УДК 681.3.06

Демсъка-Кульчицъка О. М., Перевозчикова О. Л., Січкаренко В. О.

РОЗВИТОК ЛЕКСИЧНОЇ КАРТОТЕКИ УКРАЇНСЬКОЇ МОВИ

Розглядаються теоретичні та інформаційні аспекти побудови комп'ютерної лексикографічної картотеки на підґрунті повнотекстової бази і текстових джерел на будь-яких носіях (паперових, електронних, телексних, факсимільних). Розглядаються напрями реалізації підсистем вводу, аналізу, обробки і збереження текстових матеріалів.

Лексична картотека (ЛК) - це сукупність лексичних карток із заголовними словами, текстами-ілюстраціями вживання цих слів у відповідному значенні та вказівкою на джерело ілюстративного тексту [1]. Подальший розвиток лексичної картотеки української мови в сучасних умовах передбачає перехід від так званої традиційної, на паперових носіях, картотеки до комп'ютерної чи віртуальної ЛК, що реалізована в Інституті української мови НАНУ (за завданням його директора чл.-кор. В. В. Німчука) разом з Інститутом кібернетики НАНУ.

Архітектурна будова. Комп'ютерна ЛК - це аналогічна до традиційної сукупність лексичних карток із заголовними словами, текстами-ілюстраціями вживання цих слів у відповідному значенні та вказівкою на джерело ілюстративного тексту, але зберігається вона на електронних носіях, містить додаткові інформаційні поля, і з неї за алгоритмом лематизації автоматично формується заданий реєстр. Електронна ЛК має складну архітектуру, в основі якої лежить словоцентричний підхід, тобто слово стає точкою відліку в тріаді «лексична картка - гіперкартка - ЛК як колекція гіперкарток», що показано на рис. 1.

Сучасні електронні ЛК мають досить складну архітектуру, містять зліченну множину структурних елементів. Так, у архітектурі гіперкарток їхня колекція фактично є комп'ютерною ЛК простої ієрархічної будови, де гіперкартка - це колекція всіх наявних у базі карток-ілюстрацій щодо усіх словоформ конкретної леми, гіпокартка - це колекція наявних у базі карток на конкретну словоформу конкретної гіпекарткової леми. Наприклад, для іменника таких гіпокарток повинно бути 14 одиниць, оскільки 14 відмінкових словоформ однини і множини може мати слово-іменник. Відповідно, картка є конкретною одиничною карткою гіпокарткової словоформи, яких може бути досить багато, а чисельність залежить від кількості слововживань конкретної леми у певній словоформі в опрацьованих текстах, що схематично зображено у табл. 1.

Лексична картка у паперовій картотеці могла агрегувати до 40 одиниць контекстів слововживання. Обмеженість параметрів паперової картки та її одновимірність уможливлювали лише фіксацію заголовного слова, ілюстрацію його вживання та бібліографічний опис за схемою: *автор, видання, заголовок, том, дата, сторінка*. Натомість

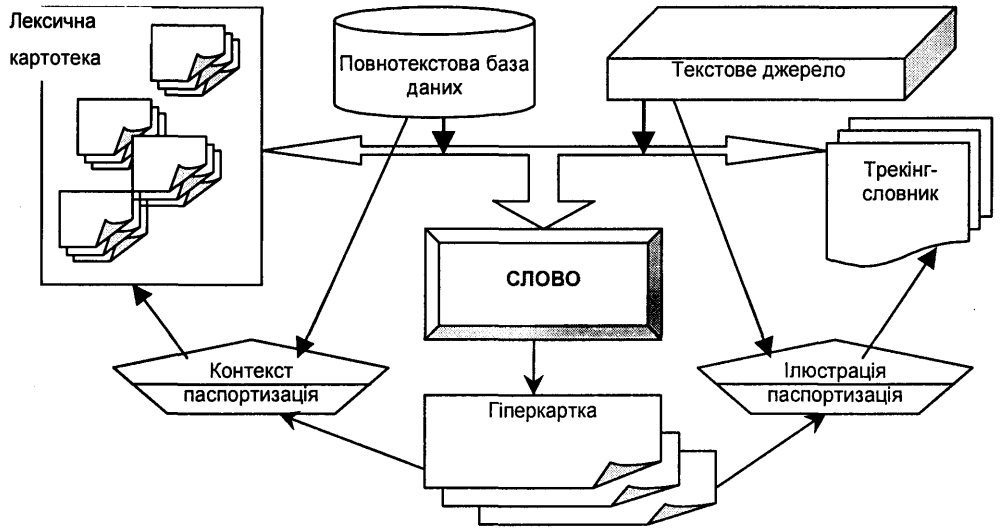

Рис. 1. Словоцентричний підхід до побудови комп'ютерної ЛК

віртуальна картка дає змогу розмістити значно більше інформації: відомості про наголос, граматичні характеристики слова, парадигму відмінювання, морфемне членування, керування, синонімію, полісемію, омонімію, паронімію тощо. Власне це є передумовою архітектурної складності електронної ЛК.

Так, складна будова ЛК передбачає наявність різних архітектурних елементів з можливим як безпосереднім, так і опосередкованим переходом від елемента до елемента, що далі наведено у табл. 1, де блок *слово/цитата/бібліографія цитати* є спрощеною схемою гіперкартки, що будується автоматично безпосередньо з ресурсів повнотекстової БД. Інші блоки можуть формуватися як безпосередньо з текстової бази даних, так і з інших лінгвістичних джерел. Наприклад, блок *лексичний параметр* формується на словниковій базі. У цьому полі подається інформація про наявність запитуваного слова у словниках української

Таблиця 1. Схема електронної лексичної картотеки складної архітектури

				переклад на … мову / мови		
						омонім(и) / омограф(и)
						антонім(и)
						синонім(и)
						варіант(и) / дублет(и)
				фразеологічне сполучення		
				вільне словосполучення		
						етимологія
						лексикографічний параметр
ремарка // ремарки						
				парадигмування		
дієслово	іменник	займенник	числівник	прикметникові слова		незмінні частини мови
морфемне членування	складоподіл	довжина слова	вимова	орфографія	наголос	частиномовна характеристика
корінь						
афікси						
						БІБЛІОГРАФІЯ ЦИТАТИ
						ЦИТАТА
						СЛОВО
						ПОВНОТЕКСТОВА БАЗА ДАНИХ

мови із збереженням графіки, аналогічно з блоком *переклад на ... мову/мови.* Але поле *наголос* може реалізуватися двояко. Один підхід зветься словниковим, тобто у систему вводиться інтерактивний словник наголосів і користувач може знайти у цьому словнику необхідне слово та відомості про його наголошування. Другий підхід полягає у побудові таблиці-сітки [2] наголосів шляхом формалізації правил наголошування слів в українській мові, за допомогою якої комп'ютер розставлятиме наголос. Це продуктивніше за словниковий підхід, оскільки уможливлює отримання інформації про наголошування слів, відсутніх у словнику наголосів через свою неологічність.

Крім того, програмна складність побудови електронної ЛК полягає в тому, що необхідно, по-перше, реалізувати комп'ютерне оброблення текстових масивів, по-друге, створити електронну повнотекстову базу даних (ПБД) і, по-третє, уможливити одержання електронних текстових документів або перетворення текстів на паперових носіях чи факсиміле в електронні тексти.

Як показано у табл. 1, електронна ЛК складної архітектури містить поля двох глобальних типів - ідентифікаційне та інформаційне. Ідентифікаційне поле, по суті, відтворює традиційну лексичну картку, подаючи слово, цитату-ілюстрацію і бібліографічні дані *(автор, назва, заголовок, том, випуск, дата, сторінка).* Інформаційні поля є складовою частиною усіх інших архітектурних елементів, крім *повнотекстової бази даних.* Це групи:

a) морфемне членування слова з окремими входами у підполя *корінь* та *афікси/'складоподіл/довжина;*
b) слова/вимова/орфографія/наголос/частиномовна характеристика;
c) дієслово/іменник/займенник/числівник/прикметникові слова/ незмінні частини мови;
d) парадигмування;
e) ремарка/ремарки;
f) лексикографічний параметр;
g) етимологія;
h) фразеологічне сполучення/вільне словосполучення;
i) варіанти/дублети/синоніми/антоніми/омоніми/омографи;
j) переклад на ... мову/мови.

Оскільки в основі будь-якої паперової чи комп'ютерної ЛК лежить текст, то, відповідно, побудова ЛК починається з формування ПБД, у якій реалізується віртуальне подання даних, що допомагає в кожний момент часу здійснювати інформаційні зрізи, які дають змогу оперативно аналізувати, порівнювати і добирати за параметрами конкретного текстового матеріалу і супровідної інформації лексичні одиниці, ілюстративну інформацію або бібліографічні дані. А на підставі базових правил можна виділяти так зване чисте слово, тобто позбавлене зайвих знаків пунктуації і можливих словодеформативних чинників. Натомість розширені правила повинні давати змогу виділяти не лише необхідне слово, але і його словоформу, щоб мати повну картину використання/функціонування того чи іншого слова в мові.

Крім того, використовуваний трекінг-словник дає змогу встановити норми виділення найінформативнішого з погляду лінгвіста матеріалу і, так би мовити, затінення неповнозначних неінформативних мовних одиниць, визначити порядок оброблення числових значень, сформувати списки ключових слів, тематичні покажчики, рубрикацію тощо. Далі через послідовність *Норма —> Загальне дерево рубрик —> Рубрикація —> Список ключових слів* реалізується найоптимальніше подання функціонування словоформ.

При створенні ПБД одним із важливих аспектів є відбір фактичного матеріалу. Існує чимало методів та методик відбору текстового матеріалу, призначених як для лексикографічних робіт, так і для інших лінгвістичних досліджень. Загалом, методики відбору текстового матеріалу для ПБД або машиночитаних корпусів текстів сформувалися в межах лінгвістики корпусу, основним завданням якої є впровадження тексту природної мови у сферу комп'ютерно-інформаційних технологій. Концептуально відбір текстового матеріалу залежить від його призначення, тематики і обсягу. Так, якщо цікавою є конкретна лексика певного часового відрізка, то доцільно відібрати тексти, апріорі багаті на такий лексичний матеріал, і за умови незначних обсягів фактичного матеріалу вводити тексти повністю. Повністю вводяться тексти при створенні комп'ютерної бази творів окремого автора. В інших випадках, особливо, коли необхідно мати загальномовну картину великого періоду існування мови обсягом понад мільйони слововживань, робиться вибірка, зазвичай береться до уваги одна третя текстового матеріалу окремого тексту певного змісту, тематики, стилю, автора тощо [3].

Реалізація першого прототипу. У лексикографії зараз важливим є найбільш повне і багате відтворення української мовної картини кінця XX століття і фіксація її в словниках сучасного періоду. Тому настільки важливі текстові джерела і фактичний матеріал, основу якого становлять:

- тексти друкованих ЗМІ з 1980 до 2001 рр. (особливо виділений початок 90-х років); попередньо не оброблені тексти художньої літератури (за історичних обставин заборонені і нові твори українських авторів);
- сучасні наукові і науково-популярні праці;
- шкільні підручники, видані здебільшого після 1995 року.

Складність побудови електронної ЛК полягає в тому, що необхідна не просто обробка текстових масивів, але і створення комп'ютерної ПБД із вирішенням задачі надходження по каналах зв'язку

електронних текстових документів чи перетворення у електронний варіант текстових джерел на паперовому носії чи у факсимільному зображенні. Насамперед це пов'язано з розходженням схем кодування і зв'язаних з ними лексикографічних *порядків для української мови [4]. Під час пошуку* необхідного слова і його контексту здійснюється статистичний аналіз, установлюються частотні показники для конкретного слова в текстовому джерелі чи у всій ПБД, для чого визначені правила добору слів-кандидатів для ЛК і реалізована базова підсистема відбору слів-кандидатів. Крім того, для обробки текстового матеріалу джерел реалізовані підсистеми:

- попереднього розбору масиву символьних даних;
- статистичної обробки;
- оцінки і розбору слів як кандидатів на розміщення в ЛК;
- пошуку і виділення контексту слова.

Спеціальний рівень захисту від вірусних атак передбачений для обробки текстових масивів великого розміру, одержуваних по мережі Інтернет чи електронною поштою від україномовних газет, журналів та інших друкованих видань. Відзначимо їхню здатність бути потенційними вірусоносіями і руйнувати не саму створену систему, а передусім ПБД.

Значний інтерес має набір правил, сформульованих для добору слів у ЛК. Саме ці, закладені в систему правила найбільше автоматизують ведення ПБД, роблячи її насправді електронним повнотекстовим джерелом, що виключає вплив сторонніх факторів, передусім людського.

Так, формування набору правил, що лежать в основі добору слів для введення в ЛК, засноване головним чином на: а) статистичному методі, б) діахронно-порівняльному аналізі фактичного матеріалу, в) семантичній інтерпретації семного набору конкретної лексичної одиниці, виходячи з виділеного контексту. Залежно від частотності словоживання в ЛК відбираються слова або з найвищою частотністю, або з мінімальною. Високий показник частотності свідчить про поширеність уживання слова, а мінімальний - про специфіку лексичного матеріалу мови у певний період його дослідження.

Діахронно-порівняльний аналіз дає змогу визначати час появи слова, пік активності функціонування в мові і відхід на периферію, тобто простежити так званий життєвий цикл слова, а використовуючи паралельно семантичну інтерпретацію, можна визначити характер життєвого циклу і специфіку вживання/функціонування слова, відібраного для введення в ЛК. Крім того, аналізуються граматичні, деривaційні, синтаксичні, стилістичні параметри конкретного слова, що дає змогу усувати фактичні чи формальні помилки в текстовому матеріалі, коректно відтворювати лек-

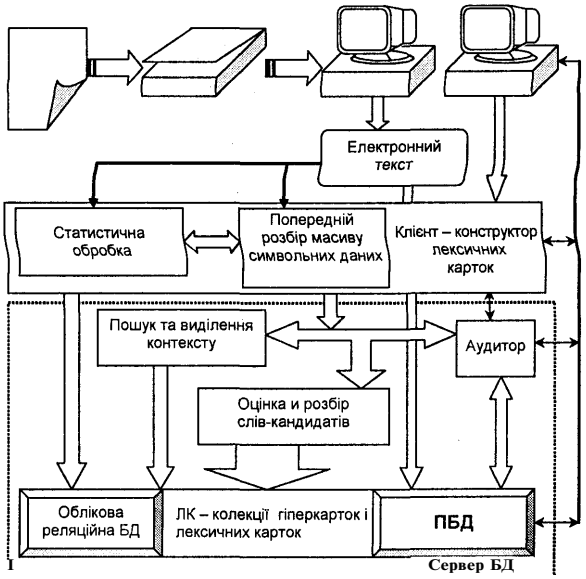

Рис. 2. Схема одержання лексичної картки в ЛК

сикографічну форму слова. Зазначимо, що її відтворення безпосередньо зв'язане з формалізацією граматики української мови, тобто зі зведенням традиційної академічної граматики до системи формальних граматичних правил для комп'ютера, використовуючи які комп'ютер визначає словоформу і співвідносну лексикографічну форму, будує коректний регістр слів, анулюючи різного роду помилки.

Також підкреслимо, що задачі формування ЛК і, отже, алгоритми їхньої реалізації відрізняються від традиційних процедур обробки ПБД для інформаційного пошуку і контент-аналізу. Так, частотний аналіз слів здійснюється для прийменників і сполучників, які зазвичай відкидаються за стоп-словником у класичних процедурах автоматичного індексування текстових джерел. Для створення ЛК слововживання цих частин мови є чи не найпоказовішим у картині зміни мови.

Надбудова над ПБД у вигляді облікової реляційної БД служить передусім для створення і збереження профілів текстових документів-джерел, виконуючи роль каталогу з зовнішніми зв'язками. Пов'язана статистична й інша інформація, що заноситься в облікову реляційну БД, допускає аналітичні (частотний аналіз слів, виділення леми і значеннєвих груп тощо) і аудиторські дії не тільки щодо слововживань, але і щодо текстового матеріалу загалом. З цією метою в межах облікової реляційної БД реалізовані відповідні блоки аналізу, побудовані на властивості об'єктності мови SQL-99 [5]. Система тригерів, обмежувальних та перевіркових умов сукупно з контрольними таблицями дають змогу організувати контроль при відновленні БД і проводити аналітичні дії над ПБД безпосередньо на стороні сервера БД, повною мірою використовуючи переваги клієнт-серверної архітектури. При цьому не тільки зменшується

міжмережний трафік за рахунок виконання більшості дій над збереженою у БД інформацією безпосередньо на сервері, але й зростає швидкість обробки даних як, мабуть, найкритичніший параметр реалізації системи.

Передбачається, що розмір ПБД за 1,5-2 роки експлуатації досягне 12-15 Гбайт, для чого необхідні спеціальні алгоритми забезпечення реактивності системи. Одним з рішень є перенесення частини логіки виконання застосування на сервер БД, що автоматично приведе до ще однієї переваги - відсутності зайвої перекомпіляції застосування у разі зміни (нарощування, модифікації тощо) правил обробки текстових джерел та процесів аудиту й аналізу ПБД.

На практиці більшість лексичних карток створюються вручну (у рукописному вигляді чи на друкарській машинці). Звідси низька продуктивність роботи і, відповідно, недостатня динаміка спостереження за розвитком слововживання, коли нормою вважається продукування 100-150 лексичних карток за день. Реалізований прототип системи дозволяє в режимі наскрізного друку робити до 8000 карток за день. Проблема полягає у формуванні ПБД, яке вимагає значних трудовитрат для введення текстових матеріалів, наприклад сканування та розпізнавання тексту з його очищенням і обробкою.

Важливим питанням залишається використання матеріальної ЛК, що нараховує в Інституті української мови НАНУ близько 13 млн одиниць. Тут можливі кілька шляхів вирішення задачі, але доцільно зберегти цю ЛК як свідчення української мови кінця XIX - початку XXI століть, аналогічно збереженню, наприклад, стародрукованих чи рукописних видань XI-XIV століть. Зазначимо, що про повноцінну роботу електронної ЛК можна говорити вже після 8-10 місяців від початку її створення. Хоча у європейській практиці створення віртуальних ЛК існують приклади відмови від паперового варіанта картотеки з моменту початку повномасштабної роботи електронної ЛК [6, 7].

У реалізованому прототипі (альфа-версія системи) програмко підтримуються переведення друкованого тексту в електронний вигляд і обробка текстових масивів у режимі неформатованого (plain) тексту чи в RTF-форматі, одержання статистичної інформації, вичленування слова і його контексту з первинною підготовкою тексту *(розпізнавання абревіатур, цифрових записів чисел, видалення «зайвих» знаків пунктуації та «прихованих» пробілів, ідентифікація слів за трекінг-словником* тощо). Визначаються характеристики тексту *(назва, тип, видавець, автор* тощо, усього близько 20 параметрів) і створюється ідентифікаційна картка тексту для розміщення в обліковій реляційній БД і при подальшій обробці у ПБД.

Система здатна працювати на різних рівнях складності для підтримки дій:

- техніка-оператора, що готує матеріали (у тому числі електронні) для реєстрації в ПБД, виконуючи тільки механічну роботу. Передредагування тексту проходить у напівавтоматичному режимі;

— експерта-лексиколога, що контролює роботу ПБД навчає комп'ютер, задаючи зразки слововживань, формулюючи і змінюючи правила добору слів, і мінімізує вплив помилок.

Такий розподіл ролей користувачів дає змогу нечисленному персоналу досягати максимального ефекту, звільняє лексикографа від рутинної роботи, переключаючи його увагу на підвищення якості подання матеріалу, що дозволяє:

— сформувати з відібраного текстового матеріалу електронну версію текстового масиву для подальшої обробки;

- визначати кількісні характеристики і частотні параметри тексту;

- для обраних слів динамічно формувати електронну картку і дублювати її на паперовому носії як виділений текстовий фрагмент за певними правилами.

Після завершення першого етапу накопичення початкової «критичної» маси текстового матеріалу і створення повноцінної ЛК передбачається використовувати її для вирішення задач комп'ютерної лексикографії і створення актуальних словників української мови (передусім тематичних). Відомі два підходи до створення словників, заснованих на інформаційних технологіях і програмних рішеннях: перший - від так званого паперового словника до його комп'ютерної версії; другий - від тексту, що вводиться в комп'ютер з електронних чи паперових носіїв, до комп'ютерного лексикографічного об'єкта. Власне на другому шляху знаходиться розробка ЛК.

Задача лематизації. Перетворюючи машиночитаний текст в електронну ЛК, доволі складно побудувати алгоритм лематизації. Щоправда, європейська лексикографічна практика відмовилася від формування леми для конкретної картки, мотивуючи це тим, що досить часто виникає потреба у прикладі на конкретну парадигматичну форму слова. За умови підпису кожної картки лемною формою слова дослідникові доводиться переглядати усі (в електронній ЛК можуть бути сотні ілюстрацій) картки на виділене слово замість того, щоб одразу працювати з ілюстративним матеріалом на конкретну форму. З огляду на це високим ступенем працездатності характеризується запропонована схема гіперкартки, коли вхід відбувається через лему і наступним кроком є вибір гіпокартки з гіперкартки, а далі конкретної картки на конкретну словоформу. Поза тим, реалізація алгоритму виведення лексикографічної форми слова, тобто алгоритму лематизації, необхідна для автоматичного формування будь-якого заданого реєстру на електронній текстовій базі.

Алгоритм лематизації характеризується двома типами складності: структурною (лінгвістичною)

і обчислювальною (програмною). Структурна складність алгоритму для української мови мотивується її внутрішньою природою: синтетичність мови, частиномовний поділ, категоріальні характеристики кожної з частин мови і відмінюваність/невідмінюваність слів.

Відомо, що в українській мові виділяється 10 лексико-граматичних класів слів *(іменник, займенник, прикметнім, числівник, дієслово, прислівник, прийменник, частка, сполучник, вигук).* Для порівняння (див. табл. 2 і 3) у схемі європейського проекту лематизації таких мов, як англійська (EN), румунська (RO), словацька (SL), чеська (CS), болгарська (BG), естонська (ET), угорська (HU), виділено такі частини мови: іменник *(Noun),* дієслово *(Verb),* прикметник *(Adjective},* займенник *(Pronoun),* означуване слово *(Determiner),* артикль *(Article),* прислівник *(Adverb},* прийменник *(Adpositiori),* сполучник *(Conjunction),* числівник *(Numeral),* вигук *(Interjection),* інші *(Residual),* абревіатура *(Abbreviation),* частка *(Particle)* [8]. Зіставивши наведені частиномовні поділи слів і врахувавши відмінюваність українських порядкових числівників за прикметниковою схемою, доцільно дещо модифікувати традиційний лексико-граматичний розподіл в українській мові з огляду на реалізацію алгоритму лематизації. Модифікований лексико-граматичний поділ охоплює такі частини мови: іменник, дієслово, прикметникові слова (прикметник + порядковий числівник), займенник, числівник (кількісний), прислівник, прийменник, сполучник, частка, абревіатура, вигук. Далі виділені частини мови поділяються на змінні і незмінні.

Незмінні частини мови не передбачають аналізу алгоритмом лематизації, оскільки формально слова, які до них належать, залишаються незмінними за будь-яких умов, крім граматичної помилки, і лексикографічне слово дорівнює лемі. Натомість наявність змінних частин мови, особливо іменника і дієслова, в українській мові мотивує чималу структурну складність. Так, у межах українського іменника традиційно визначають морфологічні граматично незалежні категорії *рід, число, відмінок,* і власне ці категорії є носіями загальнокатегорійного значення предметності. Крім того, в межах українського іменника слова розподіляються на лексико-граматичні розряди: власні/загальні назви, істоти/неістоти, конкретні/абстрактні, речовинні, збірні. Кожна з цих ознак іменника має свою специфіку інтерпретації в алгоритмі лематизації. З погляду лематизації категорії та розряди іменника групуються у так звані лематизаційні й анотаційні атрибути. Крім того, інколи певна категорійна або розрядна інформація ідентифікується комбінацією символів, що виступають теґами конкретної лексичної одиниці. Отже, категорійно-розрядний аналіз українського іменника для лематизації має схему, виведену у табл. 2 за принципом, представленим проектом *Multext-East,* який схематично за запозиченою табл. 3 аналізує категорії іменника в типологічне різних мовах. Автори статті з метою власних досліджень ввели до цієї таблиці українську мову (див. останній стовпчик).

Таблиця 2. Перелік лематизаційно залежних категорій та розрядів іменника української мови

Український термін	Англійський термін	Символ атрибута
I. Лематизаційні атрибути:		
Тип	type:	
загальна назва	common	c
власна назва	proper	P
Рід	gender:	
чоловічий	masculine	m
жіночий	feminine	f
середній	neuter	n
спільний	common	eg
подвійний	epicoenon	e
Число	number:	
однина	singular	s
множина	plural	P
двоїна	dual	d
злічувані	count	t
plurale tantum		pt
Відмінок	case:	
називний	nominative	n
родовий	genitive	g
давальний	dative	d
знахідний	accusative	a
орудний	instrumental	i
місцевий	locative	l
кличний	vocative	v
прямий	direct	r
непрямий	oblique	o
II. Анотаційні атрибути:		
Особовість	definiteness:	
наявна (так)	yes	y
відсутня (ні)	no	n
Клітини	clitic:	
наявна (так)	yes	y
відсутня (ні)	no	n
Істота/неістота	animate:	
істота - (так)	yes	y
неістота - (ні)	no	n:
конкретний	concrete	et
речовинний	material	mt
абстрактний	abstract	at
Числова характеристика	number specificity:	
збірні	collective	cl
Відміна	declension:	dl
	1	1dl
	2	2dl
	3	3dl
	4	4dl
Група	declension group:	
тверда	hard	dlh
мішана	mixed	dim
м'яка	soft	dls

Крім розбудованої категорійно-розрядної структури, іменник української мови характеризується ще й системою відмінкових закінчень, складною з погляду лематизації, з високим ступенем омонімії.

Таблиця 3. Типологія категорії іменника

		C	EN	RO	SL	CS	BG	ET	HU	UA
Тип (Type)	загальна назва (common)	c	x	x	x	x	x	x	x	x
	власна назва (proper)	p	x	x	x	x	x	x	x	x
Рід (Gender)	чоловічий (masculine)	m	x	x	x	x	x			x
	жіночий (feminine)	f	x	x	x	x	x			x
	середній (neuter)	n	x	x	x	x	x			x
Число (Number)	однина (singular)	s	x	x	x	x	x	x	x	x
	множина (plural)	p	x	x	x	x	x	x	x	x
	двоїна (dual)	d			x	x				
l.s.	злічувані (count)	t					x			x
Відмінок (Case)	називний (nominative)	n			x	x	x	x	x	x
	родовий (genitive)	g			x	x		x	x	x
	давальний (dative)	d			x	x			x	x
	знахідний (accusative)	a			x	x			x	x
	кличний (vocative)	v		x		x	x			x
	місцевий (locative)	l			x	x				x
	орудний (instrumental)	i			x	x			x	x
l.s.	прямий (direct)	r		x						x
l.s.	непрямий (oblique)	o		x						x
l.s.	партитив (partitive)	1						x		
	ілятив (illative)	x						x	x	
	inessive	2						x	x	
	елятив (ellative)	e						x	x	
	алатив (allative)	t						x	x	
	адесив (adessive)	3						x	x	
	аблятив (ablative)	b						x	x	
l.s.	транслятив (translative)	4						x		
	термінатив (terminative)	9						x	x	
l.s.	абесив (abessive)	5						x		
l.s.	комітатив (komitative)	k						x		
l.s.	адитив (aditive)	7						x		
l.s.	темпоратив (temporalis)	m							x	
l.s.	казуатив (causalis)	c							x	
l.s.	сублятив (sublative)	s							x	
l.s.	делятив (delative)	h							x	
l.s.	соціатив (sociative)	q							x	
l.s.	фактив (factive)	y							x	
l.s.	суперлятив (superessive)	p							x	
l.s.	дистрибутив (distributive)	u							x	
Означеність (Definiteness)	ні (no)	n		x			x			
	так (yes)	y		x			x			
l.s.	короткий атрикль (short_art)	s								
l.s.	повний артикль (full_art)	f								
Клітика (Clitic)	ні (no)	n		x						x
	так (yes)	y		x						x
Істота/неістота (Animate)	ні (no)	n				x			x	x
	так (yes)	y				x			x	x
Категорійність числа (Owner_Number)	однина (singular)	s							x	
	множина (plural)	p							x	
Категорійність особи (Owner_Person)	перша (first)	1							x	
	друга (second)	2							x	
	третя (third)	3							x	
Числова належність (Owned_Number)	однина (singular)	s							x	
	множина (plural)	p							x	

Наприклад, лише закінчення -а (не враховуючи варіанта **-я)** може вказувати на ряд іменникових значень, звідки й виникає проблема однозначного виведення леми:

a) називний відмінок (Н), жіночий рід, тверда або мішана група, I відміна, однина;
b) називний відмінок, середній рід, IV відміна з суфіксом -ат-, -ят-, -ам-, однина;
c) родовий відмінок (Р), чоловічий рід, тверда або мішана група, II відміна, однина;
d) родовий відмінок, середній рід, тверда або мішана група, II відміна, однина;
e) родовий відмінок, середній і чоловічий рід, II відміна, однина;
f) знахідний відмінок (З), чоловічий рід, тверда або мішана група, II відміна, однина;

Таблиця 4а. Зведена схема відмінювання іменників української мови з урахуванням наголошування, однина

<table>
<tr><td></td><td colspan="3">I відміна</td><td colspan="6">II відміна</td><td>III відміна</td><td colspan="2">IV відміна</td></tr>
<tr><td></td><td colspan="3">жіночий рід</td><td colspan="3">чоловічий рід</td><td colspan="3">середній рід</td><td>жіночий рід</td><td colspan="2">середній рід</td></tr>
<tr><td></td><td>тв. гр.</td><td>міш. гр.</td><td>м'як. гр.</td><td>тв. гр.</td><td>міш. гр.</td><td>м'як. гр.</td><td>тв. гр.</td><td>міш. гр.</td><td>м'як. гр.</td><td></td><td>-ат-
-ят-
-ам-</td><td>-ен-</td></tr>
<tr><td>Н.(n)</td><td colspan="2">-а</td><td>-я</td><td colspan="3">-∅</td><td>-о
-о'</td><td colspan="2">-е</td><td>-∅</td><td>-а'</td><td>-я'</td></tr>
<tr><td rowspan="2">Р.(g)</td><td rowspan="2">-и</td><td colspan="2">-і</td><td colspan="2">-а</td><td>-я</td><td rowspan="2">-а</td><td rowspan="2">-я</td><td rowspan="4">-і</td><td rowspan="2">-и</td><td rowspan="2">-і</td></tr>
<tr><td></td><td>-ї</td><td colspan="2" rowspan="2">-а'</td><td>-я'
-ю</td></tr>
<tr><td>Д.(d)</td><td colspan="3">-і</td><td colspan="2">-у</td><td>-ю</td><td rowspan="3">-у</td><td rowspan="3">-ю</td><td colspan="2" rowspan="3">-і</td></tr>
<tr><td colspan="2" rowspan="3"></td><td rowspan="3">-ї</td><td rowspan="3">-ові</td><td>-у'</td><td>-ю'</td></tr>
<tr><td>-еві</td><td>-еві</td></tr>
<tr><td>-еві'</td><td>-еві'
-єві</td></tr>
<tr><td>З.(a)</td><td colspan="2">-у</td><td>-ю</td><td colspan="2">-а</td><td>-я</td><td>-о</td><td colspan="2" rowspan="2">-е</td><td rowspan="2">-∅</td><td rowspan="2">-а'</td><td rowspan="2">-я'</td></tr>
<tr><td colspan="2"></td><td></td><td>-а'</td><td></td><td>-я'</td><td rowspan="2">-о'</td></tr>
<tr><td colspan="6">-∅</td></tr>
<tr><td>О.(i)</td><td>-ою</td><td colspan="2">-єю</td><td rowspan="2">-ом</td><td colspan="2">-ем</td><td rowspan="2">-ом</td><td rowspan="2">-ем</td><td rowspan="2">-ю</td><td rowspan="2">-а'м</td><td>-я'м</td></tr>
<tr><td colspan="2"></td><td>-єю</td><td colspan="2">-е'м</td><td>-ем</td></tr>
<tr><td>М.(l)</td><td colspan="3">-і</td><td colspan="3">-і
-і'</td><td rowspan="4">-і</td><td rowspan="4">-і</td><td colspan="2" rowspan="4">-і</td></tr>
<tr><td colspan="2" rowspan="3"></td><td rowspan="3">-ї</td><td rowspan="2">-ові</td><td colspan="2">-еві
-еві'</td></tr>
<tr><td colspan="2" rowspan="2">-у</td><td>-ю
-ю'
-ї</td></tr>
<tr><td>Кл.(v)</td><td>-о</td><td colspan="2">-е</td><td>-у</td><td>-ю
-ю'</td><td rowspan="2">-о</td><td colspan="3">-е</td><td>-е</td><td>-а'</td><td>-я'</td></tr>
<tr><td></td><td>-є</td><td>-е</td><td></td><td colspan="5"></td></tr>
</table>

g) знахідний відмінок, середній рід, IV відміна з суфіксом -ат-, -ят-, -ам-, однина;
h) кличний відмінок (Кл), середній рід, тверда або мішана група, II відміна, множина;
i) кличний відмінок, середній і чоловічий рід, II відміна, множина;
j) кличний відмінок, жіночий рід, III відміна, множина;
k) кличний відмінок, IV відміна з суфіксом -ат-, -ят-, -ам-, однина;
l) кличний відмінок, середній рід, IV відміна з суфіксом -ат-, -ят-, -ам-, множина.

Структурна складність іменникового відмінювання доволі чітко простежується у зведеній схемі відмінювання іменників української мови з урахуванням наголошування (табл. 4а, 4б). Очевидно, що за таких флексійних умов побудова алгоритму лематизації (особливо детермінованого) є складною задачею. Слід врахувати, що зведення форм іменника є дещо простішим завданням, ніж лематизація дієслова, структурно складнішого в українській мові за іменник.

Обчислювальна складність алгоритму лематизації пов'язана передусім з поданням правил природної мови програмними засобами, що вимагає значної формалізації граматики природної мови (у нашому випадку української), досягнення регулярності умов застосування конкретного правила. Зауважимо, що з метою зменшення обчислювальної складності доцільно послуговуватися таблицями, деревами, розміченими масивами, сукупністю словників (стоп-словник, словник тегованих лем, частотний словник тощо).

Резюме. По-перше, перехід від традиційної лексикографії до комп'ютерної починається з побудови електронної ЛК, яка може бути архітектурно простою або зі складною системою ідентифікаційних та інформаційних полів з послідовним або безпосереднім переходом від одного до іншого поля у прямому і зворотному порядку. Проста архітектурна система ЛК має ієрархічну будову від гіперкартки через гіпокартку до картки. По-друге, важливим є вибір загального методу і часткових методик добору фактичного матеріалу в конкретних випадках побудови ПБД. По-третє, простий, а тим більше детермінований алгоритм лематизації для української мови, який характеризується двома типами складності, є проблемним завданням через розбудовану категорійно-розрядну структуру українських лексико-граматичних груп слів та флексійних умов і відсутність формалізованої граматики української мови.

Таблиця 46. Зведена схема відмінювання іменників української мови з урахуванням наголошування, множина

	I відміна			II відміна						III відміна	IV відміна	
	жіночий рід			чоловічий рід			середній рід			жіночий рід	середній рід	
	тв. гр.	*міш. гр.*	*м'як. гр.*	*тв. гр.*	*міш. гр.*	*м'як. гр.*	*тв. гр.*	*міш. гр.*	*м'як. гр.*	*мати*	*-ат- -ят- -ам-*	*-ен-*
Н.(n)	-и	-і		-и	-і'		-а'	-а	-я'	-і	-а	-а'
			-ї	-и'		-ї'						
Р.(g)	-∅			-і'в			-∅			-ей	-∅	
				-ів		-ї'в						
Д.(d)	-ам		-ям	-ам		-я'м	-а'м	-ам	-я'м	-ам	-ам	-а'м
				-а'м						-а'м		
										-ям		
										-я'м		
З.(a)	-и	-і	-∅	-і'в			-а'	-а	-я'	-і	-а	-а'
			-ї	-ів		-ї'					-∅	
О.(i)	-ами		-ями	-ами		-я'ми	-а'ми	-ами	-я'ми	-ами	-ами	-а'ми
				-а'ми						-а'ми		
										-ями		
										-я'ми		
М.(l)	-ах		-ях	-ах		-я'х	-а'х	-ах	-я'х	-ах	-ах	-а'х
				-а'х						-а'х		
										-ях		
										-я'х		
Кл.(v)	-и	-і		-и'	-і'		-а'	-а	-я'	-і	-а	-а'
			-ї	-и		-ї'						

1. *Бурячок А. А.* Лексична картотека- К.: Українська енциклопедія ім. М. П. Бажана, 2000.- 750 с.
2. Сучасна українська літературна мова. Морфологія.- К.: Наукова думка, 1969.- 250 с.
3. *Carroll J. B., Davies P., Richman B.* The American Heritage Word Frequency Book.- NY: American Heritage Publishing Co. (Boston: Houghton Mifflin Company), 1971- 160 p.
4. *Первовозчикова О. Л., Сичкаренко В. А.* Интернационализация и локализация прикладных платформ и приложений в корпоративных информационных системах // Управляющие системы и машины.- 2000.- № 5/6.- С. 43-58.
5. ISO/IEC 9075-2:1999 Information technology - Database languages - SQL - Part 2: Foundation (SQL/Foundation).- Geneva: JTC1 ISO/IEC, 1999.- 1140 p.
6. Компьютерный корпус текстов русских газет конца XX века: создание, категоризация, автоматизированный анализ языковых особенностей / Б. В. Виноградова, О. В. Кукушкина А. А. Поликарпов, С. О. Савчук // Русский язык: исторические судьбы и современность.- М.: Изд-во Московского университета, 2001.-С. 393-398.
7. *Демская-Кульчицкая О. М.* Корпус текстов украинской периодики // Исследование славянских языков в русле традиций сравнительно-исторического и сопоставительного языкознания- М.: Изд-во Московского университета, 2001.- С. 26-28.
8. MULTEXT-East: Multilingual Text Tools and Corpora for Central and Eastern European Languages. Проект Копернікус. «Багатомовні текстові інструменти і корпуси для центрально-східноєвропейських мов».- http://nl.ijs.si/ME, 2001.

O. M. Demska-Kulchytska, O. L. Perevozchikova, V. A. Sichkarenko

EVOLUTION OF UKRAINIAN LANGUAGE LEXICAL REPOSITORY

The theoretic and information aspects of building the computer oriented lexicographic card-file on the base of a text database and other sources of text data located in arbitrary medium are discussed. The medium can be paper, electronics, telex, facsimile, etc. The directions of implementation the input, analyses, processing and storage subsystems are discussed.