Олена Сірук
Київський національний університет імені Тараса Шевченка

ЗАСТОСУВАННЯ КОМП'ЮТЕРНОГО ІНСТРУМЕНТАРІЮ ДЛЯ ОПРАЦЮВАННЯ ДІАЛЕКТНИХ ДАНИХ

Стаття присвячена огляду зарубіжного досвіду застосування комп'ютерного інструментарію для оброблення діалектних даних та вітчизняним здобуткам у цій царині. Спеціальну увагу приділено "Корпусу українських діалектних текстів" (КорУДіТ) та "Електронній лексичній картотеці українських діалектів" (ЕЛКУД) у контексті їх застосування для дослідження діалектних текстів.

Ключові слова: *діалектний текст, корпус, українська мова, говірка, комп'ютерна лексикографічна система, комп'ютерна діалектологія*.

Закордонний та вітчизняний доробок із комп'ютерної діалектології зацікавив нас, розробників "Корпусу українських діалектних текстів" (КорУДіТ) [Сірук 2011], як теоретично можлива до використання на українському (кириличному) матеріалі скарбниця ідей та досвіду розв'язання лінгвістичних проблем, які виникають перед дослідником-практиком.

Історія дослідження діалектних матеріалів за допомогою точних методів із застосуванням комп'ютерного інструментарію нараховує більш ніж півстоліття. Одні з перших відомих спроб були зроблені американськими дослідниками на матеріалі американських діалектів англійської мови в 50-х роках XX ст. і зафіксовані в матеріалах міжнародної конференції з комп'ютерної лінгвістики COLING 1969. Середина XX ст. стала поворотним моментом у класифікуванні та публікації даних про розмовний різновид американського варіанта англійської мови. Якщо підготовка матеріалів дослідження лексики та вимови атлантичних спільнот від 1949 р. до їхньої публікації Г. Курасом у 1961 р. робилися вручну, то укладання картотеки, сортування і підготовка до друку "Регіонального словника Техасу" Е. Етвуда (1962) та дослідження Ґ. Вудом лексики та дистрибуції слів внутрішніх

районів Півдня вже виконувалися із залученням ЕОМ, що допомогло збільшити точність обчислення частоти поширення діалектних явищ [Wood 1969 : 1–12]. Друга доповідь з конференції COLING 1969, яку можна зарахувати до сфери комп'ютерної діалектології, була присвячена комп'ютерному представленню варіативності початкових фрикативів у південних британських діалектах [Francis, Svartvik, Rubin 1969]. Автори описали свій **перший досвід створення діалектних карт** за допомогою комп'ютера і відзначили його суттєву допомогу в забезпеченні точності та зручності роботи дослідника. Розвиток комп'ютерного картографування передбачав розроблення програм автоматичного пошуку даних. Така програма була створена для "Лінгвістичного атласу Сполучених Штатів та Канади" [Shuy 1966].

Поворотними в історії розвитку комп'ютерної діалектології й основоположними для її перетворення власне на діалектометрію можна вважати праці "Співвідношення між відстанню просторовою та відстанню лексичною" [Séguy 1971 : 335–357], "Діалектометрія: принципи і методи використання цифрової таксономії в галузі лінгвістичної географії" [Goebl 1982] та "Діалектометричні студії на основі італороманського, ретороманського й галлороманського мовного матеріалу з AIS і ALF" [Goebl 1984]. Автори досліджень послідовно обстоюють застосування статистичних методів у різних аспектах діалектології, тож Сеґі та Ґьобль небезпідставно вважаються фундаторами **діалектометрії** [Heeringa, Nerbonne, Kleiweg 2002 : 445–452]. Осмислення діалектології з погляду застосування до її об'єкта точних методів обчислення та новітніх технологій незмінно актуальне [Kretzschmar 1996 : 13–39; Nerbonne, Kretzschmar 2003 : 245–255; Nerbonne, Kretzschmar 2006 : 387–397]. Зараз активно формуються комп'ютерні джерела інформації (комп'ютерні діалектні атласи, корпуси діалектних та історичних текстів, питальників) і послідовно застосовуються до діалектного матеріалу точні методи дослідження (комп'ютерне картографування, статистичне опрацювання зібраних до атласів чи репрезентованих на картах даних тощо). Математичне моделювання ареальної

варіативності діалектних рис у поєднанні з переглядом традиційного уявлення про діалект, на думку В. Кретчмара [Kretzschmar 1996 : 13–39], має всі шанси оновити наші знання про мову та її регіональні відмінності. Комп'ютерним методам і прогресу в діалектометрії присвячено спеціальні випуски журналів "Computers and the Humanities" (том 37, випуск 3, 2003) та "Literary and Linguistic Computing" (том 21, випуск 4, 2006). Проблема методів у діалектології, інноваційні підходи та сучасні теоретичні й методологічні тенденції в цій царині активно обговорювалися на XII Міжнародній конференції "Methods XII" (http://www.upei.ca/methodsxii/html/e_speakers_papers.html) та семінарах Д. Хіпа, Дж. Нербонна, В. Кретчмара. Сучасний стан розвитку комп'ютерної діалектології відображений також проблематикою симпозіуму "Минулі та сучасні процеси діалектної конвергенції. Дані та методи" (Амстердам, 2010) і нерозривно пов'язаний з іменами таких дослідників, як В. Герінґа, Ф. Гінскенс, Б. Джозеф, Ф. Манні та ін.

Дослідження з комп'ютерної діалектології знаходимо в активі групи комп'ютерної лінгвістики Університету Хайфи. До завдань проекту з комп'ютерного опрацювання діалектів входили автоматичне оброблення арабських діалектних текстів Палестини та питальників, підготовка лексиконів і глосаріїв, автоматизоване визначення діалектних меж та укладання діалектних атласів. Проект базувався на систематичних польових дослідженнях арабських діалектів Палестини, якими говорять усі громади, що мешкають у північних і центральних районах Ізраїлю та сусідніх з ними. На жаль, проект не завершено через смерть його натхненника і координатора Рафі Талмона [Talmon, Wintner 2001].

У 90-і роки XX ст. одну з найперших спроб застосування комп'ютерних методів для опрацювання **слов'янських** діалектних даних здійснили словацькі дослідники. Так, П. Жиґо звернув увагу на переваги комп'ютерних технологій і поділився досвідом опрацювання зібраного для "Загальнослов'янського лінгвістичного атласу" на території Словацької Республіки діалектного матеріалу [Жиго 1996 : 57–63]. Приблизно в цей же час ак-

тивізували розпочате ще в довоєнні часи [Doroszewski 1962 : 380–392] й епізодично здійснюване в 1970-і рр. застосування статистичних методів до діалектологічних досліджень і польські вчені [Kąś 1994 : 117–123; Zarębina 1997 : 35–50]. Змінюється підхід до діалектології в цілому: діалектні одиниці залежно від часу їх фіксації починають розглядати або в контексті історії мови, або в рамках соціолінгвістики [Językoznawstwo 2003 : 66–67]. На сьогодні в Польщі чи не вперше на слов'янських теренах практично завершено онлайнову довідкову систему "Gwary polskie. Przewodnik multimedialny" (за редакцією Г. Карась), присвячену діалектології та діалектам під кутом висвітлення не лише лінгвістичним, а й загалом культурологічно-етнографічним (http://www.gwarypolskie.uw.edu.pl).

Застосування математичних методів і комп'ютерного інструментарію до діалектного матеріалу болгарської мови було успішно здійснено болгарсько-голландською групою вчених. Дослідники виконали обчислення близькості/віддаленості болгарських говірок щодо одна одної та щодо літературного ідіому за допомогою методу відстані Левенштейна [Osenova, Heeringa, Nerbonne 2010 : 425–458; Prokić et al. 2009 : 269–298]. Ця апробована на матеріалі романських і германських мов (зокрема ірландської [Kessler 1995], нідерландської [Nerbonne et al. 1996; Heeringa 2004], сардської [Bolognesi, Heeringa 2002 : 45–84], норвезької [Gooskens, Heeringa 2004 : 189–207], німецької [Nerbonne, Siedle 2005 : 129–147], американської англійської [Nerbonne 2005], каталанської [Valls et al. 2010]) методика була вперше використана для аналізу діалектної фонетики однієї зі слов'янських мов. Залучивши ще два методи частотного аналізу (метод частоти фонів і метод частоти диференційних ознак), автори дослідили міжмовні контакти болгарських діалектів та мов країн, які межують із Болгарією.

У напрямку створення власне **діалектних корпусів** найбільшу кількість ресурсів створено для германських мов, зокрема англійської та німецької. У таких корпусах, як Newcastle Electronic Corpus of Tyneside English (NECTE), Helsinki Corpus of British

English Dialects (HC; 1,5 млн. слів), Freiburg Corpus of English Dialects (FRED; 2,5 млн. слів і 300 годин записів), застосовується орфографічний запис із позначенням деяких фонетичних рис (HC, FRED) або паралельний фонетичний запис (NECTE). У двохмільйонному скандинавському діалектному корпусі (Nordic Dialect Corpus, NDC), який містить діалектні тексти північногерманських мов, частина текстів існує одночасно в орфографічному і фонетичному записах, причому трансформація відбувається напівавтоматично за допомогою спеціально складеної програми [Johannessen et al. 2009 : 73–80]. Додано звукові файли та відеофайли; система пошуку дає можливість будувати велику кількість різних запитів. Корпусна база даних Nordic Syntax Database містить синтаксичний модуль із можливістю сортування одиниць за місцем, віком, статтю інформантів і власне синтаксичними критеріями, а також уможливлює генерування результатів у вигляді лінгвістичних карт з ізоглосами [Lindstad et al. 2009 : 283–286]. Корпуси діалектних текстів створені також у Німеччині в рамках Архіву німецької мови (Deutsches Spracharchiv, DSAv) та Бази даних розмовної німецької мови (Datenbank Gesprochenes Deutsch, DGD). DSAv складається з 28 окремих корпусів, які містять записи та задокументовані стенограми вітчизняних і зарубіжних німецьких діалектів (зокрема корпуси німецьких говірок Бразилії, Ізраїлю, Росії, Румунії, Північної Америки загалом, а також корпус слов'янських говірок Рурської області), стандартної німецької та записи розмовного мовлення в різних соціальних і ситуаційних контекстах. Ці корпуси оснащені потужною пошуковою системою й викладені на сайті Інституту німецької мови (м. Мангайм) (http://dsav-oeff.ids-mannheim.de/DSAv/). Цікавою для діалектолога є також корпусна пошуково-аналітична система Cosmas II (http://www.ids-mannheim.de/cosmas2/projekt/). Масштабні проекти і дослідження діалектів німецької мови стали можливими завдяки не тільки державній підтримці, а й фінансовій ініціативі концерну "Фольксваґен".

Потужну багаторічну державну підтримку має корпусотворення в Китаї. Корпуси діалектних текстів створено за рахунок

національних фондів "Дослідження високих технологій 863", "Програми розвитку 973" та Національної наукової фундації Китаю. Корпуси містять тексти, зображення, аудіо- та відеозаписи. Корпус регіонального акцентного мовлення (Regional Accent Speech Corpus, RASC863) представляє мовлення десяти основних діалектних різновидів китайської мови, один з яких, мандаринський, налічує 800 мільйонів мовців і виконує функції загальнонаціональної мови. Більшість китайців є "білінгвами"; вони можуть спілкуватися своєю рідною говіркою та мандаринською, модифікованою впливом рідної фонологічно, лексично й синтаксично [Li et al.]. Корпус мандаринського діалекту китайської мови є аудіокорпусом і містить записи міського мовлення. Корпус розпізнаної китайської розмовної мови складається з восьми підкорпусів, серед яких – два діалектних. Корпус південного діалектного мовлення нараховує 70 годин запису (100 тис. речень, або 8.2 GB), Корпус північного діалектного мовлення – 59 годин запису (75 тис. речень, або 5.8 GB) [Qian et al.].

Праця над корпусним представленням діалектних текстів слов'янських мов сьогодні найактивніше розвивається в Чехії, Словаччині, Словенії, Польщі, Росії, Болгарії [Сірук 2011]. Певний доробок у цій ділянці напрацьовано й українськими лінгвістами. Українська діалектологія має потужний досвід дослідження говірок за допомогою питальників, міцну традицію укладання діалектних атласів, словників, опрацювання теоретичних діалектологічних проблем. В останні роки активізувався розвиток діалектної текстології. Якщо раніше тексти рідко потрапляли в центр уваги дослідників (за винятком таких праць, як "Говори української мови: збірник текстів" [Говори 1977]) і здебільшого спорадично долучалися до монографій та підручників, то зараз з'являються дослідження, зроблені на основі діалектних текстів (наприклад, "Лінгвокогнітивні та прагматичні виміри діалектних текстів буковинських говірок" [Руснак 2009]), і текстозбірки ("Говірки Чорнобильської зони: тексти" [Говірки 1996], "Українські говори Румунії: діялектні тексти" [Павлюк, Робчук 2003]; живе мовлення носіїв говірки у фонетичній транскрипції, орфо-

графічному записі та диск з аудіофайлами пропонуються у виданні "Говірка села Машеве Чорнобильського району", присвяченому опису мови і традиційної культури одного села [Говірка 2003]; заслуговує на увагу збірник текстів "Говірки південно-західного наріччя української мови" [Говірки 2005]). Але такі праці поодинокі, не охоплюють усього українського діалектного континууму, а тексти, які є джерельною базою досліджень, здебільшого не доступні пересічному дослідникові не тільки в комп'ютерній формі, а й у паперовій через обмежений тираж.

Необхідно зазначити, що як специфіка діалектного матеріалу, так і психологічне тяжіння до традиції від початку створювали діалектологам додаткові труднощі в залученні електронного інструментарію для мовних досліджень. Сумніви стосовно доцільності його застосування посилювалися, з одного боку, малодоступністю та недосконалістю перших ЕОМ і програмного забезпечення, з другого – незнанням лінгвістами основ програмування й алгоритмізації. Тож кількість досліджень, які можна зарахувати до комп'ютерної діалектології, апріорі не була значною порівняно з іншими галузями мовознавства. У багатьох розвинених країнах вже стало традицією застосування комп'ютера в повсякденному житті, а комп'ютерних методик і програм – для лінгвістики, натомість в Україні ситуація суттєво не змінилася. Сьогодні українська комп'ютерна діалектологія перебуває на етапі формування; корпусів української діалектної мови наразі немає, як немає і публікацій з цієї проблематики. Тож за найближчу **мету** ми поставили собі розробити методику укладання "Корпусу українських діалектних текстів" і створити відповідно до цієї методики його сегмент [Сірук 2011]. Зведення сукупності виявлених за різними джерелами українських діалектних текстів, їх опрацювання як елементів єдиної лінгвістичної інформаційної системи, забезпечення оперативного доступу користувачів до цього джерела мовних даних – ось які **основні кроки** потрібно здійснити для розроблення комп'ютерної лінгвістичної системи, призначеної для багаторівневого аналізу діалектної мови, зокрема фонетичних, морфологічних, синтаксич-

них, семантичних, стильових рис на різних етапах її функціонування. Інакше кажучи, потрібно за допомогою відповідного програмного забезпечення зробити доступними для корпусного опрацювання зафіксовані в паперовому варіанті чи аудіовигляді тексти шляхом їх переведення в комп'ютерну форму; забезпечити ці тексти відповідними зовнішніми (автор, інформація про видання або приватну текстотеку, про записувачів та інформаторів, анкета текстів тощо) і внутрішніми **маркерами**, які ще називають структурними (номер, початок і кінець тексту, розділу, абзацу, речення, слова тощо), а також комплексом власне **лінгвістичних розміток** (морфологічних, синтаксичних, семантичних тощо). І теоретичні засади, і практична реалізація нашого проекту є **новаторськими** для української діалектології та корпусної лінгвістики.

Корпусна методика дослідження діалектних текстів є взаємодоповняльним поєднанням методів та інструментарію корпусної лінгвістики й текстової діалектології. Таке поєднання забезпечує комплексність діалектологічного дослідження, його масштабність, частотну перевірюваність, обґрунтованість висновків, їхню прозорість, а також швидкість отримання результатів. КорУДіТ покликаний уможливити порівняння діалектної мови з літературним ідіомом, говорів і говірок між собою, а також текстів однієї говірки аж до вивчення ідіолектів окремих її носіїв. Методика опрацювання діалектного тексту, яка використовується в КорУДіТ, дає можливість формувати й опрацьовувати паралельно три взаємопов'язані підкорпуси (затранскрибованих діалектних текстів, діалектних текстів в орфографічному записі та корпус "перекладених" літературною мовою діалектних текстів). За допомогою якісно-кількісного аналізу розмічених текстів можна зробити статистично обґрунтовані висновки щодо функціонування певного явища у певному середовищі часово-просторового континууму.

КорУДіТ розробляється в рамках створення "Корпусу текстів української мови" (КТУМ), який на сьогодні складається з анотованих підкорпусів загальним обсягом понад 6 млн. слововжи-

вань. Робота зі створення КТУМ проводиться співробітниками лабораторії комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка під керівництвом доц. Н. П. Дарчук у наукових студях з формалізації мовних досліджень. При створенні "Корпусу текстів української мови" враховувався багаторічний досвід лексикографічної та корпусної роботи фахівців лабораторії. Окрім діалектного, який виділено в окремий проект через особливу складність його структури, КТУМ містить підкорпуси:

1) української поезії;
2) української художньої прози (кінець XX ст.);
3) фольклорних текстів;
4) публіцистичних текстів за 2008–2011 рр.;
5) наукових текстів з лінгвістики, економіки, юриспруденції тощо;
6) паралельних українсько-російських текстів.

Зараз КорУДіТ працює в тестовому режимі. Здійснюється внесення діалектних текстів до бази даних корпусу, їх загальна розмітка, автоматизований морфологічний аналіз із подальшим зняттям омонімії в ручному режимі. Відпрацювання системного оброблення діалектних текстів у корпусі проводиться на матеріалах авторської діалектної текстотеки обсягом близько 100 тис. слововживань, які представляють регіон західноволинських говірок. Одночасно триває робота над удосконаленням системи внесення текстів, а також над розробленням і тестуванням систем автоматичної морфологічної розмітки й автоматизованого зняття омонімії. Продовжується пошук оптимального шляху оброблення текстових діалектних масивів, зокрема в напрямку створення програмного пакету для автономного робочого місця діалектолога, з можливістю виконувати певні завдання оброблення текстів не безпосередньо на корпусному сервері, а незалежно від наявності зв'язку з корпусом через мережу Інтернет, із подальшою синхронізацією опрацьованого в автономному режимі матеріалу з усім корпусом. Невід'ємним процесом є створення інтернет-сторінки з інструкцією для користувача та розробника і базовим пошуком у діалектному корпусі в мережі

Інтернет на лінгвістичному порталі Mova.info, а також удосконалення інтерфейсу програмного пакету адміністратора.

Ще одним проектом із комп'ютерної діалектології, який можна використовувати як текстову дослідну базу, є "Електронна лексична картотека українських діалектів" (ЕЛКУД), розроблена співробітниками лабораторії комп'ютерної лінгвістики та відділом діалектології Інституту української мови НАН України. Метою формування ЕЛКУД є зведення всієї сукупності виявлених за різними джерелами лексичних одиниць українських діалектів, їх опрацювання як елементів інформаційної системи та забезпечення оперативного доступу користувачів до цього джерела мовних даних. Лінгвістична база даних ЕЛКУД створювалася на основі друкованих діалектних словників і неопублікованих рукописних лексичних картотек інших фондів (українські матеріали до лексичної частини "Загальнослов'янського лінгвістичного атласу", до "Атласу української мови" та "Лексичного атласу української мови") за принципом формалізації мовного матеріалу, тобто представлення інформації про діалектні одиниці у структурованому, придатному для комп'ютерного опрацювання вигляді [Сірук 2010 : 134–141; Grytsenko, Siruk, Sorokin 2009 : 132–144].

Отже, перспективність і доцільність застосування точних методів для аналізу діалектних матеріалів і використання комп'ютерних технологій у діалектології є безсумнівними. Світовий досвід може прислужитися українським діалектологам для розроблення нових лінгвістичних методик і програмних продуктів на українському ґрунті, для отримання якісно нових результатів, активувати вітчизняні мовознавчі дослідження в цілому, забезпечивши тим самим поштовх до нових звершень і для зарубіжних колег, зокрема для творців текстових корпусів слов'янських мов.

1. Говірка села Машеве Чорнобильського району: У 4-х ч. – К., 2003. 2. Говірки південно-західного наріччя української мови / Упоряд. Н. М. Глібчук. – Львів, 2005. 3. Говірки Чорнобильської зони: тексти / Упоряд.

П. Ю. Гриценко, Н. П. Прилипко, О. А. Малахівська та ін. – К., 1996. 4. Говори української мови: збірник текстів / Відп. ред. Т. В. Назарова. – К., 1977. 5. *Жиго П.* Опыт использования ЭВМ в обработке материалов ОЛА // Общеславянский лингвистический атлас. Материалы и исследования. 1991–1993. Сборник научных трудов**. –** Вып. 21. – М., 1996. 6. *Павлюк М., Робчук І.* Українські говори Румунії: діялектні тексти. – Канада, 2003. 7. *Руснак Н. О.* Лінгвокогнітивні та прагматичні виміри діалектних текстів буковинських говірок. – Чернівці, 2009. 8. *Сіруκ О. Б.* Моделювання комп'ютерних систем для української мови: "Електронна лексична картотека українських діалектів" (ЕЛКУД) // UCRAINICA IV. Současná Ukrajinistika. Problémy jazyka, lietratury a kultury. Sborník vědeckých článků. Z mezinárodní konference V. Olomoucké sympozium ukrajinistů střední a východní Evropy**. –** Olomouc, 2010. 9. *Сіруκ О. Б.* "Корпус українських діалектних текстів" (КорУДіТ) // Мовні і концептуальні картини світу. – Вип. 35. – К., 2011. 10. *Сіруκ О. Б.* Слов'янська корпусна діалектологія: історія та перспективи // Мовні і концептуальні картини світу. – Вип. 36. – К., 2012. 11. *Bolognesi R., Heeringa W.* De invloed van dominante talenop het lexicon en de fonologie van Sardische dialecten // Gramma/TTT: tijdschrift voor taalwetenschap. – Vol. 9 (1). – 2002. 12. *Doroszewski W.* O statystyczne przedstawienie izoglos (pierwodruk 1935), przedruk // Studia i szkice językoznawcze. **–** Warszawa, 1962**.** 13. *Francis N. W., Svartvik J., Rubin G. M.* Computer-produced Representation of Dialectal Variation: Initial Fricatives in Southern British Dialects // International Conference in Computational Linguistics COLING 1969: Preprint № 50. 14. *Goebl H.* Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie. Österreichische Akademie der Wissenschaften. – Wien, 1982. 15. *Goebl H.* Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF. 3 Vol. – Tübingen: Max Niemeyer, 1984. 16. *Gooskens C., Heeringa W.* Perceptive Evaluation of Levensthein Dialect Distance Measurments using Norwegian Dialect Data // Language Variation and Change. – Vol. 16. – 2004. 17. *Grytsenko P., Siruk O., Sorokin V.* Electronic Lexical Card Index for the Ukrainian Dialects (ELCIUD) // NLP, Corpus Linguistics, Corpus Based Grammar Research: Fifth International Conference "SLOVKO 2009". Proceedings. L'. Štur Institute of Linguistics. – Bratislava, 2009. 18. *Heeringa W.* Measuring Dialect Pronunciation Differences using Levensthein Distance. PhD Thesis. – University of Groningen, 2004. 19. *Heeringa W., Nerbonne J., Kleiweg P.* Validating Dialect Comparison Methods // Classification, Automation, and New Media. Proceedings of the 24th Annual Conference of the Gesellschaft für Klassifikation e. V., University of Passau, March 15–17, 2000. – Springer, Berlin, Heidelberg and New York, 2002. 20. Językoznawstwo w Polsce: stan i perspektywy. – Opole, 2003. 21. *Kąś J.* Metody statystyczne w badaniach dialektologicznych // Acta Universitatis Lodziensis. Folia Linguistica**.** – Vol. 12. – 1994**.** 22. *Johannes-*

sen J. B., Priestley J., Hagen K., Åfarli T. A., Vangsnes Ø. A. The Nordic Dialect Corpus – an advanced research tool // Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009. NEALT Proceedings Series. – Vol. 4. – 2009. 23. *Kessler B.* Computational dialectology in Irish Gaelic // Proceedings of the European Association for Computational Linguistics. – Dublin, 1995. 24. *Kretzschmar W. A.* Quantitative Areal Analysis of Dialect Features // Language Variation and Change**. –** Vol. 8 (1). – 1996**.** 25. *Li Ai-jun, Yin Zhi-gang*. Standardization of Speech Corpus // http://www.codata.org/06conf/abstracts/C5/C5-Li Ai-jun.htm. 26. *Lindstad A. M., Nøklestad A., Johannessen J. B., Vangsnes Ø. A.* The Nordic Dialect Database: Mapping Microsyntactic Variation in the Scandinavian Languages // Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009. NEALT Proceedings Series. – Vol. 4. – 2009. 27. *Nerbonne J.* Variuos Variation Aggregates in the LAMSAS South // Language Variety in the South III. – University of Alabama Press, Tuscaloosa, 2005. 28. *Nerbonne J., Heeringa W., E. van den Hout, P. van der Kooi, Otten S., W. van de Vis*. Phonetic Distance between Dutch Dialects // CLIN VI, Papers from the sixth CLIN meeting. – University of Antwerpen, 1996. 29. *Nerbonne J., Kretzschmar W.* Introducing Computational Methods in Dialectometry // Computational Methods in Dialectometry. Special issue of Computers and the Humanities. **–** Vol. 37 (3). – 2003. 30. *Nerbonne J., Kretzschmar W.* Progress in Dialectometry: Toward Explanation // Literary and Linguistic Computing**. –** Vol. 21 (4). – 2006. 31. *Nerbonne J., Siedle C.* Dialektklassifikation auf der grundlage aggregierter ausspracheunterschiede // Zeitschrift fur Dialektologie und Linguistic. – Vol. 72(2). – 2005. 32. *Osenova P., Heeringa W., Nerbonne J.* A Quantitative Analysis of Bulgarian Dialect Pronunciation // Zeitschrift für Slavische Philologie. – Vol. 66(2). – 2010. 33. *Prokić J., Nerbonne J., Zhobov V., Osenova P., Simov K., Zastrow T., Hinrichs E.* The Computational Analysis of Bulgarian Dialect Pronunciation // Serdica Journal of Computing**. –** Vol. 3(3). – 2009. 34. *QIAN YueLiang, LIN ShouXun, ZHANG YongDong, LIU Yang, LIU Hong, LIU Qun*. An Introduction to Corpora Resources of 863 Program for Chinese Language Processing and Human-Machine Interaction // http://mtgroup.ict.ac.cn/~liuyang/papers/final_revised.pdf. 35. *Séguy J.* La relation entre la distance spatiale et la distance lexicale. Revue de Linguistique Romane. – Vol. 35. – 1971. 36. *Shuy R. W.* An Automatic Retrieval Program for the Linguistic Atlas of the United States and Canada // Computation in Linguistics. – Bloomington: Indiana University Press, 1966. 37. *Talmon R., Wintner Sh.* Computational processing of spoken North Israeli Arabic // Arabic Language Processing: Status and Prospects, ACL/EACL 2001 Workshop. – Toulouse, 2001. 38. *Valls E., Nerbonne J., Prokić J., Wieling M., Clua E., Lloret M.* Applying Levenshtein Distance to Catalan Dialects. A Brief Comparison of Two Dialectometric Approaches // Accepted to appear in: Verba. Anuario Galego de Filoloxía. 39. *Wood G. R.* Dialectology by Computer // International Conference

in Computational Linguistics COLING 1969: Preprint № 10. 40. *Zarębina M.* Słownictwo mieszkańców wsi Polski południowo-wschodniej (analiza statystyszna) // Z polszczyzny historysznej i współszesnej. – Rzeszów, 1997.

Статья посвящена обзору зарубежного опыта использования компьютерного инструментария для обработки диалектных данных и отечественным достижениям в этой области. Особое внимание уделено "Корпусу украинских диалектных текстов" (КорУДиТ) и "Электронной лексической картотеке украинских диалектов" (ЭЛКУД) в контексте их использования для исследования диалектных текстов.

Ключевые слова: *диалектный текст, корпус, украинский язык, говор, компьютерная лексикографическая система, компьютерная диалектология.*

The paper is devoted to the review of foreign experience of using computer tools for processing dialect materials and achievements in this sphere at home. We pay special attention to the Corpus of Ukrainian Dialect Texts (CorUDiT) and the Electronic Lexical Card Index for the Ukrainian Dialects (ELCIUD) in the context of their use as the base for dialect text research.

Key words: *dialect text, corpus, Ukrainian language, dialect, computer lexicographic systems, computational dialectology*.