РОЗДІЛ X. ПРИКЛАДНА ЛІНГВІСТИКА: НАПРЯМИ Й АСПЕКТИ ДОСЛІДЖЕННЯ

Ілля Данилюк

УДК 81’33=161.2

КОРПУС ТЕКСТІВ ДЛЯ ВИВЧЕННЯ ГРАМАТИЧНОЇ СЛУЖБОВОСТІ

У статті розглянуто основні теоретичні та практичні підходи до створення новітнього корпусу текстів української мови, призначеного, серед інших загальних лінгвістичних цілей, для вивчення функціонально-комунікативної природи службових частин мови – прийменника, частки і сполучника. Описано вимоги до корпусу, поставлені в межах проекту завдання, технічні й програмні аспекти реалізації корпусу текстів, система тегів для службових частин мови.

Ключові слова: корпус текстів, корпусний менеджер, розмітка, тег, службові частини мови.

У сучасній лінгвістиці, озброєній потужними апаратно-технічними можливостями персональних і суперкомп’ютерів, що функціонують або незалежно один від одного, або частіше у складі мережі, «хмари», можливість опрацювання величезної кількості емпіричних мовних даних відкриває величезні перспективи. Хоча ідеться насамперед про поступ у вирішенні базових завдань комп’ютерної лінгвістики – машинний переклад, розпізнавання і синтез мовлення, із частковою поразкою у яких вчені й потенційні користувачі певний час тому, по суті, змирилися, – основні теоретичні й прикладні напрями мовознавства також отримують новий імпульс через використання корпусів текстів (КТ). Вивчення мовних явищ із використанням КТ уможливлює не тільки підтвердження чи спростування окремих аспектів теорії, а й дозволяє здійснити відкриття невідомих закономірностей у функціонуванні мовних одиниць різних рівнів. Напрям *корпусної лінгвістики* упевнено посів одне з чільних місць у прикладному мовознавстві, започаткований працею Г. Кучери та В. Нельсона Френсіса *Computational Analysis of Present-Day American English*, виданій у 1967 році. Досить детально розроблено теоретичні засади напряму [Facchinetti 2007], вироблено його методи [Wallis, Nelson 2001]: анотування корпусу (розмічування), екстракція (пошук) даних у корпусів, аналіз корпусу. Сьогодні напрям представлено кількома періодичними виданнями: *Corpora, Corpus Linguistics and Linguistic Theory, ICAME Journal* і *the International Journal of Corpus Linguistics*.

У цій статті, яка стане першою у майбутньому циклі публікацій, ми ставимо собі за **мету** розкрити основні етапи створення КТ української мови для вивчення граматичних явищ, пов’язаних із функціонуванням службових частин мови – частки, прийменника і сполучника, і намагатимемося реалізувати такі **завдання**: 1) описати передумови створення КТ граматичної службовості, у тому числі проблему вибору корпусного менеджера; 2) визначити перелік завдань, що виникнуть у в’язку зі створенням такого КТ; 3) описати теоретичні засади побудови металінгвістичної і власне лінгвістичної розмітки; 4) подати систему тегів для основних службових одиниць української мови.

Під терміном лінгвістичний, або мовний, корпус текстів сьогодні розуміють великий, представлений в електронному вигляді, уніфікований, структурований, розмічений, філологічно компетентний масив мовних даних, призначений для вирішення конкретних лінгвістичних завдань [Захаров 2005: 3]. Таким завданням для кафедри української мови та прикладної лінгвістики Донецького національного університету стало багатоаспектне вивчення граматики службовості, або тих класів слів, що прийнято називати службовими частинами мови: насамперед прийменника, частки та сполучника. Першим етапом стала розробка граматики прийменника, що мала результатом розроблену класифікацію прийменникових параметрів, виданий словник [Загнітко, Ситар та ін. 2007], низку теоретичних публікацій. Цей етап було реалізовано разом із колегами з Москви, Гродно, Познані, Великотирново, Ополе у 2006-2009 роках. Тоді корпусом текстів для збирання статистичних даних функціонування прийменникових одиниць виступало зібрання текстів трьох функційних стилів – художнього, наукового, офіційно-ділового, обсягом по 1,5 мільйона слововживань у кожному. Безперечно, це був зовсім не корпус текстів, однак інших дієвих інструментів квантитативних досліджень на той час у розпорядженні кафедри не було.

Другим етапом було вивчення граматики частки, що здійснювалось за матеріалами усіх доступних фундаментальних словників української мови – тлумачних (для вияву системної динаміки часток та їхньої кваліфікації і класифікації), етимологічних, історичних, а також на ґрунті робочого КТ. Усе це дозволило встановити характерологію частки, її функційні і комунікативні вияви, простежити синтагматику і парадигматику, простежити динаміко-еволюційні параметри та з’ясувати етимологію [Загнітко, Каратаєва 2012]. Слабкою ланкою постала квантифікація, тому що робочий корпус виявився недостатнім для адекватного простеження кількісних параметрів цих дискурсивних слів. Виникло чітке відчуття потреби у реальному корпусі текстів, оскільки розроблена класифікація граматичних ознак частки не була представлена у жодному з наявних корпусів не тільки української, але й будь-якої слов’янської мов, що призвело до зародження самої ідеї створення новітнього КТ.

Загалом робота з корпусами, представленими в електронному вигляді, давно вже стала одним із основних методів лінгвістичних досліджень. Так, ще в 1960-і роки створювався Браунівський корпус (США), який містив 1 млн слів. У 1970-і роки минулого століття стартував LOB корпус (Великобританія, Норвегія), у

1980-ті роки почали створюватися такі корпусу, як: Машинний Фонд російської мови, Уппсальський корпус російської мови (Швеція) – обидва по 1 млн слів, The Bank of English, Birmingham, – 20 млн слів. (слайд). У 1990-і роки було створено British National Corpus, який включав на той час 100 млн слів, а також інші національні корпуси для угорської, італійської, хорватської, чеської, японської мови обсягом по 100 млн слів. На початку XXI ст. створювалися такі корпуси, як American National Corpus і Gigaword corpora (англійська, арабська, китайська) на 1 млрд слів, Національний корпус російської мови, над яким працюють лінгвісти Москви і Санкт-Петербурга, містить 300 млн слововживань. В Україні проблемою корпусної лінгвістики активно займаються вчені Інституту української мови НАН України, Українського мовно-інформаційного фонду, Інституту філології Київського національного університету ім. Т.Шевченка, Національного університету «Львівська політехніка» та ін.

Однак певні обмеження в доступі до наявних корпусів української мови, і насамперед відсутність у них необхідної інформації щодо граматичної природи службових змусили нас удатися до створення власного корпусу текстів, що відповідав би потребам вчених, що працюють над проблемою граматичної службовості.

Загальні вимоги до сучасного корпусу текстів, зважаючи на складність лінгвістичних завдань, є досить високими. Як вказала Орися Демська-Кульчицька, ці вимоги є такими [Демська-Кульчицька 2005: 9-19]: по-перше, корпус повинен бути достатньо великого обсягу, по-друге, він повинен бути структурованим або розміченим, по-третє, тексти, що входять до певного корпусу, повинні бути в електронному варіанті, і, по-четверте, як правило, для роботи з корпусом створюється спеціальне програмне забезпечення – *корпусний менеджер* (КМ).

Цінність будь-якого створюваного корпусу текстів, у тому числі й КТ граматичної службовості, ми услід за авторкою [Демська-Кульчицька 2005: 9-19] бачимо ось у чому:

• корпус може використовуватися багаторазово: закладені сьогодні в КТ граматичної службовості підходи дозволять використовувати його для лінгвістичних досліджень, пов'язаних не тільки з морфологією службовості, але й іншими напрямками;

• КТ показує мовні дані в їхньому реальному оточенні, що дозволяє досліджувати лексичну і граматичну структуру мови, а також безперервні процеси мовних змін, що відбуваються в мові впродовж певного відрізку часу;

• КТ повинен характеризуватися репрезентативністю, або збалансованим складом текстів, що дозволить надалі використовувати його для тестування пошукових машин, машинних граматик, систем перекладу і для вирішення інших завдань прикладної лінгвістики;

• КТ має важливе значення для викладання мови, оскільки з його допомогою можна швидко й ефективно перевірити особливості вживання незнайомого слова чи граматичної форми;

• КТ може бути своєчасно доповнений і розширений, щоб відповідати дедалі вищим вимогам майбутньої лінгвістики і вирішувати нові завдання.

У циклі статей ми маємо намір розкрити реалізацію низки завдань, які постали у процесі розробки КТ граматичної службовості:

1. Обрати класифікацію форм мовлення для української мови різних періодів;
2. Закласти у корпус ємну класифікацію функціональних стилів і підстилів української мови для сучасних прикладних досліджень;
3. Усталити для потреб корпусу класифікацію жанрів текстів, що входять до нього;
4. Визначити часові етапи (періоди) функціонування української мови для класифікації одиниць корпусу текстів за хронологічним принципом;
5. Розробити метатекстову розмітку (опису текстів) для корпусу;
6. Розробити або адаптувати корпусний менеджер до роботи з українським корпусом текстів;
7. Закласти у корпус відповідну класифікацію граматичних класів слів (частин мови) української мови з урахуванням потреб сучасних і перспективних лінгвістичних досліджень;
8. Розробити систему тегів для символічного позначення граматичних класів і підкласів слів у корпусі;
9. Розробити регулярні вирази для створення пошукових запитів у корпусі;
10. Розробити загальний алгоритм і конкретний інструмент для перетворення довільного тексту у форму вертикального файлу для корпусу тексту;
11. Розробити загальний алгоритм і конкретний інструмент автоматизованої лематизації одиниць корпусу (визначення початкової форми для словозмінних одиниць);
12. Розробити загальний алгоритм і конкретний інструмент побудови морфологічної розмітки для одиниць корпусу;
13. Компілювати і запустити корпус текстів у локальній мережі лабораторії кафедри української мови та прикладної лінгвістики ДонНУ;
14. Реалізувати на базі корпусу текстів низку прикладних завдань наукових проблем кафедри української мови та прикладної лінгвістики ДонНУ.

Отже, у циклі статей послідовно розкриємо окремі кроки цього плану.

Найважливішим складником корпусу є, звичайно, розмітка, що відрізняє його від електронних колекцій текстів, бібліотек, енциклопедій, причому чим вона багатша і повніша, тим вищою є наукова чи навчальна цінність корпусу. Відомо, що існують різні типи розмітки:

1) *екстралінгвістична:*
   - *метатекстова* (інформація про автора, назву, дату створення, обсяг, тематику тощо), що характеризує текст загалом;
   - *структурна*, яка є інформацією про структуру тексту і дозволяє відокремити одне слово від іншого, виділити розділові знаки, окремі слова, межі словосполучення, речення, абзацу, тексту;
   - *типографська,* яка для максимального відображення паперового джерела включає позначки поділу на сторінки, вказівку гарнітури й оформлення шрифту (жирний, курсив), кольору шрифту, позначки для покликань, гіперпокликань, малюнків і рисунків, таблиць і графіків тощо;
2) *лінгвістична*, що містить певну лінгвістичну інформацію про виділену одиницю (слово або речення) і включає:
   - *морфологічну розмітку*, для якої в англомовній літературі використовується термін *part-of-speech tagging (POS-tagging)*, дослівно – частиномовна розмітка. Насправді морфологічні мітки включають не тільки ознака частини мови, але й ознаки граматичних категорій, властивих цій частині мови. Це основний тип розмітки: по-перше, більшість великих корпусів є якраз морфологічно розміченими корпусами, по-друге, морфологічний аналіз розглядається як основа для подальших форм аналізу – синтаксичного і семантичного, і, по-третє, успіхи в комп'ютерній морфології дозволяють автоматично розмічувати корпуси великих розмірів;
   - *синтаксичну розмітку*, яка є результатом синтаксичного аналізу, або *парсингу (англ. parsing)*, що виконується на основі даних морфологічного аналізу. Цей вид розмітки описує синтаксичні зв'язки між лексичними одиницями і різні синтаксичні конструкції (наприклад, підрядне речення, дієслівне словосполучення тощо);
   - *семантичну розмітку,* у якій найчастіше семантичні теги позначають семантичні категорії, до яких належить слово чи словосполучення, і вужчі підкатегорії, що визначають його значення; (наприклад, *нота* – музичний термін або термін зі сфери дипломатії)
   - *анафоричну розмітку*, яка фіксує референтні зв'язки, наприклад, займенникові;
   - *просодичну розмітку,* у якій застосовуються мітки, що описують наголос та інтонацію.

До метатекстової розмітки КТ граматичної сужбовості ми заклали три форми мовлення: дві виділювані традиційно – *писемну* й *усну*, а також *мережеву* – новітню для української мови форму, існування якої можна простежувати з останніх років минулого століття. У цій формі функціонують різноманітні сайти, форуми, чати, блоги, твітер, sms тощо. На наш погляд, корпус цих текстів виявить нові мовні явища, які стануть перспективним об’єктом дослідження.

У межах кожної із форм пропонуємо розрізняти 6 функційних стилів української мови із під стилями (таблиця 1):

Табл. 1. Класифікація стилів і підстилів КТ граматичної службовості

Умовн. позн.	Стиль	Підстиль
X.	Художній	1) прозовий
		2) ліричний
		3) драматичний
		4) комбінований
N.	Науковий стиль	1) академічний
		2) науково-популярний
		3) науково-навчальний
		4) науково-інформативний
		5) науково-довідковий
		6) науково-технічний
		7) науково-оцінний
O.	Офіційно-діловий	1) законодавчий
		2) дипломатичний
		3) адміністративно-канцелярський
		4) судовий
P.	Публіцистичний стиль	1) інформаційний
		2) аналітичний
		3) художньо-публіцистичний
		4) науково-публіцистичний
		5) рекламно-довідковий
		6) політичний
K.	Конфесійний стиль	1) сакральний
		2) літургійний
		3) проповідницький
		4) науково-конфесійний
		5) навчально-конфесійний

		6) художньо-конфесійний
R.	Розмовний стиль	1) розмовно-побутовий
		2) розмовно-офіційний

У процесі розробки класифікації відмовилися від виділення епістолярного стилю, включивши його основні жанри до розмовного стилю писемної форми, оскільки він на сьогодні мало репрезентативний і його функційні вияви важко, іноді майже неможливо встановити. Безперечно, можна скористатися спадщиною митців, але це буде відносний синхронний зріз, опосередкований щонайменше п'ятдесятьма-шістдесятьма роками. Та й у своєму загалі епістолярний стиль послуговується конструкціями розмовного стилю. Інша річ, що йому притаманна особлива жанровість, але вона також у своїй репрезентативності обмежена.

Опис типології жанрів у межах кожного стилю ми плануємо подати у наступних публікаціях серії.

Металінгвістична розмітка КТ граматичної службовості також включає інформацію про час появи тексту. З погляду часових періодів функціонування української мови пропонуємо розрізняти 5 зрізів, мотивуючи поділ наявністю суттєвих правописних і граматичних відмінностей у мові:

1. Українсько-руський період – X-XV століття;
2. Староукраїнський період – XVI-XVIII століття;
3. Новоукраїнський період – XIX століття-1933 р.;
4. Українсько-радянський період 1933-1991;
5. Новітній український – З 1991 року.

Останні два періоди разом кількісно однозначно обіймуть понад 90% текстів корпусу, однак у перших трьох функціонування окремих стилів також є формально зафіксованим. Наприклад, в українсько-руському періоді можна вирізнити *науковий стиль* (це твори «Фізіолог», «Шестиднев» Іоанна Екзарха, «Християнська топографія» Козьми Індикоплевста, хроніки Іоанна Малали, Георгія Амартола, Георгія Синкела), *офіційно-діловий* (листи, грамоти); *публіцистичний* (жанри: слова і повчання); *художній* (прозові жанри – житійно-повістева, легендарна література – житія Ольги, Володимира, Києво-Печерський Патерик, «Четья» 1489; літописний жанр: «Повість врем'яних літ»).

Зі староукраїнського періоду XVI-XVIII століття збереглися численні зразки *конфесійного стилю* (Пересопницьке Євангеліє, Крехівський Апостол), *наукового стилю* (жанри: літописи, хронографи, передмови, післямови, посвяти; граматики, тлумачні і перекладні словники; науково-практичні жанри: лікарські і господарські порадники, календарі, травники, місяцеслови), *офіційно-ділового стилю* (листи, грамоти; універсали, накази, інструкції, ордери, про меморії; дипломатичний жанр: листування гетьманських канцелярій; юридичний жанр: «Литовський статут», актові книги, купчі записи, духівниці, описи майна, супліки-скарги, судові документи, записи свідчень), *публіцистичного стилю* (жанри: слова і повчання, полемічний жанр – твори Герасима Смотрицького, Василя Суразького, Івана Вишенського, Стефана Зизанія, ораторсько-проповідницький жанр: Кирило Транквіліон-Ставровецький, Лазар Баранович, Йоаникій Галятовський, Антоній Радивиловський), *художнього стилю* («Четьї-Мінеї» Данила Туптала, твори Димитрія Ростовського, апокрифи, легенди; літописно-мемуарний жанр: Острозький, Львівський, Хмільницький літописи, літопис Самовидця 1702 р., Григорія Граб'янки 1710 р., Самійла Величка 1720 р., віршовані жанри: епіграми, емблеми, віршовані передмови, історичні вірші, вірші Івана Величковського, Климентія Зіновіїва; драматичні жанри: шкільні драми Феофана Прокоповича) [Українська мова 2004: 643-654]. Стильове та жанрове різноманіття наступних періодів не потребує коментарів.

Структурна розмітка КТ граматичної службовості наразі включає дані тільки про межі тексту, це тег <doc> – </doc>, і межі речень у ньому тег <s> – </s>. Спеціальний тег <g/> позначає розділові знаки, перед якими не ставиться пробіл – це всі, крім тире. У перспективі маємо намір додати автоматично виставлені позначки абзаців, параграфів, розділів, частин тощо. Лінгвістична розмітка – завжди найскладніша і найважливіша частина корпусу. Оскільки було обрано за основний корпусний менеджер відкритий проект NoSketch Engine (http://nlp.fi.muni.cz/trac/noske), розроблений в університеті Масарика (Брно, Чехія) [Rychlý, Pavel, Smrž, Pavel 2004], то форма представлення розмітки у вигляді тегів була єдино можливою. Структура стандартного тегу є такою: на першій позиції в тегу позначка граматичного класу слова, далі позначки підкласів, усі позначки одно символьні латиницею або цифрами, за кожним підкласом закріплена позиція, яка не змінюється для різних класів (детально описати принципи тегування маємо намір у наступних публікаціях).

Наприклад, для слова *конференцією* тег має вигляд – Izzooin1m (сх. 1).

Схема 1. Розшифрування тегу для словоформи «конференцією»

іменник		*жіночий рід*		*орудний відмінок*		*Неістота*		*м'яка група*
I	Z	Z	O	O	I	N	1	M
	загальна назва		*однина*		*іменниковий тип відмінювання*		*перша відміна*	

У КТ граматичної службовості закладено такий поділ слів на граматичні класи із позначенням їх за ключовими літерами українських термінів (табл. 2):

Таблиця 2. Класи слів у КТ граматичної службовості

№	Клас	Тег
1)	(І)менник	I
2)	(Д)ієслово	D
3)	При(к)метник	K
4)	При(с)лівник	S
5)	(Ч)ислівник	C
6)	(З)айменник	Z
7)	Час(т)ка	T
8)	С(п)олучник	P
9)	При(й)менник	J
10)	(В)игук	W
11)	(А)бревіатура	A
12)	(Р)ешта	R

Нарешті, на завершення першої із серії публікацій наведемо розгорнуту систему тегів для частини мови, яка є одним із безпосередніх об’єктів вивчення – частки (табл. 3).

Таблиця 3. Система тегів для класу частки

Позиція в коді	Атрибут	Значення	Тег
0	Час(т)ка		T
1	Походження	п(е)рвинна	e
		в(т)оринна	t
2	Тип (для вторинних)	(в)ідсполучникова	v
		в(і)ддієслівний	i
		відз(а)йменниковий	a
		від(п)рислівниковий	p
3	Будова	(п)роста	p
		с(к)ладна	k
		ск(л)адена	l
4	Статус	власне час(т)ка	t
		частка-(в)исловлення	v
		(а)налог частки	a
5	Дистинктивний тип	(а)пелятивна	a
		(о)цінна	o
		(е)моційна	e
		(т)екстоструктурувальна	t
6	Функційний тип	Вказівна	f
		категорійно-комунікативні : питальна	l
		Спонукальна	m
		заперечна	q
		Стверджувальна	1
		дискурсивні : підсилювальна	2
		роз’яснювальна	3
		Обмежувальна	4
		Зіставна	5
		Відокремлювальна	6
		Видільна	7
		Порівняльна	8
		Вірогідна	9

Отже, розробка КТ граматичної службовості, як, зрештою, і будь-якого корпусу текстів є складним багатокроковим процесом. На сьогодні завершено створення україномовного інтерфейсу для корпусного менеджера NoSketchEngine Manatee/Bonito, розроблено систему тегів для металінгвістичної та власне лінгвістичної морфологічної розмітки, що буде описано у наступних публікаціях серії. Найважчим завданням попереду є визначення кількісних і якісних параметрів критерію репрезентативності для текстів корпусу, механізмів автоматичного розмічування.

Література

Баранов 2003: Баранов, А.Н. Введение в прикладную лингвистику [Текст] / А. Н. Баранов. – М. : Едиториал УРСС, 2003. – 360 с. – ISBN 5-354-00313-X

Бук 2007: Бук, С. Корпус текстів Івана Франка : спроба визначення основних параметрів [Текст] // Прикладна лінгвістика та лінгвістичні технології : MegaLing‑2006 : Зб. наук. пр. / НАН України. Укр. мовн.-інформ. фонд, Таврійськ. нац. ун-т ім. В. І. Вернадського ; за ред. В. А. Широкова. – К. : Довіра, 2007. – С. 72-82.

Демська-Кульчицька 2005: Демська-Кульчицька, О. Основи національного корпусу української мови [Текст] / О. Демська-Кульчицька. – К. : Інститут української мови національної академії наук України, 2005. – 219 с.

Захаров 2005: Захаров, В.П. Корпусная лингвистика : Учебно-метод. пособие. – СПб., 2005. – 48 с.

Зубов 2004: Зубов, А.В. информационные технологии в лингвистике : Учеб. пособие для студ. лингв. фак-тов высш. учеб. заведений [Текст] / А. В. Зубов, И. И. Зубова. – М. : Издательский центр «Академия», 2004. – 208 с.

Карпіловська 2006: Карпіловська, Є.А. Вступ до прикладної лінгвістики : комп'ютерна лінгвістика. Підручник [Текст] / Є. А. Карпіловська. – Донецьк : ТОВ «Юго-Восток, Лтд», 2006. – 188 с. – ISBN 966-374-078-7.

Корпусна лінгвістика 2005: Корпусна лінгвістика : Моногр. [Текст] / В. А. Широков, О. В. Бугаков, Т. О. Грязнухіна, О. М. Костишин, М. Ю. Кригін ; НАН України, Укр. мов.-інформ. фонд. – К. : Довіра, 2005. – 472 с.

Загнітко, Каратаєва 2012: Загнітко, А.П., Каратаєва Г.О. Словник часток : матеріали і статті [Текст] / Анатолій Загнітко, Анна Каратаєва ; Донец. нац. ун-т. – Донецьк : ДонНУ, 2012. – 381 с.

Загнітко, Ситар, Данилюк, Щукіна 2007: Загнітко, А.П., Ситар, Г.В., Данилюк, І.Г., Щукіна, І.А. Словник українських прийменників. Сучасна українська мова [Текст]. – Донецьк : ТОВ ВКФ «БАО», 2007. – 416 с.

Загнітко, Ситар, Данилюк 2012: Загнітко, А.П., Ситар, Г.В., Данилюк, І.Г., Структура і модель бази даних «Українські частки та їхні еквіваленти» [Текст] // Комп'ютерна лінгвістика : сучасне та майбутнє. Матеріали Міжнародної науково-практичної конференції. – К. : КНЛУ, 2012. – С. 21-22.

Загнитко, Ситарь, Данилюк 2011: Загнитко, А.А., Ситарь, А.В., Данилюк, И.Г. Особенности структурирования базы данных служебности ↔ вспомогательности : на материале украинских частиц [Текст] // Проблемы компьютерной лингвистики : Сборник научных трудов под ред. А. А. Кретова. – Вып. 5. – Воронеж, 2011. – С. 56-67.

Українська мова 2004: Українська мова : Енциклопедія / Редкол. : Русанівський, В. М. (співголова), Тараненко, О. О. (співголова), Зяблик М. П. та ін. – 2-ге вид., випр. і доп. – К. : Вид-во «Укр. енцикл.» ім. М. П. Бажана, 2004. – 824 с. – ISBN 966-7492-19-2.

Biber, Conrad, Reppen 1998: Biber, D., Conrad, S., Reppen, R. Corpus Linguistics, Investigating Language Structure and Use [Текст]. – Cambridge : Cambridge UP, 1998. – ISBN 0-521-49957-7

Facchinetti 2007: Facchinetti, R. (ed.) Corpus Linguistics 25 Years on. [Текст]. – New York / Amsterdam : Rodopi, 2007. – ISBN 978-90-420-2195-2

Rychlý, Pavel, Smrž, Pavel 2004: Rychlý, Pavel, Smrž, Pavel. Manatee, Bonito and Word Sketches for Czech. In Proceedings of the Second International Conference on Corpus Linguisitcs [Текст]. – Saint-Petersburg : Saint-Petersburg State University Press, 2004. – Pp. 124-132. – ISBN 5-288-03531-8.

Wallis, Nelson 2001: Wallis, S., Nelson, G. Knowledge discovery in grammatically analysed corpora [Текст] // Data Mining and Knowledge Discovery, 5. – 2001. – Pp. 307-340.

Wynne: Wynne, M. (editor). Developing Linguistic Corpora : a Guide to Good Practice [Електронний ресурс]. – Oxford : Oxbow Books, 2005. – Access mode : URL http://ahds.ac.uk/linguistic-corpora. – Title from the screen. – ISSN 1463 5194.

В статье рассмотрены основные теоретические и практические подходы к созданию новейшего корпуса текстов украинского языка, предназначенного, кроме общих лингвистических целей, для изучения функционально-коммуникативной природы служебных частей речи – предлога, частицы и союза. Описаны требования к корпусу, поставленные в рамках проекта задания, технические и программные аспекты реализации корпуса текстов, система тегов для служебных частей речи.

Ключевые слова: корпус текстов, корпусный менеджер, разметка, тег, служебные части речи.

The article discusses the basic theoretical and practical approaches to the design of new corpora for the Ukrainian, which is designed among other general linguistic purposes for studying functional and communicative nature of the syntactic parts of speech (POS) – preposition, conjunction and particle. We describe the requirements for building, objectives set in the project, technical and program aspects of the corpora, tag system for syntactic POS.

Keywords: corpora, corpora manager, tagging, tag, syntactic POS.

Надійшла до редакції 7 серпня 2012 року.