На правах рукопису
УДК 658.012.011.56

КОВТУНЕНКО ЛЮДМИЛА СЕРГІЇВНА

КОМП'ЮТЕРНО-ІНФОРМАЦІЙНІ АСПЕКТИ СУЧАСНОЇ УКРАЇНСЬКОЇ ЛЕКСИКОГРАФІЇ

Спеціальність - 05.25.05 Інформаційні системи та процеси.

Автореферат
дисертації на здобуття наукового ступеня
кандидата технічних наук

Київ - 1995

Роботу виконано в Українськ[illegible] академії наук України.

Науковий керівник:
Кандидат фізико-математичних наук В. А. Широков

Офіційні опоненти:
Доктор технічних наук, професор В. Я. Рубан
Кандидат технічних наук Л. Й. Костенко

Провідна установа:
Інститут проблем реєстрації інформації НАН України

Захист відбудеться "11" 05 1995 року о 14 год. на засіданні Спеціалізованої вченої ради Д.01.31.01 по захисту дисертацій на здобуття наукового ступеня доктора наук (кандидата наук) в Центральній науковій бібліотеці ім. В. І. Вернадського за адресою: м.Київ, просп. 40-річчя Жовтня, 3.

З дисертацією можна ознайомитися в Центральній науковій бібліотеці ім. В. І. Вернадського Національної академії наук України (м. Київ).

Автореферат розіслано "5" 04 1995 р.

Вчений секретар
спеціалізованої ради
кандидат економічних наук Чекмарьов А.О.

Актуальність дослідження.

Розв'язання різноманітних завдань, які стосуються побудови інформаційно-пошукових та експертних систем, систем штучного інтелекту, автоматизації наукових досліджень тощо передбачає вирішення проблем спілкування людини з комп'ютером природною мовою.

Традиційний погляд на мову як на стихію, непридатну для формалізації внаслідок її складності, наявності великої кількості виключень, системна роль яких не поступається правилам мови, в останній час підлягає значній ерозії, і натомість все більшого поширення набуває точка зору, що лінгвістика, тим більше в її комп'ютерній іпостасі, є наукою точною в загальнонауковому значенні (Пещак М.М., Широков В.А. Структурні моделі лексикографічних систем. Тези 3-ї Міжнародної наукової конференції "Проблеми української науково-технічної термінології": Львів, 1994). Усталенню цього погляду вирішальним чином сприяють процеси інформатизації, які в останнє десятиріччя набули глобального характеру. Можливість спілкування з комп'ютером природною мовою з абстрактної наукової проблеми перетворюється в нагальну потребу, від розв'язання якої залежить ефективне функціонування інтелектуальних систем, що становлять ядро нових інформаційних технологій (Рубан В.Я., Медведев А.А. Применение информационного языка для повышения эффективности проектирования АСУ //НТИ, сер. 2, 1977, №11, 12, с. 43-48; Проектирование и функционирование интегрального словаря по юриспруденции. //НТИ, сер. 2, 1993, №1, с. 9-16). Національні програми з інформатизації розвинених країн містять як органічні елементи розвиток методів лінгвістичного забезпечення інформаційних систем різного рівня та функціональної спрямованості. Не є винятком в цьому смислі й відповідна національна програма України (Указ Президента України "Про державну політику інформатизації України" від 31 травня 1993 р. № 186/93; Постанова

Кабінету Міністрів "Питання інформатизації" від 31 серпня 1994 р. № 605). Врахування сучасного статусу української мови, а також відомих змін, що відбулися в українському правописі та лексиці протягом останніх років, переводить проблематику розвитку методів комп'ютерної лексикографії до розряду першочергових.

Формулювання лінгвістичних проблем у термінах моделей баз даних та знань, широке розповсюдження інтелектуальних лінгвістичних систем, засобів розуміння комп'ютером природних текстів та вилучення з них знань, систем автоматичного перекладу, редагування, реферування тощо викликало до життя і активного застосування різноманітні програмні засоби - текстові та лінгвістичні процесори, гіпертекстові та експертні системи, а також велику кількість експериментальних програм, моделюючих ті чи інші мовні явища.

Без перебільшення можна стверджувати, що в поточному й наступному десятиріччях розвитку інформатики створення людино-машинних інтерфейсів, заснованих на природній мові, стане одним з головних. Це, у свою чергу, висуває перед комп'ютерною лексикографією серйозні завдання, від вдалого розв'язання яких безпосередньо залежить якість і ефективність функціонування розроблюваних інформаційних систем.

Об'єкт дослідження. Створення інтелектуальних автоматизованих систем базується на формалізованих знаннях про їх об'єкти. Слова природної мови в цьому відношенні виступають ефективним об'єктом досліджень й дозволяють розробку таких формальних методів, які допускають інтерпретації у вигляді чітких алгоритмів. Розвиток алгоритмічної бази сучасної лінгвістики створює необхідну передумову для автоматизації лінгвістичних досліджень, послідовно ускладнюючи завдання, розв'язання яких стало можливим за допомогою комп'ютерної техніки.

Об'єктом дослідження стає словниковий фонд природної (тут української) мови як необхідного посередника людського спілкування, комунікативна функція якої поширилася і на людино-машинні системи.

Включення природної мови до компонентів нових інформаційних технологій спонукає до усвідомлення самої мови як системи інформаційної з необхідністю інтерпретації феноменів лексики у термінах інформаційних наук.

Предметом дослідження є, власне, українська лексика, алгоритмічний апарат та програмні засоби її граматичного, морфологічного аналізу, автоматизовані засоби побудови парадигматичної класифікації української мови. Ми досліджуємо лексикографічні феномени, розглядувані в їх зв'язку та через призму комп'ютерної техніки. При цьому висвітлюються два основні аспекти взаємодії словникарства з інформатикою. У першому комп'ютер розглядається як технологічне знаряддя для розв'язання класичних проблем лексикографії таких, як типологія та класифікація словників, розробка теорії структури словникових статей, лексико-семантичні дослідження тощо, а також для укладання словників традиційного типу. У другому аспекті сама лексикографія виступає у функції прикладної дисципліни при створенні елементів лінгвістичного забезпечення інформаційно-комп'ютерних систем, зокрема: машинних тезаурусів, автокоректорів, систем автоматичного перекладу та реферування тощо.

Завдання дослідження полягає у створенні алгоритмічного апарату мовних класифікацій та розробці комп'ютерних систем, які були б орієнтовані на комплексне розв'язання проблем сучасної української лексикографії. Такі стендові (на відміну від "вбудованих" машинних словників) лексикографічні системи характеризуються розвиненим філологічним інтелектом, певною функціональною повнотою, зручним інтерфейсом і поєднують як технологічні, так і дослідницькі можливості в лексикографічній роботі.

Алгоритмічне представлення мовного об'єкта дозволяє виявити певні закономірності у поведінці лексикографічної системи, до якої він входить, взаємозв'язок її компонентів, вивчити її характеристики, виражені на рівні форми.

Проблема лінгвістичного забезпечення комп'ютерних систем полягає у створенні досить складних програмних засобів для обробки текстової інформації природною мовою, укладання словників, а також автоматизації роботи лексикографів. Така інформація представлена у вигляді текстів та різних типів словників (орфографічних, тлумачних, перекладних тощо). Виникає потреба в побудові якісних комп'ютерних словників як інструменту для роботи з мовою і над мовою, а це включає щонайменше дослідження й редагування текстів, вивчення мови і переклад, пошук інформації та автоматизовану обробку даних, спілкування з комп'ютером тощо.

В традиційній комп'ютерній лексикографії виділяються три головні напрямки. Перший - це автоматична побудова словників шляхом комп'ютерної обробки тексту. Вказаний напрям розвивався з самого початку як допоміжний засіб, що полегшував рутинну роботу у філологічних дослідженнях, і застосовувався при побудові словників, реєстрів, конкордансів тощо.

Другий напрямок пов'язаний з розробкою комп'ютерних словників як внутрішніх елементів інформаційних систем. Вони виконують роль інформаційно-пошукових тезаурусів і застосовуються в різних типах автоматизованих систем, при штучному перекладі тощо. Для словників такого типу характерна розвинена та строго формально визначена структура, досить багатий сервіс.

Нарешті, третій напрямок розглядає комп'ютерні словники як машинні версії традиційних. У межах цього напрямку ставляться і розв'язуються такі проблеми, як типологія словників, теорія структури словникових статей; виникає можливість проведення лексико-семантичних досліджень.

У сучасній комп'ютерній лексикографії виникла і набирає все більш окреслених рис тенденція до синтезу всіх трьох напрямків з явним виділенням когнітивного аспекту лінгвістичної діяльності.

Наукова і практична діяльність потребує усвідомлення та обґрунтування великої кількості різних правил, засобів, якими лексикографи оперують при створенні словників. Вивчення цих законів та їх алгоритмізація дозволяє інтерпретувати лексикографічну діяльність у вигляді автоматизованої системи. Такі системи в науковій літературі називають лінгвістичним або лексикографічним процесором, один з варіантів якого запропоновано у дисертації.

Мета дослідження - розробити лексикографічний процесор природної мови. Основним системним інструментом такого процесора виступає комп'ютерний словник, що являє собою програмний комплекс з алгоритмічною частиною, базами даних, знань, засобами системного сервісу та статистичної обробки.

Функціональна структура такого процесора ґрунтується на інтеграції в діалоговому режимі ряду процедур, які дозволяють виконувати певні операції над текстами та словниками: формулювати умови відбору слів до словника, визначати структури словникових статей, автоматично формувати стандартну граматичну й стилістичну характеристики слова, транскрибувати слово відповідно до законів звукової системи мови, автоматично виділяти шрифти для лівої та правої частин словника, який передається до видавництва для публікації, будувати індекси та інверсійні словники тощо. Для програмної реалізації цих функцій необхідно поєднати наукову діяльність філологів, обробку мовного матеріалу з сучасними комп'ютерними засобами та програмним забезпеченням так, щоб лексикографічний процесор дійсно став інструментом комп'ютерного конструювання й підготовки словників.

Важливим етапом цього процесу є структурування філологічних правил обробки текстів і словників у тісному зв'язку з правилами комп'ютерної

алгоритмізації на основі дослідження та формалізації представлення відповідних лінгвістичних фактів.

Методом дослідження є комп'ютерне моделювання лінгвістичних та лексикографічних феноменів. Систематично використовується формальне представлення мовного матеріалу та трансформація його до структурованих інформаційних систем (баз даних та знань). Формальна репрезентація об'єктів мови дає можливість послідовного використання методів алгоритмічного аналізу з подальшою програмною реалізацією інформаційних та лінгвістичних функцій.

Наукова новизна. Запропонований автором формалізований метод обробки комп'ютерних текстів на базі автоматичного орфографічного словника за допомогою алгоритмічно побудованих класифікацій та правил українського правопису не має аналогів серед відповідних розробок і досліджень у галузі комп'ютерної україністики.

Розроблений лексикографічний процесор являє собою програмно-лінгвістичний механізм, створений на базі складноструктурованої системи правил та класифікацій над мовним знаком, що виступає у вигляді інтегрованого багатокомпонентного програмного комплексу, орієнтованого на розв'язання проблем генерації та ведення словникових систем різних типів, а також виконання лексикографічних досліджень.

На основі формалізації підходу до аналізу слова як конструкції, складеної з літер і літеросполук, алгоритмічно побудовано правила парадигматизації та лематизації текстових слів (ці правила до певної міри виявилися універсальними і не потребують втручання спеціалістів-філологів на етапі поповнення словника новими словами і під час роботи з текстами) без введення "зайвих і складних" позначень та таблиць, як це зроблено, наприклад, в Грамматическом словаре русского языка (А.А. Зализняк, -М. : Русский язык, 1980) і відповідному комп'ютерному словнику.

На базі запропонованих автором алгоритмів побудовано одну з ланок лексикографічного процесора - принципово новий граматичний автоматизований

словник, який є основним інструментом для роботи з україномовними текстами та словниками.

Практична значимість та реалізація. За допомогою розробленої автором комп'ютерної лексикографічної системи була одержана парадигматична класифікація української мови, яку покладено в основу всієї системи автоматизованого укладання україномовних словників. Система дозволяє одержувати повну парадигму змінюваних частин мови, виходячи як з реєстрового слова, так і з будь-якої текстової словоформи, що разом з іншими функціями робить її ще й граматичним довідником. В результаті автоматизованої обробки орфографічної словникової бази даних за допомогою розробленого лексикографічного процесора та допоміжних для поліграфічної системи програм за короткий час було здійснене видання "Орфографічного словника української мови" (обсягом близько 120 тисяч слів), який відповідає існуючому українському правопису.

Апробація роботи. Основні положення та результати дослідження доповідались на науковій конфіренції "Проблеми створення машинних фондів мов" (Київ, 1991), I-III Міжнародних наукових конференціях "Українська науково-технічна термінологія" (Львів, 1992-1994), на міжнародній науковій конференції "Автоматизовані системи інформаційно-бібліотечного обслуговування" (Київ, 1994).

На захист виносяться наступні положення:

1. На основі дослідженої внутрішньої структури словникових статей різних типів словників пропонується автоматизована структурна модель лексикографічних систем.

2. Запропоновано формалізований метод побудови словникових баз даних, виходячи з текстової форми словника.

3 Розвинено алгоритмічний апарат і програмне забезпечення для одержання парадигматичної класифікації української мови.

4. Розроблено варіант лексикографічного процесора як інтегрованої комп'ютерно-інформаційної системи, орієнтованої на комплексне розв'язання проблем сучасної української лексикографії.

Структура роботи. Дисертація складається з вступу, трьох глав, у яких викладено основний зміст, висновків та списку використаної літератури.

ЗМІСТ РОБОТИ

У першому параграфі **першої глави** дається огляд існуючих підходів та результатів у галузі комп'ютерної лексикографії, а також поглиблений аналіз відомих підходів та програмних засобів, що стосуються розглядуваної тематики.

Другий параграф першої глави присвячено розробці формальних засад теоретичної лексикографії та викладу елементів структурної теорії словникових систем. Вона базується на дихотомічній структурі слова як мовного знака. В розробленій моделі зазначена структура представлена на зразок переважної більшості виданих словників, де в реєстрі повнозначні частини мови подані вихідними (канонічними) словоформами, а в моделі становлять підмножину $S0(L) \subset S(L)$, де через S(L) позначено множину слів конкретної природної мови L, описаних виданими досі словниками і включених до лексикографічної моделі. З наведеного випливає можливість представлення словникової системи у вигляді такої декомпозиції:

$$D[S(L)] = \{ S0(L); DF[S(L)], DC[S(L)] \}, \qquad (1)$$

де через DF[S(L)] позначено формальну частину опису множини S(L), а через DC[S(L)] - її змістовну частину.

Узагальненість дефініцій DF[S(L)] та DC[S(L)] дозволяє досить оперативно користуватися системою (1) при лексикографічній кваліфікації слова і як одиниці тексту, і як одиниці словника.

Дихотомічність структури слова як мовного знака відображається в моделі за допомогою виділення у словникових статтях реєстрової та інтерпретаційної частин. Як відомо, реєстрова частина називається лівою, а інтерпретаційна - правою частиною. Звідси загальна формула представлення будь-якого словника в лексикографічній моделі матиме такий вигляд:

$$V(L) = \{ \Lambda(L);\ P(L),\ H \} \qquad (2)$$

В останній формулі через V(L) позначено словник як множину словникових статей; $\Lambda(L)$ є множиною лівих частин словникових статей словника V(L); P(L) - множина правих частин словникових статей цього ж словника. H - відображення множини $\Lambda(L)$ на P(L):

$$H : \Lambda(L) \rightarrow P(L) \qquad (3)$$

Терміни "ліва" й "права" частини є до певної міри умовними, оскільки межа між фрагментами структури словникової статті, які позначаються цими термінами, не завжди однолінійна. Лексикографічне розмежування лівої і правої частин, таким чином, стосується не стільки формально позиційного їх розташування в словниковій статті, скільки відображення функціонального протиставлення форми та змісту в слові.

У структурі друкованих словників, як і в лексикографічній моделі, і, відповідно, в комп'ютерних словникових системах, словникова стаття починається реєстровим словом x, яке, разом з тим, вважається її ідентифікатором. Отже, формулу (2) словника можна деталізувати таким чином:

$$V(L) = \bigcup_{x \in S0(L)} V(x) \qquad (4)$$

$$\Lambda(L) = \bigcup_{x \in S0(L)} \Lambda(x); \qquad P(L) = \bigcup_{x \in S0(L)} P(x),$$

де V(x) - словникова стаття, яка очолюється реєстровим словом x, а Λ(x) і P(x) відповідно ліва та права частини цієї словникової статті. Таким чином, H (Λ(x)) = P(x). На множині V(L) визначається частковий порядок, індукований "лексикографічним" упорядкуванням множини S(L). Для словника V(L) може існувати автоморфізм. тобто відображення

$$A : V(L) \rightarrow V(L) \qquad (5)$$

яке констатує наявність відсилочних словникових статей.

Таким чином, відображення H та A породжують макроструктуру словника V(L). Крім цього, кожен словник має ще й свою внутрішню мікроструктуру, яка відображає у неявному вигляді семантику предметної області, що є об'єктом конкретного словника. Вказана мікроструктура стосується будови об'єктів Λ(x) і P(x).

Розглянемо приклади побудови словникових структур для україномовних словників. Найпростішою є мікроструктура *орфографічного* словника , для якого:

$$\Lambda(U) = S0(U) \quad \text{і} \quad \Lambda(x) = x. \qquad (6)$$

Права частина P(L) орфографічного словника являє собою множину парадигматичних показників. Для кожного x відповідна P(x) є спеціальним чином організована послідовність флексій або квазифлексій. Фактично системна роль правої частини в орфографічному словнику зводиться до зіставлення кожній реєстровій одиниці x особливостей її словозміни, завдяки чому будується повна парадигма [x].

На відміну від паперового варіанту орфографічного словника, в якому функція H, що об'єднує ліву та праву його частини в єдину словникову статтю :

$$H(\Lambda(x)) = P(x),\ V(x) = \{ \Lambda(x), H(\Lambda(x)) \}$$

не є явно визначеною, в комп'ютерному його аналізі зазначена функція грає активну роль. Її можна інтерпретувати як оператор, який відображає реєстрову одиницю x в її повну парадигму [x] :

$$H : x \rightarrow [x] \qquad (7)$$

Отже, в комп'ютерному варіанті орфографічного словника H(x) = [x]. Іншими словами, в інформаційній системі функція H виконує роль інтерфейса між Λ(U) та P(U). Будова функції H є досить складною і являє собою сукупність взаємодіючих блоків.

Після того, як реєстровій одиниці приписані значення атрибутів, що розташовані у відповідних вузлах, тобто після граматичної ідентифікації лексеми x, вона потрапляє до парадигматичного блока. В ньому реєстрова одиниця, перетворюючись, приймає вид, адаптований для парадигматичного аналізу :

$$x \rightarrow i(x) \ * \ \omega(x) , \qquad (8)$$

де через ω (x) позначено квазіфлексію слова x довжиною до п'яти літер, а зірочкою позначається конкатенація. Власне саме ця частина слова піддається парадигматичному аналізу, що являє собою набір правил виводу, згідно з якими за ω(x) будується повна парадигма [x].

В останньому блоці за одержаною парадигмою будується права частина P(x) словникової статті з даним словом x.

Аналогічно розглянуто приклади побудови структури орфоепічного і тлумачного словників української мови. Проаналізовано зв'язок між елементами структури словників, структурами відповідних баз даних та поліграфічним оформленням словникових статей. Завдяки цьому виявляється можливою розробка технологічної системи для укладання словників різних типів за допомогою автоматизації таких функцій :

— генерація узагальненої абетки словника;

— породження структури словникових статей конкретного словника і формування відповідної бази даних;

— граматична та лексико-семантична ідентифікація об'єктів словника, включаючи функції лематизації та автоматичної побудови парадигми для повнозначних відмінюваних частин мови;

— проведення лексикологічних досліджень;

— злиття словників та одержання підсловника з даного словника;

— виконання коректорських і редакторських робіт;

— одержання оригінал-макету готового словника ;

— ряд сервісних інформаційно-пошукових функцій.

Зауважимо, що навіть на одному словнику можна згенерувати декілька неізоморфних структур, які відображають різні інформаційно-лексикографічні аспекти цього словника та його можливі функціональні проекції.

В другій главі розвинено алгоритмічний апарат аналізу української лексики з метою автоматизованої побудови повної парадигми для повнозначних відмінюваних частин мови та лематизації (редукції довільної словоформи до вихідної (канонічної) форми).

Об'єктом і базою для аналізу виступає комп'ютерний орфографічний словник української мови з точки зору української граматики, представленої комп'ютерними україномовними текстами та правилами українського правопису. Словник подано в його традиційному вигляді, де $\Lambda(U) = S0(U)$, $\Lambda(x) = x$, а $P(x)$ представляє сукупність флексій або квазіфлексій, відображаючи фонетичні та морфонологічні процеси, що відбуваються у слові при словозміні. Завдання полягає у побудові граматичних класифікацій та присловникових граматичних характеристик у залежності від кінцевих літеросполук лівої та флексій/квазіфлексій правої частини словника, побудові та програмній реалізації операторів λ та ρ таких, що :

$$\rho\, x = [x],\ \forall\, x \in S0(U);\ \lambda\, \xi = x,\ \text{де}\ \xi \in [x] \qquad (9)$$

Встановлено множину змінних, що параметризують елементи парадигми. Такими змінними виступають граматичні категорії, набір та області визначення яких залежать від частини мови. Для іменників, цими змінними є "ВІДМІНОК", "ЧИСЛО"; для прикметників, займенників, порядкових числівників - "РІД", "ВІДМІНОК", "ЧИСЛО"; для кількісних числівників - "ВІДМІНОК"; для дієслів - "ЧАС", "СПОСІБ", "ВИД" та "ОСОБА".

Множина наборів значень вказаних граматичних змінних утворює своєрідну гратку G, вузли якої $g \in G$, заповнюються елементами парадигми лексеми, що належить до певної відмінюваної частини мови. Гратка G породжує класифікацію на S(U), тобто відображення :

$$f : S(U) \to G, \qquad (10)$$

таке, що $\forall\, \xi \in S(U)\ \exists\ g \in G$, що $f(\xi) = g$.

Оскільки різні частини мови відмінюються за різними правилами, і, навіть, окремі одиниці однієї і тієї ж частини мови мають різні варіанти змін, то для розв'язання основної задачі потрібно побудувати систему алгоритмізованих правил, які відповідають різним частинам мови і окремим виняткам.

Представимо базу даних комп'ютерного орфографічного словника як сукупність записів ORF, кожен з яких являє собою окрему словникову статтю орфографічного словника : $V(x) \Rightarrow ORF$. Кожен запис складається з двох полів: WORD, яке відповідає лівій частині словника : $\Lambda(x) = x \Rightarrow WORD$, та ZAK, яке відповідає його правій частині : $P(x) \Rightarrow ZAK$. Формат запису має такий вигляд :

$$ORF = \{\, WORD,\ ZAK \,\} \qquad (11)$$

База даних з описаною структурою сформована автоматично, для чого розроблена спеціальна керуюча програма, яка забезпечила транспорт вихідного текстового файлу до бази даних. Такий підхід до обробки непрепарованого комп'ютерного словника зумовив створення програмного трикомпонентного комплексу "об'єкт—керуюча програма —база даних", який став багатофункціональним середовищем для словникових структур.

Сформована описаним способом система являє собою основу для подальшого лінгвістичного аналізу. Зазначений аналіз полягає у встановленні процедур структуризації та породженні лінгвістичних класифікацій на множині D (WORD) або, що те саме, на $\Lambda(U) = S0(U)$, в залежності від P(x) — розподілі реєстрових одиниць на групи однорідних мовних фактів як частин мови.

У другому параграфі другої глави детально представлено алгоритми побудови парадигматичних класів для відмінюваних частин мови та створено алгоритмічний апарат фонетико-морфологічного аналізу. В третьому параграфі наводяться результати цього аналізу у вигляді таблиць парадигматичної класифікації відмінюваних частин мови.

Зазначена класифікація, яка відображає морфонологічні процеси сучасної української літературної мови, апробована на широкому мовному матеріалі, представленому в Орфографічному словнику української мови.

Алгоритми й програмне забезпечення парадигматичної класифікації розроблені для кожної відмінюваної частини мови; всередині іменників та прикметників подається класифікація за числами й родами.

У четвертому параграфі реалізовано алгоритми граматичної параметризації класів відмінюваних частин мови та побудовано комп'ютерні правила, які відображають формальні властивості словозміни.

Створено спеціальну програмну процедуру морфологічного рівня, для перевірки сполучуваності основи та флексії з урахуванням фонетичних змін, що

відбуваються при відмінюванні. Ця ланка досліджень розпадається на два етапи: відображення коренево-суфіксальної сполучуваності із закінченнями та префіксально-кореневих змін при словотворі.

Автоматизований аналіз здійснювався у межах кожного парадигматичного класу слів і дозволив виділити та алгоритмізувати граматичні трансформації кожної реєстрової одиниці, забезпечити розпізнавання фонетико-морфологічних перетворень у залежності від якості основи слова, програмно сформулювати лінгвістичні правила формалізованого представлення граматичної будови слова.

Процедура реконструкції вихідної форми з будь-якої текстової словоформи викликає побудовані та програмно реалізовані правила лематизації, а потім вихідна форма за алгоритмами парадигматизації відтворює процес побудови усіх словозмінних форм, включаючи і ту словоформу, яка зумовила весь цей алгоритм. Крім того, в складних (але припустимих) варіантах взаємодії в процесі обробки даних можлива запланована зміна структури її правил поведінки.

В третій главі розглядаються прикладні аспекти комп'ютерної лексикографічної системи української мови, використання правил та класифікацій в автоматизованій обробці україномовних текстів. Описуються принципи побудови та функції автоматизованої словникової системи для роботи над комп'ютерним словником з автоматичним розподілом реєстрових одиниць за граматичними характеристиками певного парадигматичного класу, яка включає такі режими:

1) перегляд, корегування існуючих і введення нових словникових статей;

2) фіксування наголосів та їх контроль безпосередньо на екрані;

3) автоматичну побудову всіх словоформ для обраної реєстрової одиниці (блок парадигматизації);

4) реконструкцію реєстрової одиниці, виходячи з будь-якого слова з україномовного тексту (блок лематизації);

5) сортування за алфавітом;

б) індексацію для швидкого пошуку за заданими критеріями, наприклад:

— пошук слова за кількома початковими літерами.

— сортування словника.

Крім того, розроблено програму транспорту з словникової бази даних до видавничої системи, за допомогою якої здійснюється комп'ютерна верстка словника, виконуються локальні корекції безпосередньо в поліграфічному корпусі без повторної його генерації та готується оригінал-макет видання. В результаті автоматизованої обробки орфографічної словникової бази даних за допомогою розробленого варіанту лексикографічного процесора та допоміжних для поліграфічної системи програм у 1994 році було здійснене видання "Орфографічного словника української мови".

Автоматизована словникова система є складовою частиною лексикографічного процесора з лінгвістичним моделюючим та програмним забезпеченням, яке за допомогою наведеного вище автоматичного морфологічного словника виконує інструментальні функції та включає три підсистеми : укладання словників, робота з україномовними текстами, робота з словниками.

Підсистема укладання словників призначена для створення комп'ютерної структури будь-якого словника, модифікації структури існуючого словника, введення, знищення та корегування словникових статей створених словників.

Підсистема роботи з україномовними текстами включає свою підсистему створення та редагування українських текстів, перегляд тексту на екрані комп'ютера, злиття кількох обраних користувачем текстів в один файл, розбиття тексту на файли, друкування тексту на принтері, знищення текстових файлів, видавання статистичних даних, процедуру обробки вибраного тексту, в основі якої лежать функції лексикографічного процесора.

До підсистеми входить коректор, який може автоматично виправляти помилки у тексті, і є зручним та достатньо швидко реагуючим засобом при обробці текстів для укладання словників.

Підсистема роботи із словниками передбачає кілька режимів, серед яких, зокрема, є інструментарій для роботи над словниками призначений для перегляду, знищення, друкування та з'єднання/розбиття словникових баз даних. Остання функція дозволяє конструювати нові словникові статті з різних словникових баз.

Ця підсистема включає також функції побудови інверсійних текстових словників та/або словників у вигляді баз даних для обраного з цією метою словника; створення індексних файлів для забезпечення швидкого пошуку за заданим користувачем критерієм; здійснення групового пошуку словникових статей; утворення різних "проекцій" словника, підсловників, відбираючи в них лексику за заданими критеріями; побудови транскрипцій словникових статей за алгоритмізованими правилами української вимови.

Передбачено й режими роботи з лівою та правою частинами словника і з наголосами; транспорт словників із бази даних до текстового файлу з подальшим автоматичним виділенням шрифтів та надрядкових знаків.

Розроблена словникова система створює передумови для комплексної автоматизації лексикографічної діяльності, починаючи з етапа одержання словникової картотеки та проектування структури словника і закінчуючи етапами автоматичного набору, верстки та тиражування.

Традиційне різноманіття типів словників перетворюється в автоматизованій системі на різноманіття режимів їх експлуатації. Автоматизований словник як основа лексикографічного процесора стає генеральним реєстром слів української мови, споряджених відповідною інформацією, що забезпечує знаходження необхідних даних за лексичними, граматичними, стилістичними, орфографічними та орфоепічними, словотворчими й іншими особливостями кожного окремого слова.

ОСНОВНІ РЕЗУЛЬТАТИ

Досліджено формальну структуру словникових систем, алгоритми їх автоматичного аналізу та побудови на цій основі лексикографічних баз даних.

Розвинено алгоритмічний апарат і програмне забезпечення граматичної словникової системи на базі внутрішньої структури орфографічного та орфоепічного словників української мови,.

На основі запропонованої структурної моделі словникових систем розроблено формалізований метод побудови граматичних баз даних, виходячи з текстової форми орфографічного словника.

Розроблено автоматизовану систему дослідження явищ відмінювання в українській мові.

Одержано парадигматичну класифікацію української мови на основі формального аналізу слова як конструкції, складеної з літер та літеросполук. Зазначену класифікацію апробовано на широкому мовному матеріалі.

Створено програмне забезпечення автоматизованої побудови парадигм для повнозначних відмінюваних частин мови.

Досліджено алгоритми та створено програмне забезпечення для реконструкції канонічної словоформи з довільного текстового слова.

На базі одержаних алгоритмів побудовано одну з ланок лексикографічного процесора - граматичний автоматизований словник, який є інструментом для роботи з україномовними текстами та укладання словників різних типів. За допомогою та при використанні цієї словникової системи здійснене нове академічне видання Орфографічного словника української мови обсягом близько 120 тисяч слів. — К.: Довіра, 1994. — 864 с.

ПУБЛІКАЦІЇ З ТЕМИ ДИСЕРТАЦІЇ

1. Русанівський В.М., Пещак М.М., Широков В.А., Костишин О.М., Буркат Є.В., Ковтуненко Л.С. Структура та конфігурація апаратно-програмного комплексу національного машинного фонду української мови та літератури (Республиканская научная конференция "Проблемы создания и внедрения информационных технологий в Академии наук УССР: тезисы докл./ Киев, Институт математики, 1990. С. — 4-5).

2. Широков В.А., Костишин О.М., Буркат Є.В., Ковтуненко Л.С. Об одной модели интегрированной информационной системы организационного управления (Республиканская научная конференция "Проблемы создания и внедрения информационных технологий в Академии наук УССР: тезисы докл./ Киев, Институт математики, 1990. — С. 9-11).

3. Широков В.А., Костишин О.М., Буркат Є.В., Ковтуненко Л.С. Об определении информационной емкости объектов, наделенных сложной семантической структурой (Республиканская научная конференция "Проблемы создания и внедрения информационных технологий в Академии наук УССР: тезисы докл./ Киев, Институт математики, 1990. — С. 51-53).

4. Ковтуненко Л.С. Автоматизированный грамматический словарь украинского языка (Міжнародна наукова конференція "Проблемы створення машинних фондів мов: тези допов./ Київ, Український мовно-інформаційний фонд АН України, 1991. — С. 33-34).

5. Ковтуненко Л.С., Ярун Г.М. Граматична ідентифікація у термінологічних базах даних (І міжнародна наукова конференція "Українська науково-технічна термінологія" : тези допов. /Львів, Львівський політехнічний інститут, 1992. — С. 210-211).

6. Ковтуненко Л.С., Ярун Г.М. Автоматизована словникова система (II міжнародна наукова конференція "Українська науково-технічна термінологія" : тези допов./ Львів, Львівський політехнічний інститут, 1993. — С. 115-116).

7. Ковтуненко Л.С., Ярун Г.М. Комп'ютерний аналіз та синтез орфографічного словника української мови (III міжнародна наукова конференція "Українська науково-технічна термінологія": тези допов./ Львів, Державний університет "Львівська політехніка", 1994. — С. 246-247).

АННОТАЦИЯ

Ковтуненко Л.С. Компьютерно-информационные аспекты современной украинской лексикографии.

Диссертация на соискание ученой степени кандидата технических наук по специальности 05.25.05 "Информационные системы и процессы". Центральная научная библиотека им. В.И. Вернадского Национальной академии наук Украины, Киев, 1995.

Предложена структурная модель словарных систем и формальный метод построения словарных баз данных. Разработано программное обеспечение для парадигматической классификации украинского языка и реконструкции канонических словоформ. Создана автоматизированная грамматическая система, с помощью которой подготовлено издание Орфографического словаря украинского языка.

Ключові слова: комп'ютерна лексикографія, словникові структури, парадигматизація, лематизація, програмне забезпечення.

L.S. Kovtunenko. Computer- & informational aspects of modern ukrainian lexicography.

"Informational systems and processes" of 05.25.05 speciality thesis submitted for an academic degree of technical sciences. V.I. Vernadsky Central Scientific Library of the National Academy of Sciences of Ukraine, Kiev, 1995.

The structural model of the dictionary's systems and the formal method of dictionary's data bases is proposed. The software for the paradigmatic classification of ukrainian language as well as reconstruction of canonic wordforms were elaborated. The automized grammar system which was the main tool for preparing the new edition of the Ukranian orphographic dictionary was created.

Keywords: computery lexicography, dictionary's structures, paradigma, lematization, software of the lexicographic systems.