РОЗДІЛ ІХ. ПРИКЛАДНА ЛІНГВІСТИКА: НАПРЯМИ Й АСПЕКТИ ДОСЛІДЖЕННЯ

Ілля Данилюк

УДК 81’33=161.2

КОРПУС ТЕКСТІВ ДЛЯ ВИВЧЕННЯ ГРАМАТИЧНОЇ СЛУЖБОВОСТІ: КЛАСИФІКАЦІЯ ГРАМАТИЧНИХ КЛАСІВ І ПІДКЛАСІВ

У статті, яка є другою публікацією у циклі, присвяченому опису теоретичних і практичних засад створення корпусу текстів для вивчення граматичної службовості у межах наукової теми кафедри української мови і прикладної лінгвістики ДонНУ, описано закладену в корпус класифікацію граматичних класів і підкласів. Подано принцип побудови тегів, наведено їх повну систему для першої версії корпусу.

Ключові слова: корпус текстів, частини мови, граматичний клас, граматичний підклас, тег.

Пропонована стаття є продовженням розпочатого циклу публікацій, присвяченого опису теоретичних і практичних засад створення корпусу текстів для вивчення граматичної службовості у межах наукової теми кафедри української мови і прикладної лінгвістики ДонНУ (перша публікація [Данилюк 2013]). У ній ми ставимо собі за **мету** розкрити класифікацію частин мови й виділених у їхніх межах категорій, або, іншими словами, типологію граматичних класів і підкласів для одиниць корпусу. Конкретними **завданнями** є: 1) опис загальної системи тегування у корпусі; 2) подання граматичних класів й одиниць неграматичної природи; 3) наведення класифікації підкатегорій у межах кожного з виділених класів.

Нагадаємо, що 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]. У корпусі використовуються різні види розмітки, з яких лінгвістична є завжди найскладнішою і найважливішою. Корпусний менеджер на основі NoSketch Engine [Rychlý, Smrž 2004] передбачає форма представлення розмітки у вигляді тегів. Наведемо повторно приклад тегу і його розшифрування для слова *конференцією* – тег *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*	

Нагадаємо структуру тегу в корпусі: на першій позиції позначка класу слова, далі позначки підкласів. Класифікація класів має такий вигляд (табл. 1):

Таблиця 1. Класи слів у КТ граматичної службовості

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

Перші десять класів – традиційно виділювані частини мови. До проблеми їх класифікації в українській мові ми зверталися неодноразово [Данилюк 2006; Данилюк 2010], остання, так звана «шкільна», була обрана з

міркувань сумісності корпусу з низкою вже здійснених, поточних і майбутніх кафедральних досліджень. Два останні класи включили той загал слів, які не входять до десятки основних класів. Літера в дужках в українських термінах – підстава для вибору символу на позначення класу в системі тегів. Виділення підкласів спирається на низку класичних [Безпояско 1993; Вихованець 2004; Курс сучасної української літературної мови 1951; Кучеренко 1961, 1964; Леонова 1983; Сучасна українська літературна мова 1969] і новітніх [Загнітко, Каратаєва 2012; Загнітко, Ситар та ін. 2007] досліджень.

Дотримуючись традиційного принципу, що всі позначки є односимвольними, ми, тим не менше, вирішили вперше, наскільки нам відомо, використати підхід, за якого, по-перше, на кожній позиції в системі тегів підкласів (першій, другій, третій і далі) для всіх класів слів використовується унікальний символ (іншими словами, наприклад, на другій позиції для певного підкласу іменника, числівника, частки, абревіатури й усіх інших класів буде унікальний символ, а не повторюваний, як це спостерігаємо в інших системах). По-друге, порядок підкласів є не довільним, а таким, що кожний підклас, якщо він виділяється в різних класів, займає одну й ту ж позицію (наприклад, підклас *походження* для частки, сполучника, прийменника й вигуку – на 2-ій позиції, підклас *рід* для іменника, дієслова, прикметника, числівника, займенника, абревіатури – на 3-ій позиції тощо). Перевагою такого підходу ми бачимо суттєве спрощення пошукових запитів – по-перше, не треба додатково задавати символ класу, якщо досліджується його унікальний підклас (наприклад, відсполучниковий тип частки), а по-друге, можна легко будувати запити на аналіз окремого підкласу безвідносно до класу слова (наприклад, виділити усі словоформи родового відмінка усіх граматичних класів). У цьому новизна системи тегів корпусу службовості.

Система тегів для класу іменника (I) має такий вигляд (табл. 2):

Таблиця 2. Система тегів для класу іменника

Позиція в коді	**Атрибут**	**Значення**	**Тег**
0	(І)менник		I
1	Назва	(в)ласна	V
		(з)агальна	Z
2	Рід	(ч)оловічий	C
		(ж)іночий	Z
		(с)ередній	S
3	Число	(о)днина	O
		(м)ножина	M
		(S)ing. tantum	S
		Pl. (t)antum	T
4	Відмінок	(н)азивний	N
		(р)одовий	R
		(д)авальний	D
		(з)нахідний	Z
		(о)рудний	O
		(м)ісцевий	M
		(к)личний	K
5	Тип відмінювання	(і)менниковий	I
		(п)рикметниковий	P
6	Істота	(і)стота	I
		(н)еістота	N
7	Відміна	І відміна	1
		ІІ відміна	2
		ІІІ відміна	3
		ІV відміна	4
8	Група	(т)верда	T
		(м)’яка	M
		м(і)шана	I

У корпусі закладено таку систему тегів для класу дієслова (D) (табл. 3):

Таблиця 3. Система тегів для класу дієслова

Позиція в коді	**Атрибут**	**Значення**	**Тег**
0	(Д)ієслово		D
1	Тип	(о)собове	O
		(і)нфінітив	I

		дієпри(к)метник	K
		дієпри(с)лівник	S
		(ф)орма на –но-то	F
		дієс(л)ово стану	L
		зв’(я)зка	J
2	Рід	(ч)оловічий	C
		(ж)іночий	Z
		(с)ередній	S
3	Число	(о)днина	O
		(м)ножина	M
		(S)ing. tantum	S
		Pl. (t)antum	T
4	Відмінок	(н)азивний	N
		(р)одовий	R
		(д)авальний	D
		(з)нахідний	Z
		(о)рудний	O
		(м)ісцевий	M
		(к)личний	K
5	Вид	(д)оконаний	D
		(н)едоконаний	N
6	Стан	(а)ктивний	A
		(п)асивний	P
7	Час	(т)еперішній	T
		(м)инулий	M
		м(а)йбутній	A
		(д)авноминулий d	
8	Спосіб	(д)ійсний	D
		(н)аказовий	N
		(у)мовний	U
9	Особа	1-а	1
		2-а	2
		3-я	3
10	Перехідність	(п)ерехідне	P
		(н)еперехідне	N
11	Валентність	0-валентне	0
		1-валентне	1
		2-валентне	2
		3-валентне	3
		4-валентне…	4

Система тегів для класу прикметника (K) має такий вигляд (табл. 4):

Таблиця 4. Система тегів для класу прикметника

Позиція в коді	**Атрибут**	**Значення**	**Тег**
0	При(к)метник		K
1	Ступінь порівняння	(н)ульовий	N
		в(и)щий	Y
		н(а)йвищий	A
2	Рід	(ч)оловічий	C
		(ж)іночий	Z
		(с)ередній	S
3	Число	(о)днина	O
		(м)ножина	M
4	Відмінок	(н)азивний	N
		(р)одовий	R
		(д)авальний	D
		(з)нахідний	Z
		(о)рудний	O

		(м)ісцевий	M
		(к)личний	K
5	Розряд	(я)кісний	J
		(в)ідносний	V
		п(р)исвійний	R
6	Безвідносна міра якості	н(у)льовий	U
		п(о)мірний	O
		на(д)мірний	D
7	Форма	повна (с)тягнена	S
		повна (н)естягнена	N
		(к)оротка	K

Система тегів для класу прислівника (S) має такий вигляд (табл. 5):

Таблиця 5. Система тегів для класу прислівника

Позиція в коді	**Атрибут**	**Значення**	**Тег**
0	При(с)лівник		S
1	Ступінь порівняння	(н)ульовий	N
		в(и)щий	Y
		н(а)йвищий	A
2	Тип	(о)бставинний	O
		оз(н)ачальний	N
		мо(д)альний	D

Система тегів для класу числівника (C) має такий вигляд (табл. 6):

Таблиця 6. Система тегів для класу числівника

Позиція в коді	**Атрибут**	**Значення**	**Тег**
0	(Ч)ислівник		C
1	Розряд	власне кількісни(й)	J
		(д)робовий	D
		з(б)ірний	B
		(п)орядковий	P
2	Рід	(ч)оловічий	C
		(ж)іночий	Z
		(с)ередній	S
3	Число	(о)днина	O
		(м)ножина	M
4	Відмінок	(н)азивний	N
		(р)одовий	R
		(д)авальний	D
		(з)нахідний	Z
		(о)рудний	O
		(м)ісцевий	M
		(к)личний	K
5	Тип відмінювання	(і)менниковий	I
		(п)рикметниковий	P

Система тегів для класу займенника (Z) має такий вигляд (табл. 7):

Таблиця 7. Система тегів для класу займенника

Позиція в коді	**Атрибут**	**Значення**	**Тег**
0	(З)айменник		Z
1	Розряд	і(м)енниковий	M
		п(р)икметниковий	R
		(ч)ислівниковий	C
		прис(л)івниковий	L
2	Рід	(ч)оловічий	C

		(ж)іночий	Z
		(с)ередній	S
3	Число	(о)днина	O
		(м)ножина	M
4	Відмінок	(н)азивний	N
		(р)одовий	R
		(д)авальний	D
		(з)нахідний	Z
		(о)рудний	O
		(м)ісцевий	M
		(к)личний	K
5	Тип відмінювання	(і)менниковий	I
		(п)рикметниковий	P
6	Група	а(т)рибутивний	T
		ат(р)ибутивно-присвійний	R
		відно(с)ний	S
		в(к)азівний	K
		(з)аперечний	Z
		з(в)оротний	V
		н(е)означений	E
		озна(ч)альний	C
		осо(б)овий	B
		ототожн(ю)вальний	J
		п(и)тальний	Y
		прис(в)ійний	W
		уза(г)альнювальний	H

Система тегів для класу сполучника (Р) має такий вигляд (табл. 8):

Таблиця 8. Система тегів для класу сполучника

Позиція в коді	**Атрибут**	**Значення**	**Тег**
0	С(п)олучник		P
1	Походження	п(е)рвинний	E
		в(т)оринний	T
2	Тип (для вторинних)	с(у)рядний	U
		під(р)ядний	R
3	Будова	(п)ростий	P
		с(к)ладний	K
		ск(л)адений	L
4	Розряд	(г)радаційний	H
		до(п)устовий	P
		(є)днальний	J
		з(і)ставний	I
		на(с)лідковий	S
		порі(в)няльний	W
		пояснювальний	B
		приєднувальний	C
		причиновий	E
		протиставний	F
		розді(л)овий	L
		умовний	Q
		цільовий	V
		часовий	Y
5	Група за способом вживання	од(и)ничні	Y
		по(в)торювані	W
		парні	B
6	Тип сполучуваних одиниць	слова	W
		речення	X

Система тегів для класу прийменника (J) має такий вигляд (табл. 9):

Таблиця 9. Система тегів для класу прийменика

Позиція в коді	Атрибут	Значення	Тег
0	При(й)менник		J
1	Походження	п(е)рвинний	E
		в(т)оринний	T
2	Тип	в(л)асне прийменник	L
		у знач(е)нні прийменника	E
		у (ф)ункції прийменника	F
		за аналогією	H
		оказіональний	J
3	Будова	(п)ростий	P
		с(к)ладний	K
		ск(л)адений	L
4	Сполучуваність	(н)азивний	N
		(р)одовий	R
		(д)авальний	D
		(з)нахідний	Z
		(о)рудний	O
		(м)ісцевий	M
		(к)личний	K
5	Значення	адресатне	C
		атрибутивне	F
		допусту	H
		інструментальне	K
		кількісно-атрибутивне	L
		локативне	M
		мети	Q
		міри і ступеня дії	S
		наслідку	1
		об’єктне	2
		порівняльне	3
		причини	4
		способу дії	5
		суб’єктне	6
		темпоральне	7
		умови	8

Система тегів для класу вигуку (W) має такий вигляд (табл. 10):

Таблиця 10. Система тегів для класу вигуку

Позиція в коді	Атрибут	Значення	Тег
0	(В)игук		W
1	Походження	п(е)рвинний	E
		в(т)оринний	T
2	Тип	віді(м)енниковий	M
		(в)іддієслівний	W
		відзайменни(к)овий	K
		відпр(и)слівниковий	Y
		іншомовний	B
		фразеологічний	Q
3	Будова	(п)ростий (однослівний)	P
		с(к)ладний (через дефіс)	K
		ск(л)їдений (з кількох слів)	L
4	Значення	емоційний	1
		наказовий	2
		спонукальний	3
		апелятивний	4

Система тегів для класу абревіатури (А) має такий вигляд (табл. 11):

Таблиця 11. Система тегів для класу абревіатури

Позиція в коді	Атрибут	Значення	Тег
0	(А)бревіатура		A
1	Тип	літерна	H
		звукова	Q
		складова	U
		мішана	W
2	Рід	(ч)оловічий	C
		(ж)іночий	Z
		(с)ередній	S
3	Число	(о)днина	O
		(м)ножина	M
4	Відмінок	(н)азивний	N
		(р)одовий	R
		(д)авальний	D
		(з)нахідний	Z
		(о)рудний	O
		(м)ісцевий	M
		(к)личний	K
5	Будова	подільна	U
		неподільна	X

Система тегів для класу R (решта) має такий вигляд (табл. 12):

Таблиця 12. Система тегів для класу R (решта)

Позиція в коді	Атрибут	Значення	Тег
0	(Р)ешта		R
1	Тип	скорочення	1
		звуконаслідування	2
		слово в іншомовному написанні	3
		номер списку	4
		лапки	5
		дужки	6

Система тегів для класу частка було наведено у [Данилюк 2013].

Отже, виділені 12 класів одиниць у корпусі текстів для вивчення граматичної службовості описані системою тегів, що характеризується використанням унікальних символів для кожної позиції, а підкласи, як то рід, число, відмінок тощо, закріплені за певним місцем у структурі тегу. У продовженні серії публікацій ми маємо намір детально описати металінгвістичну розмітку, а саме жанрову класифікацію текстів, включених до корпусу, механізм автоматичного морфологічного аналізу й алгоритм побудови вертикального файлу для корпусного менеджера Manatee.

Література

Безпояско 1993: Безпояско, О.К. Граматика української мови. Морфологія : Підручник [Текст] / О. К. Безпояско, К. Г. Городенська, В. М. Русанівський. – К. : Либідь, 1993. – 336 с.

Вихованець 2004: Вихованець, І.Р. Теоретична морфологія української мови : Академ. граматика укр. мови [Текст] / І. Р. Вихованець, К. Г. Городенська / За ред. І. Вихованця. – К. : Унів. вид-во «Пульсари», 2004. – 400 с.

Данилюк 2013: Данилюк, І.Г. Корпус текстів для вивчення граматичної службовості [Текст] // Лінгвістичні студії: Зб. наук. праць. Випуск 26 / Укл. : Анатолій Загнітко (наук. ред.) та ін. – Донецьк : ДонНУ, 2013. – С. 225-230.

Данилюк 2010: Данилюк, І.Г. Прикладна морфологія : Навчальний посібник [Текст] / І. Г. Данилюк. – Донецьк : ДонНУ, 2010. – 165 с. – ISBN 978-966-639-441-8

Данилюк 2006: Данилюк, І.Г. Синкретизм у системі частин мови [Текст] : автореф. дис. … к. філол. н. – Донецьк, 2006. – 20 с.

Захаров 2005: Захаров, В.П. Корпусная лингвистика : Учебно-метод. пособие [Текст] / В. П. Захаров. – СПб., 2005. – 48 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.

Курс сучасної української літературної мови 1951: Курс сучасної української літературної мови [Текст] / за ред. Л. А. Булаховського : В 2 т. – К. : Рад. шк., 1951 ; Т.1. – 519 с. ; Т.2. – 407 с.

Кучеренко 1961: Кучеренко, І.К. Теоретичні питання граматики української мови : Морфологія [Текст] / І. К. Кучеренко : В 2 ч. – К. : Вид-во Київ. ун-ту. 1961. – Ч.1. – 172 с. ; 1964. – Ч.2. – 159 с.

Леонова 1983: Леонова, М.В. Сучасна українська літературна мова : Морфологія [Текст] / М. В. Леонова. – К. : Вища шк., 1983. – 264 с.

Сучасна українська літературна мова 1969: Сучасна українська літературна мова : Морфологія [Текст] / За заг. ред. І. К.Білодіда. – К. : Наук. думка, 1969. – 583 с.

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.

В статье, которая является второй публикацией, посвященной описанию теоретических и практических основ создания корпусу текстов для изучения грамматической служебности в рамках научной темы кафедры украинского языка и прикладной лингвистики ДонНУ, описана заложенная в корпус классификация грамматических классов и подклассов. Подан принцип построения тегов, приведена их полная система для первой версии корпуса.

Ключевые слова: корпус текстов, части речи, грамматический класс, грамматический подкласс, тег.

The article, which is the second part in a series devoted to the description of the theoretical and practical principles of creating the text corpus for the study of syntactic grammar within the scientific theme of Department of Ukrainian Language and Applied Linguistics in Donetsk National University, describes grammatical classes and subclasses classification embedded in the corpus. Principles of tags, their complete system for the first version of the corpus are given.

Keywords: text corpus, parts of speech, grammatical class, grammatical subclass, tag.

Надійшла до редакції 21 серпня 2012 року.