УДК 811.161.2´374.73:004.656

О. Б. Сірук

ОПРАЦЮВАННЯ ДІАЛЕКТНИХ ДАНИХ У «КОРПУСІ УКРАЇНСЬКИХ ДІАЛЕКТНИХ ТЕКСТІВ»

Сірук О. Б. Опрацювання діалектних даних у «Корпусі українських діалектних текстів».

У статті йдеться про методику побудови «Корпусу українських діалектних текстів» (КорУДіТ) у рамках «Корпусу текстів української мови» (КТУМ); спеціальну увагу приділено особливостям представлення діалектних текстів у КорУДіТ. Тестування системної обробки діалектних текстів проведено на матеріалах авторської діалектної текстотеки, яка представляє регіон західноволинських говірок; у дослідженні окреслено тему перспектив розвитку КорУДіТ.

Ключові слова: діалектний текст, корпус, розмітка, українська мова.

Сирук Е. Б. Обработка диалектных данных в «Корпусе украинских диалектных текстов».

В статье идет речь о методике построения «Корпуса украинских диалектных текстов» (КорУДиТ) в рамках «Корпуса текстов украинского языка» (КТУЯ). Специальное внимание уделяется особенностям представления диалектных текстов в КорУДиТ. Тестирование системной обработки диалектных текстов проводится на материалах авторской диалектной текстотеки, которая представляет регион западноволынских говоров; затрагивается тема перспектив развития КорУДиТ.

Ключевые слова: диалектный текст, корпус, разметка, украинский язык.

Siruk O. B. Dialect Data Processing in a Corpus of Ukrainian Dialect Texts.

The paper is devoted to the methodology of building a Corpus of Ukrainian Dialect Texts (CorUDiT) in the framework of the Corpus of Texts of the Ukrainian Language (CTUL). We pay special attention to the peculiarities of the representation of dialect texts in CorUDiT. The testing of the systematic treatment of dialect texts is done on the author's dialect text collection representing the West Volynian dialect region. Our investigation also focuses on the future trends of the development of CorUDiT.

Key words: dialect text, corpus, marking, Ukrainian language.

Важливість застосування текстових корпусів та комп'ютерного інструментарію для дослідження мови вже не потрібно доводити на сучасному етапі розвитку української мовознавчої науки. Корпус текстів забезпечує дослідника репрезентативною вибіркою, одиниці якої можуть бути проаналізовані за своїми статистичними та лінгвістичними характеристиками. Текст у широкому розумінні цього терміна, текст як «писемний або усний мовленнєвий масив, що становить лінійну послідовність висловлень, об'єднаних у ближчій перспективі смисловими і формально-граматичними зв'язками, а в загальнокомпозиційному, дистантному плані – спільною тематичною і сюжетною заданістю» [1, с. 679], є найповнішою базою для фундаментальних лінгвістичних експериментів.

Але якщо тексти стильових різновидів літературного ідіому більш-менш послідовно залучаються науковцями як дослідницька база, то діалектні тексти все ще порівняно рідко стають фундаментом лінгвістичного дослідження. Проте тільки в рамках тексту можливий всебічний ґрунтовний аналіз його одиниць, його синтаксичної будови, семантичної та структурної цілісності, вираженої відповідними смисловими і формально-граматичними засобами зв'язку, а також його стильових ознак. Корпусна методика дослідження діалектних текстів –

застосування методів та інструментарію корпусної лінгвістики до царини текстової діалектології – є синтезом, взаємодоповнюючим поєднанням цих двох напрямків, яке забезпечує комплексність діалектологічного дослідження, його масштабність, частотну перевіреність і обґрунтованість висновків, їхню прозорість, а також швидкість отримання результатів. З огляду на це заслуговує на увагу закордонний та вітчизняний доробок з комп'ютерної діалектології як скарбниця ідей та досвіду розв'язання лінгвістичних проблем, можлива до застосування на українському (кириличному) матеріалі [2].

Ідею створення діалектного корпусу текстів як оптимальної бази для збереження та глибшого опрацювання текстового матеріалу ми реалізуємо у вигляді «Корпусу українських діалектних текстів» (КорУДіТ). Цей корпус є складовою «Корпусу текстів української мови» (КТУМ), колективною працею фахівців лабораторії комп'ютерної лінгвістики Київського національного університету імені Тараса Шевченка. Метою діалектного корпусного проекту є відновлення семантичної безперервності в дослідженні говірок шляхом доповнення текстовими даних, отриманих за допомогою питальників, а також забезпечення дослідника базою (структурованим мовним матеріалом та комп'ютерним інструментарієм) для багаторівневого аналізу діалектної мови, зокрема фонетичних, морфологічних, синтаксичних, семантичних, стильових рис на різних етапах її функціонування. Зокрема, це передбачає такі кроки, як зведення сукупності виявлених за різними джерелами українських діалектних текстів, подальше опрацювання цих текстів як елементів єдиної лінгвістичної інформаційної системи, а також забезпечення оперативного доступу користувачів до цього джерела мовних даних. Наразі ми працюємо над розробленням методики укладання КорУДіТ і створенням його сегмента на власноруч підготованому текстовому матеріалі західноволинських говірок обсягом близько 100 тисяч слововживань. Для тестування корпусу ми залучаємо також матеріали діалектологічної практики студентів Інституту філології.

Наше завдання полягає в тому, щоб за допомогою відповідного програмного забезпечення зробити доступними для корпусного опрацювання зафіксовані в паперовому чи аудіовигляді діалектні тексти шляхом їх переведення в комп'ютерну форму. Потрібно забезпечити ці тексти відповідними зовнішніми (автор, інформація про видання або приватну текстотеку, про записувачів та інформаторів, анкета текстів тощо) та 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 та ін.). Тексти в КорУДіТ додаються у трьох видах (відповідно до яких формуються три взаємопов'язані підкорпуси):

1) фонетична транскрипція (підкорпус транскрибованих діалектних текстів найбільше придасться для фахівців-діалектологів): *Ду ха́ти зайди́ / с'ві́тиц'і йак л'ӳстро^{у} // ти́л'ко^{у} йідно́ мн'а́со^{у} ни русте́ і сме^{и}та́на // а йак ў л'ох зайде́ш / во́чи рузбіга́йуц'і / вс'о там йе // вс'о до́бре аж^{ш} хо́че^{и}ц'і //* ;

2) літературизований (орфографічний) запис (підкорпус діалектних текстів в орфографічному записі має ширшу аудиторію, зокрема позафілологічну; передбачається можливість досліджувати стилізацію художнього тексту під говірки та розмовний стиль в рамках корпусу усної української мови): *Ду х**а**ти зайд**и** – св**і**тиці як л**ю**стро. Т**и**лько їдн**о** мн**я**со ни русте і смет**а**на. А як в льох зайд**е**ш – в**о**чи рузбіг**а**ючі – всьо там є. Всьо д**о**бре аж х**о**чеці*;

3) запис, максимально наближений до літературної мови (підкорпус «перекладених» сучасною літературною мовою діалектних текстів може застосовуватися так само, як і літературизований запис; він необхідний для автоматичної розмітки тексту, зокрема для максимально плідної роботи автоматичного морфологічного аналізу, розрахованого перш за все на тексти літературного стандарту української мови): *До хати зайди – світиться, як люстро. Тільки одне м'ясо не росте і сметана. А як в льох зайдеш – очі розбігаються – все там є. Все добре, аж хочеться*.

Для автоматизованого формування паралельних текстів передбачено спеціальну програму, яка на базі фонетичної транскрипції або орфографічного запису створює два інших тексти відповідно до розроблених для групи тестових західноволинських говірок правил базової трансформації текстів. Ці правила можна застосовувати для текстів інших говірок та поповнювати за потреби щодо конкретної говірки.

Для збереження логіки викладу та подальшого дослідження лінгвістичних змін в комунікації на стику двох різних стильових різновидів мови ми включаємо до корпусу як текст інформатора, так і текст записувача (здебільшого, це окреслення теми розмови або запитання, на які відповідає інформатор). Якщо немає слів записувача, то діалог або полілог розриваються, діалектна канва спотворюється, виникає небезпека ототожнення з нормами говірки повторених за інформатором питальних фраз і конструкцій, які в інших комунікативних ситуаціях не вживаються, послідовно замінюючись іншими, не характерними для літературного ідіому, але питомими для говірки.

Сьогодні українська комп'ютерна діалектологія перебуває на етапі формування; корпусів української діалектної мови наразі немає, як немає і публікацій з цієї теми. Тож і теоретичні засади, і практична реалізація нашого корпусного проекту є новаторськими для української діалектології та корпусної лінгвістики.

Корпус діалектних текстів української мови розрахований на фахівців та усіх, хто цікавиться діалектним шаром української лексики; його можна використовувати як довідкову систему та з навчальною метою.

За допомогою якісно-кількісного аналізу розмічених текстів можна зробити статистично обґрунтовані висновки щодо функціонування певного явища у певному середовищі часо-просторового континууму. КорУДіТ покликаний забезпечити дослідника текстовим матеріалом та програмним інструментарієм для порівняння діалектної мови з літературним ідіомом, говорів та говірок між собою, а також текстів однієї говірки аж до вивчення ідіолектів окремих її носіїв. У поєднанні з іншими лінгвістичними програмними засобами КорУДіТ може бути підґрунтям для комплексного дослідження української мови на певному етапі її розвитку. Корпус є базою для створення різного типу вибірок, на основі яких можна укласти діалектні комп'ютерні алфавітні, тезаурусні, гніздові, частотні словники та картотеки, мережеві інформаційно-пошукові системи, а також зробити проекцію корпусних даних на лінгвістичні карти.

Методика обробки діалектного тексту, яка використовується в КорУДіТ, дає змогу сформувати й опрацьовувати паралельно три взаємопов'язані підкорпуси (затранскрибованих діалектних текстів, діалектних текстів в орфографічному записі та корпус «перекладених» літературною мовою діалектних текстів). Методика створення корпусу діалектних текстів завдяки своїй гнучкості придасться для формування корпусу текстів усного мовлення та паралельних корпусів.

Розвиток проекту передбачає виконання таких базових завдань, як шліфування методики опрацювання діалектних текстів на засадах багаторівневого максимально точного їх маркування; розроблення відповідного програмного забезпечення та онлайн-інтерфейсу, «дружніх» як до діалектологів-фахівців, так і до користувачів ширшого профілю; наповнення корпусу діалектними текстами всіх регіонів України та українських говірок за кордоном.

Планується також підтримка відносно стабільного поповнення КорУДіТ говірковими текстами, зібраними під час діалектологічної практики, для забезпечення діахронічності корпусу. Додавання звукових файлів з можливістю їхньої автоматизованої трансформації в друкований текст є досить трудомістким, тож його реалізація не є нашим першочерговим завданням. А ось представлення корпусу в мережі Інтернет задля його постійного поновлення та вдосконалення шляхом зворотного зв'язку з широким колом користувачів, зокрема з діалектологами та спеціалістами з корпусної лінгвістики, є одним з пріоритетних напрямків розвитку КорУДіТ.

Література

1. Українська мова : [енциклопедія] / За ред. В. М. Русанівського, О. О. Тараненка та ін. – К., 2004.

2. Siruk O. Corpus of Ukrainian Dialect Texts (CorUDiT) as a component of a Corpus of Texts of the Ukrainian Language (CTUL) / Olena Siruk // Prace Filologiczne LX. – Warszawa, 2012.

Стаття надійшла до редакції 18.11.2011 р.