Н. Дарчук, к. філол. н., ***В. Сорокін****
Київський національний університет ім. Т. Шевченка (Київ)
УДК 161.2. 81'33

КОРПУС ТЕКСТІВ ЯК ДЖЕРЕЛО ДЛЯ МОВОЗНАВЧИХ І ЛІТЕРАТУРОЗНАВЧИХ ДОСЛІДЖЕНЬ

Стаття присвячена аналізу шляхів формування корпусів текстів різних стилів (поетичного, публіцистичного, наукового), вказується на засади створення параметризованих баз даних, характеризується програмне забезпечення таких одиниць.

Стан сучасної лінгвістики останніх десятиліть Ю. Д. Апресян [1] визначає красивою метафорою «золотий вік лексикографії» через безпрецедентний ріст числа й різноманітності словників, удосконалення методики лексикографування, принципи системного опису лексики і навіть переорієнтацію значної частини сучасної теоретичної лінгвістики на опис окремих лексем. Такий прорив у мікросвіт значення одного слова спонукає філологів до всебічного лінгвістичного «портретування», а значить інтегрального опису мови, що аж ніяк неможливо без досягнень комп'ютерної лінгвістики в цілому і корпусної зокрема.

Справді, фонетичні, просодичні, морфологічні, синтаксичні, семантичні, прагматичні, стилістичні, комунікативні, сполучуванісні властивості лексем є вивідними безпосередньо з текстів, тому стають невичерпним джерелом для словників і граматик.

Незаперечним є й той факт, що в останні десятиліття фахівціями галузей, що мають справу з комп'ютерним аналізом текстів, гостро відчувається потреба в точних функціональних характеристиках мовних одиниць у різних типах текстів. Для проведення теоретичних і прикладних (машинний переклад, аналіз і синтез мовлення, анотування і реферування тексту тощо) та дидактичних досліджень спеціалісти в галузі когнітивної лінгвістики, стилістики, семасіології, функціональної граматики, словотвору не мають узагальнених даних про закономірності функціонування мовних одиниць у мовленні. Отже, необхідність розбудови корпусної лінгвістики – ознака нашого часу, завдяки їй спеціалісти одержать усю необхідну лінгвістичну інформацію для подальшого її опрацювання у філологічних студіях.

На нашу думку, існує кілька шляхів розвитку корпусної лінгвістики, але спільним і неодмінним для них є накопичення текстів різних стилів і часових зрізів з відповідним зовнішнім (джерело, автор, рік видання тощо) і внутрішнім, пов'язаним зі структурацією тексту, його маркуванням (номер розділу, абзацу, речення, слова, початок і кінець тексту, абзацу, речення тощо).

Перший шлях полягає у програмному забезпеченні пошуку окремої словоформи у певному джерелі або всьому корпусі текстів та формуванні дослідницької ілюстративної текстової бази. Така система може функціонувати без використання лінгвістичних алгоритмів і словників або з використанням словників як допоміжного матеріалу для формування запитів, хоча сам корпус лінгвістично не розмічається.

Другий шлях пов'язаний з розробкою програмного забезпечення для самостійного маркування корпусу текстів дослідником. За допомогою спеціально розроблених програм та інтерфейсу користувач-дослідник може «портретувати» слово: визначати його морфологічну, синтаксичну, лексичну характеристику згідно з контекстом уживання. Цей шлях забезпечує можливість, з одного боку, одержання однозначної, точної у лінгвістичному смислі інформації, з іншого – створення лексикографічної картки за результатами розмітки.

Третій шлях передбачає наявність комп'ютерних інструментів – пакету програм для різнобічної лінгвістичної обробки текстового матеріалу. В цьому випадку корпус текстів лінгвістично розмічається в автоматичному режимі з подальшою автоматизованою обробкою для зняття омонімії. Завдяки цьому користувач одержує інформацію, передбачену програмним забезпеченням та базами даних (словниками з морфологічною, синтаксичною, лексичною, енциклопедичною інформацією), створює за текстами свої власні алфавітні й частотні словники, робить автоматично транслітерацію, фонематичний або фонетичний запис, сегментує текст на морфи, морфологічно індексує текст/слово, будує синтаксичні дерева залежностей речення тощо.

Кожен з цих шляхів має свої позитивні риси і недоліки. До **позитивних** рис першого можна віднести швидкість створення текстових корпусів; другого – точність характеристики одиниць аналізу та при узгодженій роботі спільними зусиллями фахівців – створення сучасної автоматизованої лексикографічної картотеки; третього – ефективність і швидкість в одержанні різноманітної інформації. До **недоліків** – ресурсну обмеженість першого, повільність роботи другого та недостатню точність третього, тому що 100% бездоганної обробки тексту автоматично досягти практично неможливо.

Займаючись проблемами корпусної лінгвістики, викладачі Інституту філології Київського національного університету ім. Т. Шевченка обрали третій шлях, розробивши відповідну концепцію, суть якої полягає у параметризації текстових корпусів за якісними і кількісними ознаками, в розробленні методичних засад створення комп'ютерних інструментів для лінгвістів, а також пакетів програм для укладання електронних картотек, частотних словників, тезаурусів, словників морфемно-словотворчих гнізд. Це розвиває проект корпусного опрацювання вшир та вглиб.

Ми не будемо спинятися на проблемах і вимогах до корпусу текстів з точки зору користувача, розуміючи, що в його побудові є своя стратегія, свої особливості: репрезентативність, повнота, структуризація матеріалу. Дійсно, як показує лексикографічна практика, правильно сформований текстовий масив – запорука якісних мовознавчих досліджень, текстова база – джерело створення словників і полігон лінгвістичних досліджень функціонування мовних одиниць усіх рівнів. Тому підбір текстового матеріалу, організація його за стилями, зонами, хронологією є обов'язковою умовою організації корпусної системи.

Мета нашого проекту полягає у комплексній та аспектній розробці корпусу текстів таких стилів:

- поетичного (300 тис. слововживань);
- публіцистичного (300 + 300 тис. слововживань);
- наукового (тексти з екології – 300 тис. , біології – 150 тис., фізики – 150 тис., мовознавства – 500 тис., літературознавства – 150 тис., філософії – 150 тис., юридичної лексики – 150 тис.).

Корпуси текстів є **дослідницькими**, оскільки призначені для вивчення різних аспектів функціонування мовної системи і зорієнтовані на широкий клас лінгвістичних задач; **статичними**, оскільки відображають певний часовий стан мовної системи. Наприклад, авторські корпуси поетичної мови – це колекції текстів окремих поетів кінця XX століття, публицистичні тексти – репрезентують мовлення газет за короткий проміжок часу (2004 рік). Водночас ці тексти дають можливість дослідити мовні явища в динаміці, тому вони є **динамічними**, а також інтерактивними, тому що користувач у проведенні дослідження може виділити з генерального корпусу робочий корпус, який включає лише частину генерального корпусу, напр., тексти якоїсь зони (Політика, Суспільство, Економіка, Спорт) або якоїсь газети «Дзеркало тижня», «Україна молода», «Українська правда».;

Невеликий за всіма мірками корпус текстів (більше 2 млн. словоформ) побудований таким чином, щоб користувач міг одержати різноаспектну інформацію про мовну одиницю (морфема, слово, словосполучення, речення):

- зовнішню – анотування тексту (автор, джерело, рік видання тощо);
- внутрішню – структурація тексту (номер розділу, абзацу, речення, слова тощо);
- граматичну, лексико-граматичну, лексичну;
- кількісну – про функціонування її як інваріантної форми (лема, морфема) і варіантної (словоформа, морф) зі статистичними характеристиками (абсолютна та середня частота, міра коливання середньої частоти, коефіцієнт стабільності).

Така сукупність характеристик дозволила створити **параметризовану базу даних,** під якою укладачі розуміють багатоаспектну і багатофункціональну систему, яка включає:

1. корпус текстів – джерело різного роду словників;
2. серію алфавітно-частотних словників з усією інформацією про слово: граматичною, лексико-граматичною, стилістичною та статистичною (абсолютна та середня частота, міра коливання середньої частоти, коефіцієнт стабільності);
3. алфавітно-частотні словники слів та слововживань спільної лексики.
4. словники неолексем;
5. словники синтаксичних моделей керування: дієслівних, іменникових, атрибутивних;

6. серію морфемних та словотвірних словників з частотними характеристиками морфа/морфеми, за якими можна вивчати комбінаторно-дистрибутивну будову, словотвірне значення кожної афіксальної морфеми в текстах;
7. словники тропів: епітетів, метафор, метонімії, порівнянь, синекдохи, оксюморонів, гіпербол;
8. словники синонімів, антонімів, фразеологізмів, тезауруси;
9. словник-конкорданс як допоміжний інструмент для формування лексико-семантичної, синтаксичної та стилістичної характеристики слова.

Ми розглядаємо параметризовану базу як реальну можливість для створення універсального (багатоцільового) словника. Але цілком закономірним постає питання: що є параметром у структурі бази? У зв'язку з тим, що термін цей неоднозначний і досить розповсюджений, ми визначаємо його як особливе словникове представлення мовних характристик одиниці. Ю. М. Караулов визначає цей термін так: «під параметром розуміємо деякий квант інформації про мовну структуру, який ...може... виступати у сполученні з іншими квантами (параметрами) і знаходити специфічне вираження у словниках» [2].

У базі передбачено параметри:

- граматичні (частина мови і категоріальні значення, напр., рід, число, відмінок, особа тощо);
- структурні (напр., моделі морфної структури слів різних частин мови; моделі керування дієслівні, іменні, атрибутивні тощо);
- лексико-семантичні, які віддзеркалюють системні відношення (синонімія, антонімія, омонімія, паронімія, ідеографія, тропіка);
- статистичні.

З цього неповного переліку видно, що:

1) деякі параметри є синкретичними (напр., частиномовний пов'язаний з категоріальним або структурним);

2) параметр завжди відноситься до лексеми/словоформи в цілому;

3) параметр може описувати окрему одиницю (напр., структуру або граматичну категорію);

4) параметр взагалі може відображати кількісні (статистичні) характеристики мовних одиниць.

5) параметри представляють багатовимірну класифікацію .

Комп'ютерна підтримка функціонування параметризованої бази забезпечується пакетом алгоритмів і програм автоматичного морфологічного і синтаксичного аналізу української мови, морфного сегментування тексту, укладання повних і часткових словників за частотою, алфавітом, а також пакетом статистичної обробки лінгвістичних даних. Створення алфавітного словника здійснюється в двох режимах: автоматичному (без зняття омонімії) і автоматизованому (зі зняттям лексичної і лексико-граматичної омонімії в деяких випадках ця функція також передбачена пакетом програм). Для виконання таких завдань, як створення словників неолексем, синонімічних груп, тезауруса, поетичних тропів та ін. передбачене програмне забезпечення в автоматизованому режимі, оскільки їх здійснення потребує контроль за значенням. При цьому обов'язковою є функція конкордансної ілюстрації з корпусу текстів. Компонентами програмного забезпечення є такі словники української мови: орфографічний, граматичний, морфемний, синонімічний, антонімів, фразеологічний, тлумачний, кожен з яких має комп'ютерне представлення, в розробці якого брали участь автори цього проекту. Програмно передбачене автоматичне порівняння авторських текстових словників зі словниками мови скорочує роботу лінгвіста у десятки разів.

Отже,

- Потреба в точних функціональних характеристиках мовних одиниць у різних стилях та жанрах гостро відчувалася в останні десятиліття фахівцями ряду галузей, що мають справу з комп'ютерним аналізом текстів, з укладанням навчальних словників, з теорією та практикою функціонування мови. Пропонована параметризована база даних є одним із перших кроків у заповненні цієї лакуни.
- Систематизована у серіях словників і списках інформація є важливою для граматичних, стилістичних, літературознавчих і семантичних досліджень лексичного фонду української мови в його статиці та динаміці.

- Запропонована методика опрацювання лінгвістичних даних в електронній базі є узагальненням комплексу теоретичних і прикладних ідей сучасного мовознавства. Технологія конструювання бази робить її надзвичайно ефективним та раціональним інструментом (вона зберігає багато часу та людських ресурсів) для спеціалістів-філологів різного профілю.
- Електронна база даних зі своєю методологією і технологією допомагає ефективно й оперативно здійснювати масштабні комплексні філологічні дослідження на рівні сучасної наукометрії.
- Ця база даних використовується у навчальному процесі на філологічних факультетах університетів, забезпечуючи необхідний теоретичний і технологічний рівень підготовки фахівців.
- Створена нами база даних починає використовуватися як готовий продукт і як модель для конструювання аналогічних баз університетами України.

Література

1. Апресян Ю. Д. О толковом словаре управления и сочетаемости русского глагола// Словарь. Грамматика. Текст. – М., – 1998. С.13.
2. Караулов Ю. Н. Лингвистическое конструирование и тезаурус литературного языка. М., 1981. С.51.

О. Зубань [*]
Київський національний університет ім. Т. Шевченка (Київ)
УДК 81’33

ПАРАМЕТРИЗОВАНА БАЗА ДАНИХ ЯК ІНСТРУМЕНТ ДОСЛІДЖЕННЯ КОРПУСУ ТЕКСТІВ

Розвиток теорії і практики прикладної лінгвістики, зокрема методів комп'ютерного моделювання, дозволили по-новому сформулювати лексикографічне завдання, а саме – створення автоматизованої системи лінгвістичного аналізу тексту. Структура цієї системи базується на модульно-словниковій ідеології: системно-структурні відношення одиниць кожного рівня мовної системи представлені в окремому модулі: фонетичному, морфемному, словотвірному, морфологічному, синтаксичному, семантичному.

У традиційній лінгвістиці наукова значущість, вагомість результатів дослідження визначається перш за все репрезентабельністю матеріалу дослідження: чим більше мовних фактів, тим достовірніші спостережувані закономірності, але традиційна форма збирання і систематизації інформації (переважно паперова картотека та використання різноманітних паперових словників) сьогодні не задовольняє потреб дослідників-філологів. Потрібні нові інформаційні технології, які б оптимізували роботу дослідника. Тому в українському мовознавстві на сьогодні нагальною є проблема укладання електронних лінгвістичних словників, які мають формат параметризованих електронних баз даних, оснащених пошуково-класифікаційними програмними аналізаторами, що забезпечують:

- ефективне та оперативне проведення лінгвістичного аналізу;
- можливість аналізу великих лексичних масивів;
- отримання точних формальних характеристик мовних одиниць різних рівнів як у системі мови, так і реляційно-функціональних особливостей їх у тексті.

У сучасній комп’ютерній лінгвістиці досить плідно розвивається лексикографічний напрямок. Зокрема в українському мовознавстві сьогодні створено чимало різногалузевих електронних словників, з різноманітними системами навігацій та інтерфейсами, частина з яких стала вже лінгвістичним комерційним продуктом і доступна для широкого кола користувачів. Такі бази даних покликані виконувати функцію своєрідних довідників для лінгвіста-дослідника і, без сумніву, є надзвичайно важливими для організації повномасштабного дослідження мови, але вони є **статичними**, їх не можна використовувати в режимі автоматизованого аналізу тексту. Тому особливої уваги сьогодні заслуговують ті електронні лінгвістичні продукти, які спрямовані на аналіз тексту і мають статус **динамічних** пошукових систем, які здатні в автоматичному або автоматизованому режимі вилучати інформацію про мовні одиниці з будь-якого параметризованого тексту.

Текст як основна форма збереження і передачі інформації, будучи результатом мовленнєвого акту, є інвентарем мовних одиниць, які комбінуються в ньому за законами