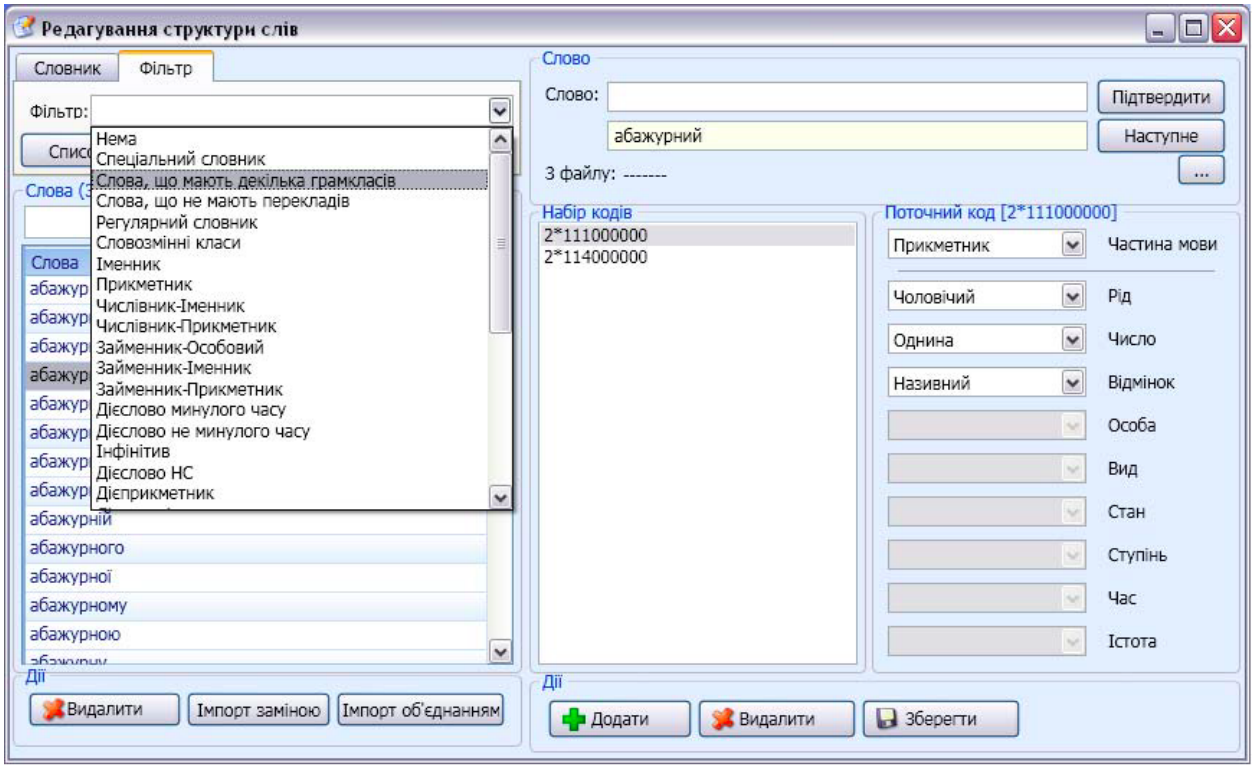
Для дослідницьких цілей в АРМ передбачено автоматичну побудову словників (на основі базового словника): словника службових слів; оберненого словника, який є основою для побудови синтетичних словників

квазізакінчень: лематизаційного і парадигматичного; окремих словників для кожного лексико-граматичного класу, це дає змогу зручного корегування і виправлення помилок при визначенні кодів оператором; словника квазізакінчень. Повнота словника квазізакінчень визначається експериментальним шляхом. Вважається, що словник є повним, якщо при додаванні нових словоформ до базового словника, словник квазізакінчень не змінюється. Після визначення лексико-граматичного класу автоматично з'являються лише ті граматичні категорії, які визначаються для даного лексико-граматичного класу, інші граматичні категорії заблоковані. Це дозволяє уникнути зайвих помилок при кодуванні граматичної інформації з боку лінгвіста:

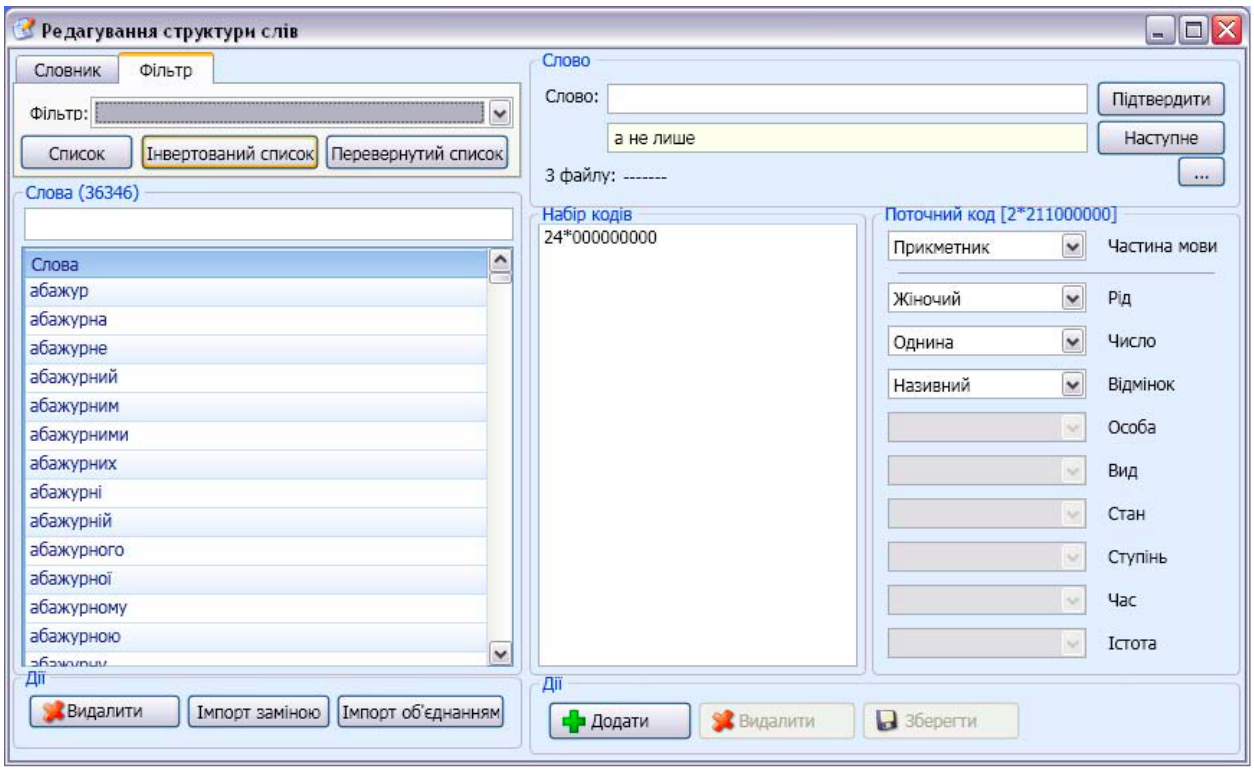


Крім того, АРМ має можливість: фільтрації даних за різноманітними граматичними ознаками; експорту даних в текстовий формат за прямим та інвертованим впорядкуванням; імпорту даних з різних текстових форматів (що допомагає лінгвісту вводити та перевіряти дані). Для дослідницьких цілей в АРМ передбачено автоматичну побудову підсловників на основі базового словника; словник службових слів, до даного словника входять всі службові частини мови; низка словників за окремими частинами мови. Це є зручною

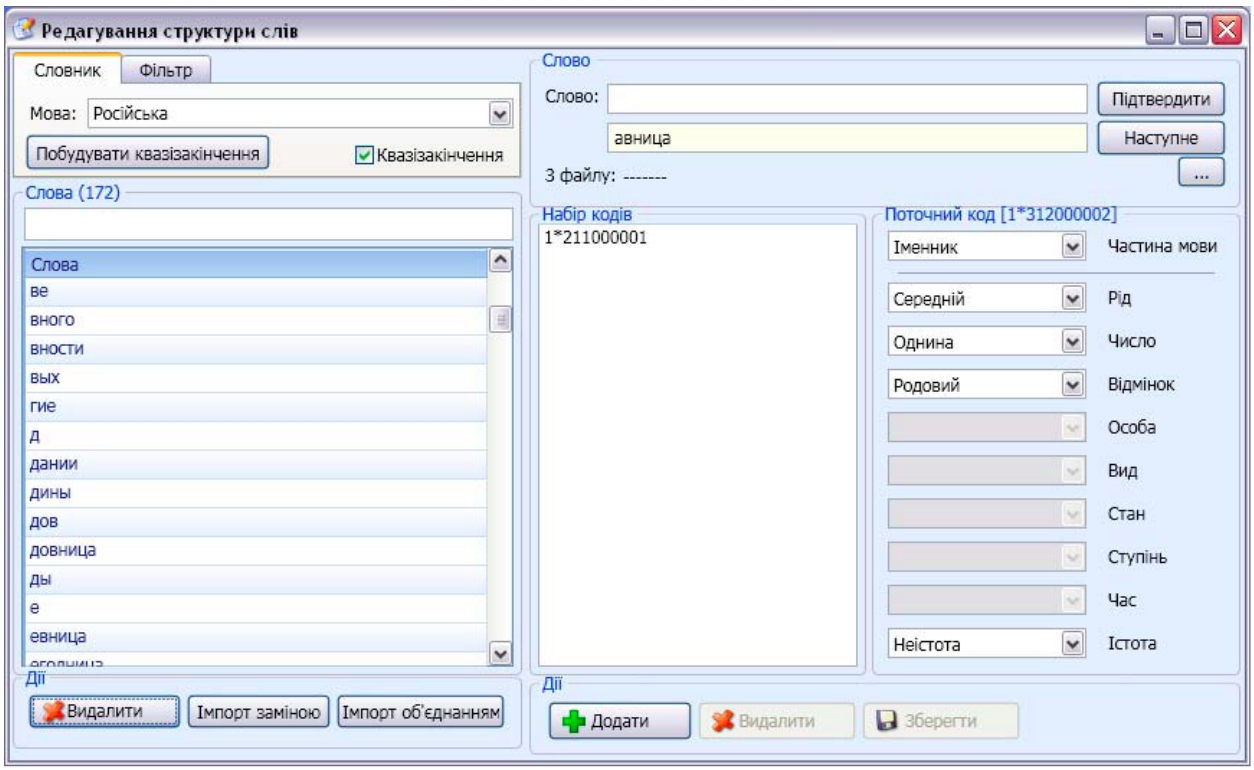
функцією при редагуванні базового словника, оскільки для словозмінних частин мови можна відслідкувати всю парадигму слова: перевірити коректність коду, додати відсутні словоформи тощо; словник словоформ, що мають декілька граматичних класів, тобто словник слів, яким притаманна міжчастинно-мовна омонімія:



Крім того, АРМ має такі функції, як: автоматичне формування базового словника в прямому алфавітному порядку (режим СПИСОК), з подальшим його виведенням в текстовий формат; автоматичне формування базового словника в зворотному алфавітному порядку, тобто з кінця слова (режим ІНВЕРТОВАНИЙ СПИСОК); автоматичне формування базового словника, в якому слова представлені в перевернутому форматі (режим ПЕРЕВЕРНУТИЙ СПИСОК). Перераховані словники є основою для побудови синтетичних словників квазізакінчень: лематизаційного і парадигматичного:



Як вже зазначалося, кінцевим результатом є автоматичне формування словника службових слів та словника квазізакінчень (режим КВАЗІЗАКІНЧЕННЯ), які безпосередньо представлені у вигляді аналітичних граматичних словників (АГС) як складових лінгвістичного забезпечення автоматичного морфологічного аналізу. Словник квазізакінчень будується на основі базового словника для кожної вхідної мови. На малюнку у правому віконці представлений фрагмент квазізакінчення для російського базового словника словоформ. Для автоматичної побудови словника квазізакінчень необхідно вибрати мову та запустити процедуру ПОБУДУВАТИ КВАЗІЗАКІНЧЕННЯ. Крім того, АРМ дає можливість експорту даних в текстовий формат за прямим та інвертованим впорядкуванням (режим ЕКСПОРТ); можливість імпорту даних з різних текстових форматів (режими ІМПОРТ ЗАМІНОЮ та ІМПОРТОБ’ЄДНАННЯМ). Введення цих режимів дозволяє розпаралелити роботу фахівців-лінгвістів по різним робочим місцям з подальшим об’єднанням результатів їх роботи:



Власне для автоматизації роботи фахівця-лінгвіста пропонується АРМ (автоматичне робоче місце) лінгвіста, особливістю якого є уніфіковане представлення граматичної інформації для трьох мов. З цією метою визначаються частини мови та інші морфологічні характеристики (герундій, артикль, допоміжне дієслово, тощо), що охоплюють перелік можливих характеристик для всіх мов вказаної групи. 1. Оцінка ефективності системи автоматичного морфологічного аналізу (АМА) залежить від обсягів словника, що застосовується, несуперечливості інформації, швидкості опрацювання текстів і можливості аналізу нових слів. 2. Розроблене автоматизоване робоче місце лінгвіста дозволяє максимально оптимізувати розробку аналітичного словника для автоматичного морфологічного аналізу, а саме: застосований принцип флективного аналізу на основі позиційно-цифрового кодування допомагає зменшити обсяги словника квазізакінченість, забезпечує компактність збереження лінгвістичних даних, а відтак дозволяє підвищити ефективність системи автоматичного морфологічного аналізу; робить АМА відкритим, оскільки уможливлює аналіз «невідомих для системи» нових слів; способи поповнення бази словоформ – «уведення слова вручну» та «опрацювання повнотекстових масивів» – надає морфологічній моделі змісту,

який може відповідати лінгвістичній реальності. 3. Поповнення словника словоформ – як дослідного масиву (базового словника) для формування списку квазіфлексій – завдяки створеному АРМ лінгвіста можна здійснювати двома шляхами: безпосередньо вводячи словоформу вручну або ж за допомогою текстового файлу. Така можливість дозволяє спиратися як на знання мови, скажімо, вводячи унікальні класи словозміни вручну, так і на реальні тексти, які дозволяють реалізувати знання-орієнтовний підхід і до формування морфологічної моделі мови. 4. Єдина система параметризації граматичної інформації дозволяє за допомогою одного АРМ (програмного модуля) опрацьовувати англомовні, російськомовні та україномовні тексти, що є однією з головних вимог до розробки багатомовної системи машинного перекладу.

Додаток 3

СЛОВНИК ТЕРМІНІВ З ПРИКЛАДНОЇ МОРФОЛОГІЇ

Автоматичний морфологічний аналіз (АМА) – початкова частина автоматичного опрацювання текстів; аналіз слів, застосовуваний з метою їх членування на морфеми або сполучення морфем та отримання граматичної інформації, необхідної на наступних етапах опрацювання (Герольд Бєлоногов); визначення леми (базової, канонічної форми слова) та її граматичних характеристик (Віктор Бочаров); в обчислювальній лінгвістиці – аналіз окремих словоформ поза контекстом, у результаті якого послідовність словоформ вхідного тексту замінюється послідовністю інформації про ці словоформи (Володимир Волошин); процес, у результаті якого кожна словоформа тексту набуває свого коду частини мови і значення граматичних категорій (рід, число, відмінок, вид, час, особа тощо) та який забезпечує проведення морфемного, синтаксичного й семантичного аналізів, неможливих без попереднього визначення частин мови (Наталія Дарчук); вихідний модуль систем АПТ (АСОТ), у результаті здійснення якого комп'ютер для кожного слова в тексті визначає його граматичний клас, або частиномовну належність, та в межах граматичних класів – граматичний підклас, або граматичні підкласи, тобто розряди слів зі спільними змістовими, формальними та функціональними властивостями (здебільшого це слова, належні до різних граматичних категорій у межах окремих частин мови) (Євгенія Карпіловська); в автоматичній обробці тексту природною мовою за допомогою комп'ютера – процедуру, унаслідок якої з форми, зовнішнього оформлення слова в тексті можна одержати відомості про будь-які рівні мовної структури (Юрій Марчук).

АМА на основі графемного аналізу – алгоритм визначення граматичних класів, який використовує списки кінцевих буквосполучень.

АМА на основі операції логічного множення – алгоритм визначення граматичних класів, який використовує словник основ слів і бульові вектори

(сукупність нулів й одиниць – показники відсутності й наявності тієї чи тієї графеми у флексії).

АМА на основі словника основ – алгоритм визначення граматичних класів, який використовує словник основ слів і низку допоміжних морфологічних таблиць.

АМА на основі словника словоформ – алгоритм визначення граматичних класів, який зводиться до пошуку у словнику і вибору відповідної інформації.

АМА на основі флективного аналізу – алгоритм визначення граматичних класів, який використовує списки квазіфлексії.

АОТ – морфологічний модуль, який володіє словниками досить великого обсягу; при пошуку в словнику використовує кінцевий автомат, що дозволяє знаходити слово за лінійний від його довжини час (дуже швидко); написаний на C++, компілюється під Linux і під Windows; володіє розвиненою системою додавання нових слів; має в розпорядженні російський, німецький та англійський лексикон; поширюється безкоштовно під ліцензією LGPL у початкових кодах.

Безсловникова («незалежна») технологія опрацювання текстової інформації – представлення всіх необхідних відомостей про мовні одиниці у вигляді алгоритмічних правил.

Внутрішньопарадигматичні омоніми – омонімічні слововживання в межах певної граматичної категорії.

Внутрішньочастиномовні міжпарадигматичні омоніми – омонімічні слововживання, у яких лексико-граматичне значення збігається, але відрізняються леми.

Внутрішньочастиномовні омоніми – омонімічні слововживання з однаковою частиномовною належністю.

Граматична категорія – клас однотипних граматичних значень (категорії роду, числа, відмінка, особи, часу, способу, стану, виду).

Граматична форма – це мовний знак, за допомогою якого виражається граматичне значення.

Граматичні омоніми – формально тотожні граматичні форми або конструкції, що мають різне граматичне значення.

Дерево квазіфлексій – дерево, яке репрезентує квазіфлексії з певною кінцевою графемою.

Доморфологічний етап АМА – підготовчий етап, у процесі якого аналізований текст необхідно розбити на речення, у кожному реченні виокремити слова, розділові знаки, інші елементи тексту (числа, формули, таблиці, смайлики тощо).

Експериментальні системи АМА – перші системи АМА, створювані у 50-60-х рр., що включали такі етапи: автоматичне виділення основи у словоформі тексту; пошук основи у словнику основ; порівняння структури словоформи з даними про її основу, які містяться у словнику основ.

Елементи формалізації письмових текстів: пробіли – для виділення меж між словами, великі літери й розділові знаки – для виділення меж між реченнями й складовими частинами речень, абзацні відступи – для виділення меж між зв’язаними за смислом групами речень тощо.

ECAIT (Експериментальна система автоматичного індексування текстів) – одна з перших систем АМА, що включала послідовні блоки: морфологічний аналіз, синтаксичний аналіз, семантико-синтаксичний аналіз прийменникових конструкцій та варіювання смислового запису запиту; інструментами аналізу є: словники основ, закінчень, омонімічних основ; таблиці семантико-синтаксичної сполучуваності компонентів прийменникових конструкцій; зняття лексичної омонімії; семантичний аналіз іменних безприйменникових конструкцій; таблиці семантичної

сполучуваності іменників і прикметників; алгоритми АМА, які визначають певну послідовність перевірок і звертань до словника і таблиць.

Засоби вираження граматичного значення: флексія: *поле – поля*; суфікс: *ягня – ягн**ят**и*; префікс: *читати – **про**читати*; постфікс: *будувати – будувати**ся***; чергування: *сон – **сн**у*; наголос: *вода – в**о́**ди*; службові слова: ***хай** думає*; суплетивізм: *я – **мен**е*; порядок слів: *День змінює ніч. / Ніч змінює день*.

Зміна граматичних закінчень – основний спосіб утворення різних форм слів зі зміною їх роду, числа, відмінка й особи.

Змістове (смислове) автоматизоване опрацювання текстової інформації – перетворення фрагментів тексту з аналізом його змісту, встановленням логіко-семантичних відношень між його компонентами (використовується додаткова, семантична інформація, виражена в тексті імпліцитно).

Квазіфлексія – кінцівка словоформи, що дозволяє однозначно встановлювати частиномовну приналежність словоформ тексту та їх граматичну характеристику.

Лематизація – автоматичне ототожнення різних словозмінних форм одного слова передбачає їх зведення до вихідної, словникової форми / угруповання словоформ одного слова; нормалізація словникової форми.

Машинне слово – ланцюжок графем від пробілу до пробілу, у тому числі пунктуаційні знаки.

Метод індексування на основі семантичного аналізу – метод, який утворюють два етапи: індексування за дескрипторним словником у режимі «Слово», який супроводжується частковим морфологічним аналізом і лематизацією; індексування за допомогою автоматичного інформаційно-пошукового тезауруса.

Морфологічне слово – це сукупність усіх словоформ (граматичних форм слів); якщо така сукупність упорядкована, вона утворює парадигму.

Морфологія – це розділ лінгвістики, який досліджує структуру слів та їх морфологічні характеристики; частина граматичної будови мови, що охоплює частини мови (граматичні класи слів), морфологічні (граматичні) категорії цих частин мови та їх форми; комп'ютерна морфологія аналізує слова програмними засобами (Віктор Бочаров).

Морфологія слова – тільки те, що належить до його форми: закінчення, суфікси, флексії, корені й інші частини словоформи (Володимир Волошин).

Надстемінг – коли під час стематизації два слова скорочуються до однієї основи, хоча це не мало б статися.

Недостемінг – коли під час стематизації два слова отримують різні основи, хоча б мали мати одну спільну.

Незалежний АМА – алгоритм визначення граматичних класів, який відбувається без звертання до словника, лише за рахунок таблиць афіксів; це вивчення комбінаторики флексій та інших афіксів, ідея якого полягає у максимальному використанні інформації про флексію з урахуванням аломорфії та варіантності, щоб можна було б звести їх до однієї морфеми.

Промислові системи АМА – системи опрацювання текстової інформації, створювані у 70-80-х рр. – до сьогодні, що включають у себе: системи автоматичного перекладу, системи інформаційного пошуку, системи автоматичного редагування текстів, системи автоматизованого анотування й реферування літератури.

РЕФЕРАТ – система автоматичного індексування, що включає такі етапи: виокремлення найбільш інформативних слів і словосполучень із тексту; розшифрування абревіатури; заміна слів, основи яких мають дескриптори у машинному словнику, на код дескриптора; зняття омонімії.

Системи автоматичного перероблення тексту (АПТ) / автоматизовані системи опрацювання тексту (АСОТ) – один із різновидів лінгвістичних інтелектуальних комп'ютерних систем, що моделюють

розумову діяльність людини у процесі розв'язання теоретичних і практичних завдань.

Словникова технологія опрацювання текстової інформації – створення допоміжних лінгвістичних баз даних словників, правил ідентифікації мовних одиниць.

Слово – мінімальна формально виокремлювана одиниця зв'язного письмового тексту.

Словотвірна основа – це така основа, з якої додаванням суфіксів і закінчень можна отримати правильні (тобто наявні у словнику) словоформи; початкова частина буквеного коду, що лишається після відсікання максимальної кількості суфіксів та відповідає умові продуктивності.

Словотвірні класи – класи слів, що характеризуються однаковим переліком суфіксів і сполучень суфіксів, поєднуваних з їх словотвірною основою.

Словоформи – це граматичні форми одного слова, тотожні лексично (спільне лексичне значення), але протиставлені граматичним значенням.

Способи вираження граматичного значення: 1) синтетичний (граматичні значення в межах морфологічного слова виражаються за допомогою афіксів: закінчення, суфікс, префікс, інтерфікс тощо): *дядьк**ові**, кошен**яти**, **від**бігти*; 2) аналітичний (показник граматичних значень – службове слово): ***буду** мріяти, мріяв **би**, **хай** мріє*; 3) аналітико-синтетичний (поєднання двох попередніх – афіксальне граматичне оформлення слова + аналітичні елементи): ***на** вікн**і**, **в** університет**і*** (грамема М.в.); 4) суплетивний (творення граматичних форм від різних коренів): ***ти** – **тебе** – **тобі**, **поган**ий – **гір**ший*.

Стемінг – це процес скорочення слова до основи шляхом відкидання допоміжних частин, таких як закінчення чи суфікс.

Фактори вибору методу АМА: 1) система мови (якщо найбільше розвинений синтетичний спосіб вираження граматичного значення (словозміна), то початковим етапом АМА є аналіз структури словоформи; якщо найбільше розвинений аналітико-синтетичний або аналітичний спосіб вираження граматичного значення (сполучення різних слів або словоформ), то аналіз слова являє собою пошук за словником заздалегідь визначених морфологічних характеристик кожного слова чи словоформи (подальший аналіз відбувається за допомогою оточення (дистрибуції) слова)); 2) система письма і друку (адже АМА призначений для писемного тексту, у якому має значення, буквене це чи складове письмо; як співвідносяться усне / писемне мовлення; спосіб членування тексту спеціальними засобами); 3) тематика тексту як результату мовленнєвої діяльності й засобу комунікації.

Флективні класи – класи відмінюваних слів, що виокремлюються на основі аналізу їх синтаксичної функції та систем відмінкових, особових та родових закінчень.

Формальне автоматизоване опрацювання текстової інформації – перетворення фрагментів тексту без аналізу його змісту.

LanguageTool – відкритий програмний засіб перевірки граматики, який використовує морфологічний модуль і працює на основі спеціальних правил.

libmorphukr – модуль морфологічного аналізу української мови, призначений для перевірки правопису окремих слів (правильно – неправильно), лематизації (побудови нормальних форм слів за довільною формою), вилучення граматичних описів тих форм, із якими збігся поданий рядок, морфологічного синтезу форм за нормальною формою і граматичним описом, також побудови списку можливих правильних накреслень для неправильного слова (підказка).

MCR DLL v2.0 – автоматичний морфологічний аналізатор, призначений для опрацювання російськомовних текстів.

SMART – система індексування тексту документів, в основі якої система поділу слів тексту на флексію і основу; словник еквівалентностей (тезаурус), призначений для заміни еквівалентних слів одним або кількома номерами понять, які слугують ідентифікаторами змісту замість основ слів; тезаурус у вигляді ієрархії понять, що забезпечує пошук для даного поняття загальнішого або вужчого чи асоційованого з ним поняття; словники статистичних і синтаксичних словосполучень; система обслуговування словників.

StarLing – автоматичний морфологічний аналізатор, що дає можливість ознайомитися з комп'ютерними базами даних за словниками Ожегова, Залізняка і Мюллера, а також проаналізувати будь-яке російське й англійське слово та отримати його повну акцентовану парадигму.

Додаток 4

ПЕРЕЛІК ПИТАНЬ ДО ЗАЛІКУ З ПРИКЛАДНОЇ МОРФОЛОГІЇ

1. Поняття автоматизованого опрацювання текстової інформації. Словникова і безсловникова технології опрацювання текстової інформації.
2. Поняття автоматичного морфологічного аналізу тексту. Його завдання.
3. Морфологія: визначення, завдання, предмет, об’єкт, ключові поняття.
4. Словоформа: визначення, приклади.
5. Морфологічне слово: визначення, приклади.
6. Граматичне значення: визначення, приклади.
7. Граматична категорія: визначення, приклади.
8. Способи вираження граматичного значення: різновиди, приклади.
9. Засоби вираження граматичного значення: різновиди, приклади.
10. Словотвірні класи: визначення, приклади.
11. Флективні класи: визначення, приклади.
12. Чинники визначення АМА.
13. Експериментальні системи АМА: ECAIT.
14. Промислові системи АМА: SMART, РЕФЕРАТ, метод індексування на основі семантичного аналізу.
15. Напрями у розробці сучасних систем АМА.
16. Основні завдання сучасних і майбутніх систем АМА.
17. Доморфологічний етап як підготовчий процес будь-якого АМА.
18. Елементи формалізації письмових текстів.
19. Поняття машинного слова. Складнощі його виокремлення.
20. Поняття лематизації. Її алгоритми на матеріалі української мови.
21. Поняття стемінгу. Варіанти його алгоритмів. Надстемінг і недостемінг.
22. Проблема граматичної омонімії у процесі АМА.
23. Граматичні омоніми: основні різновиди, приклади.
24. Основні підходи до граматичного уоднозначнення у процесі АМА.

25. Автоматичний морфологічний аналіз на основі графемного аналізу: завдання, етапи, недоліки й переваги в порівнянні з іншими методами АМА.
26. Автоматичний морфологічний аналіз на основі флективного аналізу: завдання, етапи, недоліки й переваги в порівнянні з іншими методами АМА.
27. Автоматичний морфологічний аналіз на основі словника словоформ: завдання, етапи, недоліки й переваги в порівнянні з іншими методами АМА.
28. Автоматичний морфологічний аналіз на основі словника основ: завдання, етапи, недоліки й переваги в порівнянні з іншими методами АМА.
29. Автоматичний морфологічний аналіз на основі операції логічного множення: завдання, етапи, недоліки й переваги в порівнянні з іншими методами АМА.
30. Автоматичний морфологічний аналізатор LanguageTool: функціонал, ефективність.
31. Автоматичний морфологічний аналізатор libmorphukr: функціонал, ефективність.
32. Автоматичний морфологічний аналізатор AOT: функціонал, ефективність.
33. Автоматичний морфологічний аналізатор StarLing: функціонал, ефективність.
34. Автоматичний морфологічний аналізатор MCR DLL v2.0: функціонал, ефективність.
35. Місце АМА серед дисциплін прикладної лінгвістики.

Додаток 5

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ВАСИЛЯ СТУСА
ФІЛОЛОГІЧНИЙ ФАКУЛЬТЕТ
КАФЕДРА ЗАГАЛЬНОГО ТА ПРИКЛАДНОГО МОВОЗНАВСТВА
І СЛОВ'ЯНСЬКОЇ ФІЛОЛОГІЇ

РОБОЧА ПРОГРАМА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ
ПРИКЛАДНА ЛІНГВІСТИКА V. ПРИКЛАДНА МОРФОЛОГІЯ

ступінь освіти	**Бакалавр**
галузь знань	**03 Гуманітарні науки**
спеціальність	**035 Філологія (035 Philology)**
освітня програма	**035.10 Прикладна лінгвістика (Applied Linguistics)**

Вінниця 2020

Робоча програма навчальної дисципліни «Прикладна лінгвістика V. Прикладна морфологія» для здобувачів вищої освіти спеціальності «035 Філологія (035 Philology)», освітньої програми «035.10 Прикладна лінгвістика (Applied Linguistics)», СО «Бакалавр»

Розробник: Гарбера Ірина Володимирівна, кандидат філологічних наук, доцент кафедри загального та прикладного мовознавства і слов'янської філології

Робоча програма затверджена на засіданні кафедри загального та прикладного мовознавства і слов'янської філології

ВСТУП

Робоча програма навчальної дисципліни «Прикладна лінгвістика V. Прикладна морфологія» складена відповідно до освітньої програми «035.10 Прикладна лінгвістика (Applied Linguistics)» спеціальності «035 Філологія (035 Philology)», СО «Бакалавр».

Метою вивчення навчальної дисципліни є формування компетентностей у діяльності, пов'язаній з автоматичним морфологічним аналізом письмових текстів різних жанрів і стилів (з прикладною та науково-дослідною метою), застосуванням комп'ютерних технологій для розв'язання лінгвістичних завдань, проєктуванням лінгвістичних програмних продуктів, зокрема автоматичних морфологічних аналізаторів.

Навчальна дисципліна формує міждисциплінарні взаємозв'язки із іншими дисциплінами, такими як: «Основи програмування», «Теорія алгоритмів», «Вступ до прикладної лінгвістики», «Комп'ютерна лінгвістика», «Сучасна українська мова», «Іноземна мова».

Вивчення навчальної дисципліни передбачає формування та розвиток у здобувачів вищої освіти наступних компетентностей та програмних результатів навчання:

Загальні компетентності (ЗК):

ЗК-2. Здатність зберігати та примножувати моральні, культурні, наукові цінності і досягнення суспільства на основі розуміння історії та закономірностей розвитку предметної області, її місця у загальній системі знань про природу і суспільство та у розвитку суспільства, техніки і технологій, використовувати різні види та форми рухової активності для активного відпочинку та ведення здорового способу життя.

ЗК-5. Здатність до пошуку, опрацювання та аналізу інформації з різних джерел.

ЗК-6. Уміння виявляти, ставити та вирішувати проблеми.

ЗК-7. Здатність працювати в команді та автономно.

ЗК-9. Здатність до абстрактного мислення, аналізу та синтезу.

ЗК-11. Навички використання інформаційних і комунікаційних технологій.

ЗК-12. Здатність проведення досліджень на належному рівні.

Спеціальні (фахові, предметні) компетентності (СК):

СК-1. Усвідомлення структури філологічної науки (лінгвістики) та її теоретичних основ.

СК-2. Здатність використовувати у професійній діяльності знання про мову як особливу знакову систему, її природу, функції, рівні.

СК-3. Здатність використовувати в професійній діяльності знання з теорії та історії мов, що вивчаються.

СК-5. Здатність до збирання й аналізу, систематизації та інтерпретації мовних фактів, інтерпретації та перекладу тексту.

СК-6. Здатність вільно оперувати спеціальною термінологією для розв’язання професійних завдань.

СК-9. Здатність здійснювати формалізований аналіз мовного матеріалу та використовувати комп’ютерні технології для дослідження мовних явищ і/або мовленнєвих процесів.

Програмні результати навчання (ПРН):

ПРН-1. Вільно спілкуватися з професійних питань із фахівцями та нефахівцями державною та іноземною мовами усно й письмово, використовувати їх для організації ефективної міжкультурної комунікації.

ПРН-2. Ефективно працювати з інформацією: добирати необхідну інформацію з різних джерел, критично аналізувати й інтерпретувати її, впорядковувати, класифікувати й систематизувати.

ПРН-3. Організовувати процес свого навчання й самоосвіти.

ПРН-5. Використовувати інформаційні й комунікаційні технології для вирішення складних спеціалізованих задач і проблем професійної діяльності.

ПРН-6. Розуміти основні проблеми філології та підходи до їх розв’язання із застосуванням доцільних методів та інноваційних підходів.

ПРН-7. Знати й розуміти систему мов, що вивчаються, вміти застосовувати ці знання у професійній діяльності.

ПРН-9. Аналізувати мовні одиниці, визначати їхню взаємодію та характеризувати мовні явища і процеси, що їх зумовлюють.

ПРН-12. Знати й розуміти основні поняття, теорії та концепції прикладної лінгвістики, уміти застосовувати їх у професійній діяльності.

ПРН-13. Збирати, аналізувати, систематизувати й інтерпретувати факти мови й мовлення й використовувати їх для розв’язання складних задач і проблем у спеціалізованих сферах професійної діяльності та/або навчання.

ПРН-14. Мати навички участі в наукових та прикладних дослідженнях у галузі філології (лінгвістики та інформаційних технологій).

ПРН-16. Розбивати інформацію на компоненти, розуміти їх взаємозв’язки та організаційну структуру, бачити помилки й огріхи в логіці міркувань, різницю між фактами і наслідками, оцінювати значущість даних.

ПРН-17. Застосовувати методи теоретичної та прикладної лінгвістики для розв’язання складних професійних завдань.

ПРН-18. Створювати лінгвістичний алгоритм для розв’язання поставлених завдань дослідження.

ПРН-19. Проєктувати лінгвістичні програмні продукти різних типів відповідно до поставлених завдань.

1. ОПИС НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

<table>
<tr><td>Найменування показників</td><td>Опис підготовки фахівців</td><td>Характеристика навчальної дисципліни
денна форма навчання</td></tr>
<tr><td rowspan="2">Кількість кредитів – 4</td><td rowspan="2">Галузь знань –
03 Гуманітарні науки</td><td>цикл професійної та практичної підготовки</td></tr>
<tr><td>фахова</td></tr>
<tr><td rowspan="2">Кількість годин всього – 120</td><td rowspan="2">Спеціальність –
035 Філологія (035 Philology)</td><td>Курс підготовки – 2</td></tr>
<tr><td>Семестр – 4</td></tr>
<tr><td>з них:</td><td rowspan="2">Освітня програма – 035.10 Прикладна лінгвістика (Applied Linguistics)</td><td rowspan="2">Лекції – 34 год.</td></tr>
<tr><td>аудиторних – 68</td></tr>
<tr><td>для самостійної роботи студента – 52</td><td rowspan="2">Рівень вищої освіти:
перший</td><td rowspan="2">Лабораторні – 34 год.</td></tr>
<tr><td rowspan="3">Кількість змістовних модулів – 3</td></tr>
<tr><td rowspan="2">Ступінь освіти:
бакалавр</td><td>Самостійна робота – 52 год.</td></tr>
<tr><td>Вид контролю – залік</td></tr>
</table>

2. ПРОГРАМА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

ЗМІСТОВИЙ МОДУЛЬ 1. ТЕОРІЯ АВТОМАТИЧНОГО МОРФОЛОГІЧНОГО АНАЛІЗУ

Тема 1. Автоматичний морфологічний аналіз серед дисциплін прикладної лінгвістики. ***Тема 2.*** Експериментальні та промислові системи автоматичного морфологічного аналізу. ***Тема 3.*** Доморфологічний аналіз як початковий етап автоматичного морфологічного аналізу. ***Тема 4.*** Лематизація та стемінг. ***Тема 5.*** Проблема граматичної омонімії у процесі автоматичного морфологічного аналізу.

ЗМІСТОВИЙ МОДУЛЬ 2. ОСНОВНІ МЕТОДИ АВТОМАТИЧНОГО МОРФОЛОГІЧНОГО АНАЛІЗУ

Тема 1. Автоматичний морфологічний аналіз на основі графемного аналізу. ***Тема 2.*** Автоматичний морфологічний аналіз на основі флективного аналізу. ***Тема 3.*** Автоматичний морфологічний аналіз на основі словника словоформ. ***Тема 4.*** Автоматичний морфологічний аналіз на основі словника основ. ***Тема 5.*** Автоматичний морфологічний аналіз на основі операції логічного множення.

ЗМІСТОВИЙ МОДУЛЬ 3. СУЧАСНІ АВТОМАТИЧНІ МОРФОЛОГІЧНІ АНАЛІЗАТОРИ

Тема 1. Автоматичний морфологічний аналізатор LanguageTool. ***Тема 2.*** Автоматичний морфологічний аналізатор libmorphukr. ***Тема 3.*** Автоматичний морфологічний аналізатор AOT. ***Тема 4.*** Автоматичний морфологічний аналізатор StarLing. ***Тема 5.*** Автоматичний морфологічний аналізатор MCR DLL v2.0.

3. МЕТОДИ І ФОРМИ КОНТРОЛЮ, КРИТЕРІЇ ОЦІНЮВАННЯ ЗНАНЬ ЗДОБУВАЧІВ ВИЩОЇ ОСВІТИ

Організаційно-навчальна робота представлена такими формами: виконання післялекційних завдань, виконання лабораторних робіт, виконання модульних контрольних робіт (МКР – виділені окремим пунктом у системі оцінювання, бо мають узагальнюючий, підсумковий характер).

Самостійна робота представлена такими формами: конспектування статей.

Індивідуальна робота представлена такими формами: підготовка й захист короткої презентації / повідомлення за обраною темою.

Система оцінювання знань

Змістовий модуль 1		**Змістовий модуль 2**		**Змістовий модуль 3**	
організаційно-навчальна робота	**15**	*організаційно-навчальна робота*	**15**	*організаційно-навчальна робота*	**15**
самостійна робота	**3**	*самостійна робота*	**3**	*самостійна робота*	**3**
індивідуальна робота	**6**	*індивідуальна робота*	**6**	*індивідуальна робота*	**6**
МКР №1	**9**	*МКР №2*	**9**	*МКР №3*	**10**
Поточний контроль (max 100 балів)					

4. РЕКОМЕНДОВАНА ЛІТЕРАТУРА ТА ІНФОРМАЦІЙНІ РЕСУРСИ

Основна література:

1. Данилюк І. Прикладна морфологія. Донецьк, 2010. 216 с.
2. Романюк Ю. Прикладна морфологія. Черкаси, 2009. 117 с.

Допоміжна література:

1. Бабина О. Корпусный метод автоматического морфологического анализа флективных языков. *Вестник Южно-Уральского государственного университета. Серия «Лингвистика»*. №25(284), выпуск 15. Челябинск, 2012. С. 38-44.
2. Баранов А. Введение в прикладную лингвистику. М., 2001. 360 с.
3. Белоногов Г. Компьютерная лингвистика и перспективные информационные технологии. М., 2004. 159 с.
4. Буньо Г. Сучасні методи вирішення проблеми граматичної омонімії в тексті. *Наукові записки. Серія: Філологічна*. 2014. Вип. 49. С. 12-16.
5. Волошин В. Комп'ютерна лінгвістика. Суми, 2004. 382 с.
6. Гельбух А., Сидоров Г. К вопросу об автоматическом морфологическом анализе флективных языков. Ел. режим доступу: www.dialog-21.ru/Archive/2005.
7. Грязнухіна Т., Нікула М. Система автоматичного морфологічного аналізу українського наукового тексту. Проблеми українізації комп'ютерів. Матеріали 2-ї міжнародної конференції. Київ, 1993. С. 42-46.
8. Дарчук Н. Комп'ютерна лінгвістика. К., 2008. 351 с.
9. Дарчук Н. Морфологічне анотування Корпусу української мови. *Комп'ютерна лінгвістика: сучасне і майбутнє. Матеріали Міжнародної науково-практичної конференції*. К., 2012. С. 16-19.
10. Дарчук Н. Комп'ютерне анотування українського тексту: результати і перспективи. К., 2013. 543 с.

11. Карпіловська Є. Вступ до прикладної лінгвістики: Комп'ютерна лінгвістика. Донецьк, 2006. 188 с.
12. Марчук Ю. Компьютерная лингвистика. М., 2007. 317 с.
13. Міщенко Н. Система програм морфологічного аналізу науково-технічних текстів. *Наукові записки. Серія «Філологічні науки»*. Випуск 95(2). К., 2011. С. 538-542.
14. Морфологический анализ научного текста на ЭВМ. К., 1989. 262 с.
15. Николаев И., Митренина О., Ландо Т. Прикладная и компьютерная лингвистика. М., 2016. 315 с.
16. Ножов И. Морфологическая и синтаксическая обработка текста (модели и программы). М., 2003. 140 с.
17. Партико З. Прикладна і комп'ютерна лінгвістика: Вступ до спеціальності. Львів, 2008. 224 с.
18. Соснина Е. Введение в прикладную лингвистику. Ульяновск, 2012. 110 с.
19. Сучасна українська літературна мова. Морфологія. К., 1969. 250 с.
20. Antworth E. PC-KIMMO: a two-level processor for morphological analysis. *Occasional Publications in Academic Computing*. No. 16. Dallas, 1990. 273 p.

Інформаційні ресурси:

1. http://www.pldonnu.pp.ua
2. https://moodle.donnu.edu.ua/course/view.php?id=1639