Орися Демська-Кульчицька

ОДИН З АСПЕКТІВ МОРФОЛОГІЧНОЇ АНОТАЦІЇ (ДО ПРОБЛЕМИ ПОБУДОВИ ТЕҐА)

Сучасні корпуси текстів природних мов, щоб виконувати покладені на них функції, крім розмітки на рівні первинних даних, тобто формального опису структури документа, повинні бути анотованими, чи доповненими, лінгвістичною анотацією. Традиційно у корпусній лінгвістиці під лінгвістичною анотацією розуміють а) довільну лінгвістичну інформацію про лінгвально релевантні одиниці текстових даних, подану через формальний код, б) практику введення формалізованої лінгвістичної інформації в електронний текст і в) наявність такої інформації у тексті.

У принципі, лінгвістична анотація може забезпечувати будь-яку інформацію про мову тексту, але практика засвідчує, що в сучасних корпусних дослідженнях найчастіше йдеться про морфологічну та синтаксичну анотації, які, по суті, „визначають граматичні параметри тексту" [11]. Актуальність морфолого-синтаксичної анотації у корпусних дослідженнях зумовлена аплікативними характеристиками загальномовних корпусів, призначення яких слугували емпірикою для морфологічного, синтаксичного, лексичного, лексикографічного тощо опису мови.

Дві основні тенденції виявляються щодо морфолого-синтаксичної анотації у корпусній лінгвістиці: перша - до об'єднання морфологічної і синтаксичної анотації в морфолого-синтаксичн\. або граматичну, анотацію [12], де морфологічне маркування. як правило, розглядають як перший етап граматичної анотації, або як підґрунтя морфолого-синтаксичної анотації [9; 11]; друга - до диверенціації морфологічної та синтаксичної анотацій, які трактують як самостійні [10].

З лінгвістичного погляду, здійснення морфологічної анотації корпусних текстових даних попередньо передбачає:

а) побудову тегів, тобто спеціальних кодів, які через формальний запис експлікують граматичні значення слів, до яких вони приписані;

б) створення тегсету, чи набору формальних кодів з відповідною семантикою, засобами яких адекватно детерміновано для кожної лексичної одиниці тексту її відношення до морфологічної системи мови;

в) визначення принципів уведення тегів у текстові дані.

У нашій статті предметом аналізу є пункт а, тобто йдеться про принципи побудови морфологічних тегів та їхню семантизацію.

Морфологічні теґи є передусім засобом формалізації морфологічної інформації і призначені для програмного оброблення. Програмна орієнтованість накладає низку вимог на форму, зміст і засади побудови кожного конкретного теґа. Узагальнюючи, К. Лу виводить чотири основні критерії побудови тегів:

1) довжина: короткі символи зручніші для аналізу, ніж довгі, наприклад, DDI порівняно з SINGULAR_DETERMINER;

2) експліцитність: символи, які легко інтерпретувати і розуміти, є зручнішими для використання, ніж ті, які важко інтерпретувати і розуміти, наприклад: ІМЕННИК порівняно з IBC34;

3) аналітичність: символи, які підлягають декомпозиції на логічні складники, кращі, ніж ті, які не можна декомпонувати, наприклад, теґ NP1 з набору тегів BNC може бути розкладений на N = іменник (на відміну від V = дієслово), P = власна [назва] (vs N = загальна назва), 1 = однина (на відміну від 2 = множина).

Дотримання критерію аналітичності дозволяє здійснювати корпусні дослідження навіть за умови різного рівня їхньої деталізації. Так, символом N* можна задати усі іменники, і далі, деталізуючи: N*1- усі іменники в однині, NP* - усі іменники власні назви etc.

4) однозначність: у межах теґа унікальні символи співвідносяться з унікальними значеннями, наприклад: *іменник = N, прикметник = A, займенник - P* etc. і ці символи - *N*, *A, P* - на першій позиції у коді за жодних умов не можна використовувати на позначення якихось інших значень [11] у цій позиції.

Форма і семантика морфологічного теґа для конкретної мови залежить від системи лексико-граматичних розрядів цієї мови та набору релевантних морфологічних значень. А визначені у результаті морфологічного аналізу ознаки сегментних одиниць тексту „повинні бути інструментом дослідження зв'язку між лексикою і граматикою, між словником і використанням його у мовленні, між парадигматикою і синтагматикою" [4] і роль саме такого єднального елемента між зазначеними явищами виконують частини мови [3]. Тому в основу морфологічної класифікації як первинний детермінативний параметр кладуть значення лексико-граматичного розряду слова чи його частиномовну належність.

Системна специфіка конкретної мови, лінгвістична традиція частиномовної диференціації слів та програмна орієнтованість визначення набору лексико-граматичних розрядів слів для

морфологічної анотації детермінує різний набір граматичних класів як для різних мов, так і різний набір у межах однієї мови. Наприклад, для здійснення морфолого-синтаксичної анотації корпусу польської мови IPI PAN [12] виділено 21 лексико-граматичнии клас:

1) іменник (rzeczownik);
2) прикметник (przymiotnik);
3) ступеньований прислівник (przysłówek stopniowany);
4) приприйменниковий прислівник (przysłówek przyprzyimkowy);
5) числівник (liczebnik);
6) особовий займенник (zaimek osobowy);
7) форма дієслова неминулого часу (czasownik nieprzeszły);
8) форма майбутнього часу дієслова *бути* (czasownik przyszły *być) -,*
9) аглютинативна форма дієслова *бути* (aglutynat);
10) псевдодієприкметник і псевдодієприслівник (pseudoimiesłów);
11) форма наказового способу дієслова (rożkaznik);
12) безособова форма дієслова (bezosobnik);
13) інфінітив (bezokolicznik);
14) дієприслівник теперішнього часу (imiesłów przysłówkowy współczeny);
15) дієприслівник минулого часу (imiesłów przysłówkowy oprzedni);
16) герундій (odsłownik);
17) активний дієприкметник (imiesłów przymoitnikowy czynny);
18) пасивний дієприкметник (imiesłów przymoitnikowy bierny);
19) прийменник (przyimek);
20) сполучник (spójnik);
21) частко-прислівник (partykuło-przysłówek).

У запропонованій системі граматичні класи не відповідають класичному розподілові „частин мови" у польській граматиці, вони значно деталізованіші. Виділення детальних класів вмотивоване морфологічними категоріями, притаманними окремим флексемам [8], тому традиційні дієслівні лексеми поділено на кілька окремих одиниць: а) незмінні флексеми (безособова форма дієслова й інфінітив); б) змінні за родами і числами (дієприкметник і дієприслівник різного типу); в) змінні за числом і часом форми дієслова. Крім того, спеціально виділено два окремі класи лише для дієслова *бути.* Перший клас - для опису простого майбутнього часу, оскільки у польській мові це - єдине дієслово, яке має таку форм\. і другий клас - для аглютинативних форм з дієсловом *бити,* тобто (*е)т* -(*е)ś* і т.ін. [12].

Не відповідає традиційному і набір лексико-граматичних класів слів, виділених вченими Інституту мовознавства ім. О. О. Потебні з метою реалізації автоматичного аналізу наукового тексту [1].

Прийнята нами граматична класифікація є детальнішою щодо традиційного поділу слів за лексико-граматичною належністю у

російській мові:

1) іменник чоловічого роду (существительное мужского рода: *алгоритм., контроль);*
2) іменник жіночого роду (существительное женского рода: *задача, плата);*
3) іменник середнього роду (существительное среднего рода: *число, умножение);*
4) іменник без розрізнення роду (существительное без разграничения рода: *данные);*
5) повний прикметник (полное прилагательное: *прямоугольный, временный);*
6) повний дієприкметник (полное причастие: *включающие, используемые);*
7) дієслово (глагол: *рассматривается, содержит);*
8) прислівник (наречие: *например, возможно, быстро);*
9) короткий прикметник (краткое прилагательное: *возможен);*
10) короткий дієприкметник (краткое причастие: *показан, предоставлен);*
11) займенник-іменник (местоимение-существительное: *который, они);*
12) займенник-прикметник (местоимение-прилагательное: *этот, их);*
13) прийменник (предлог: *между, в, на, для);*
14) сполучник (союз: *и, но, если);*
15) дієприслівник (деепричастие: *применяя);*
16) частка (частица: *лишь, не);*
17) кількісний числівник (количественное числительное: *два, несколько);*
18) порядковий числівник (порядковое числительное: *второй, пятый)]*
19) абревіатура (аббревиатура: *ЭВМ, ФОРТРАН);*
20) власні назви (имена собственные: *Минск-32, Томсон).*

У аналізованій системі лексико-граматичних класів додатково виокремлено числа *(3, 17, 2),* символи (*N, k, t),* формули (*F=ma*), скорочення слів *(табл., рис.)* і скорочення словосполучень *(т.д., т.п.),* слова, записані літерами неросійського алфавіту (GO, *ТО),* але вони не інтерпретовані як лексико-граматичні розряди.

Такий підхід прийнято передусім „для відображення специфіки текстів рефератів на граматичному рівні" [1]. Інші відмінності, наявні у класифікації, зумовлені аплікативною специфікою анотованого матеріалу. Наприклад, виокремлені повні та короткі прикметники і дієприкметники, мотивовані завданням дослідження характеру предикативних та іменних зв'язків у тексті, а „ці форми прикметників і дієприкметників якраз і виконують у тексті поряд з дієсловом предикативну функцію" [1]. В окремий клас виділено дієприслівник, який у реченні виступає обставиною, на відміну від інших дієслівних

форм. А дослідження атрибутивних зв'язків ставить вимогу диференціації іменникових та прикметникових займенників, перші із яких є носіями іменникової функції у тексті, а другі - атрибутивної. І нарешті, в іменнику виділено чотири класи, визначальною ознакою яких є, по-перше, наявність/відсутність категорії роду, і, по-друге, триелементна родова парадигма: чоловічий, жіночий, середній рід іменників.

У 1996 р. у межах проекту EAGLES було розроблено *Рекомендації для морфолого-синтаксичної анотації корпусу довільної природної мови,* де, власне, запропоновано модель „базові <--> додаткові" граматичні класи і визначено 11 базових лексико-граматичних розрядів слів, плюс пунктуація і спеціальні одиниці, якими можуть бути, наприклад, формули (EAGLES 1996):

1) іменник (noun: N)
2) дієслово (verb: V)
3) прикметник (adjective: AJ)
4) займенник (pronoun/determiner: PD)
5) артикль (article: AT)
6) прислівник (adverb: AV)
7) адпозиція (adposition: AP)
8) сполучник (conjunction: C)
9) числівник (numeral: NU)
10) вигук (interjection: I)
11) спеціальна одиниця (unique/unassigned: U)
12) інші (residual: R)
13) пунктуація (punctuation: PU)

В українській мові класифікаційне виділення лексико-граматичних розрядів слів для автоматичного морфологічного аналізу українського наукового тексту здійснене в Київському національному університеті ім. Тараса Шевченка спільно з науковцями відділу структурно-математичної лінгвістики Інституту мовознавства ім. О.О.Потебні НАН України. В основу граматичної класифікації для цього автоматичного морфологічного аналізу покладено „прийняту в сучасних українських граматиках класифікацію слів, що передбачає диференціацію іменників, дієслів, прислівників, числівників, займенників, ад'єктивів, сполучників, прийменників, часток" [4]. Але, як зазначають автори проекту, властива українському науковому текстові омонімія граматичних класів „знайшла своє відображення в класифікації 13 диз'юнктивних класів слів" [4]:

1) іменник / прикметник: *дані, пряма, голосні;*
2) іменник / дієслово: *дати:*
3) іменник / прислівник: *с.и 1. правоа;*
4) іменник / займенник: *кому*
5) іменник / прийменник - *поза, шляхом;*
6) прислівник сполучник частка: *так, як;*
7) дієслово / прислівник *тримано, переконано;*

8) займенник / сполучник: *що*
9) прикметник / прислівник: *вороже;*
10) прикметник / числівник: *сьома;*
11) дієслово / прикметник: *ослабла, живе;*
12) займенник-іменник / займенник-прикметник: *це, все;*
13) прислівник / сполучник: *коли.*

Крім зазначених класів, для відображення специфіки українського наукового тексту, окремо виділено клас абревіатур, символів, чисел, формул і текстових скорочень.

Зауважимо, що не лише в корпусній лінгвістиці наявна така неоднозначність у виділенні лексико-граматичних розрядів слів. Про аналогічну ситуацію в класичному мовознавстві говорять, зокрема, О. Безпояско, К. Городенська і В.Русанівський: „Виділення частин мови за набором різнорідних ознак також не сприяло створенню єдиної і послідовної класифікації частин мови. До цього спричинилося те, що по-різному визначався набір класифікаційних ознак, а крім того, значна частина мовознавців надає перевагу одній з кількох ознак, що в свою чергу спричинилося до модифікації гетерогенної класифікації. Крім того, прибічники гетерогенної класифікації зіткнулися з тим, що в повному наборі різнорідні ознаки властиві лише чотирьом класам слів, тоді як решті слів притаманна якась одна з цих ознак, за якою вони й були виділені в традиційний граматиці в окремий клас слів. Ця непослідовність спричинилась до того, що традиційна система частин мови - це поєднання гетеро- і гомогенної класифікацій частин мови, тому що чотири частини мови - іменник, прикметник, дієслово і прислівник - виділені за набором трьох основних ознак (семантичної, морфологічної і синтаксичної), тоді як решта, зокрема, займенник, числівник, прийменник, сполучник, частка - за якоюсь однією ознакою із вказаних трьох" [2].

Зрозуміло, що сьогодні проаналізувати більшість із наявних лексико-граматичних класифікацій для морфологічної анотації вже неможливо через їх численність. Тому ми зупинилися на аналізі, по-перше, базових проектів, якими є проект EAGLES, по-друге, спільному проекті автоматизації аналізу наукового тексту, реалізованому українськими науковцями для російської та української мови, і, по-третє, морфолого-синтаксичній анотації для корпусу IPI PAN, збудованої за принципом стандартності анотаційних схем для корпусів природних мов.

Наступні структурні елементи коду виводять із детермінованих морфологічних значень для кожної з визначених частин мови. Репертуар цих значень можна формувати шляхом ідентифікації формальних ознак, які забезпечуватимуть співвіднесення одиниць рівня слова в електронному тексті з відповідним лексико-граматичним класом і відрізнятимуть їх передусім за формальним параметром від одиниць з іншим лексико-граматичним значенням.

Наприклад, для побудови морфологічних теґів у проекті корпусу IPI PAN визначено як „важливі для опису польських форм слів" [12]

категорії і до кожної з них формалізовано набір релевантних значень:

• **число** (liczba): однина (pojedyncza = sg: *oko),* множина (mnoga = pl: *oczy);*

•**відмінок** (przypadek): називний (mianownik = nom: *woda),* родовий (dopełniacz - gen: *wody),* давальний (celownik = dat: *wodzie),* знахідний (biernik = acc: wodę), орудний (narzędnik = inst: *woda),* місцевий (miejscownik = loc: *wodzie),* кличний (wołacz = voc: *wodo)]*

• **рід** (rodzaj): чоловічий (męski osobowy = ml: *papież, kto;* męski zwierzęcy = m2: *baranek;* męski rzeczowy = rn3: stół), жіночий (żeński = f: *stula),* середній (nijaki zbiorowy = n1: *dziecko;* nijaki zwykły = n2: *okno, co;* przymnogi osobowy = pl: *wujostwo;* przymnogi zwykły = p2: *skrzypce;* przymnogi opisowy = p3: *spodnie);*

• **особа** (osoba): перша (pierwsza = pri: *bredzę),* друга (druga = sec: *bredzisz),* третя (trzecia = ter: *bredzi);*

• **ступінь** (stopień): нульовий (równy = pos: *cudny),* вищий (wyzszy = comp: *cudniejszy),* найвищий (najwyższy = sup: *najcudniejszy);*

• **аспект** (aspekt): недоконаний (niedokonany. = imperf: *isc),* доконаний (dokonany = perf: *zazs);*

•**негація** (negacja): незанегована (niezanegowana = aff: *pisanie, czytanego),* занегована або заперечна форма (zanegowana = neg: *niepisanie, nieczytanego);*

•**депреціативність** (deprecjatywnośc): недепреціативна форма (niedprecjatywna = ndepr: *chłopi),* депреціативна (deprecjatywna = depr: *chłopy);*

• **акцентованість** (akcentowośc): акцентована, чи наголошена, форма (akcentowana = akc: *jego, niego),* ненаголошена (nieakcentowana = nakc: *go, -n);*

•**післяприйменниковість** (poprzyimkowosc): післяприйменникова форма (poprzyimkowa = praep: *niego, -ń),* непіслярийменникова форма (niepoprzyimkowa = npraep: *jego, go);*

• **акомодативність** (akomodacyjnośc): узгодження (uzgadniająca = congr: *dwaj),* керування (rządząca = rec: *dwóch, dwu);*

• **аглютинативність** (agłutynacyjność): неаглютинативна форма (nieagłutynacyjna = nagi: *niósł, dlaczego),* аглютинативна (agłutynacyjna = agi: *nioss-, dlaczego-);*

• **вокалічність** (wokaliczność): вокалічна форма (wokaiłczna = wok: *-em, -es, ze),* невокалічна (niewokałiczna =nwok: *-m, s', z).*

Визначивши для кожного морфологічного значення символьне позначення і об'єднавши їх у межах довільного тега для елементів рівня слова, отримаємо, наприклад, такі коди для слів польської мови з відповідними лексико-граматичною характеристикою і морфологічними значеннями:

n, sg, inst, ml - *іменник, и формі однини, орудного відмінка, чоловічого роду, особи;*

n, sg, inst, m2 - *іменник у ф< рмі однини, орудного відмінка, чоловічого роду, тварини:*

n, sg, inst, n2 - *іменник, у формі однини, орудного відмінка, звичайного середнього роду;*

n, sg, inst, ź - *іменник, у формі однини, орудного відмінка, жіночого роду [12].*

Академічна граматика сучасної української мови на підставі сукупності релевантних ознак морфологічного, синтаксичного і лексико-граматичного рівнів виділяє частини мови: іменник, прикметник, числівник, займенник, дієслово, прислівник, прийменник, частка, сполучник і вигук, тобто десять лексико-граматичних розрядів, з яких п'ять: іменник, прикметник, числівник, займенник, дієслово - повнозначні змінні частини мови, одна: прислівник - повнозначна незмінна, три: прийменник,частка, сполучник,- неповнозначні незмінні частини мови і еквіваленти висловів: вигук [6].

Академічну лексико-граматичну класифікацію слів сучасної української мови беремо за основу у визначенні набору частин мови для морфологічної анотації українського корпусного тексту. Але програмна орієнтованість анотації мотивує розширення академічного набору. Таке розширення зумовлене ще й доповненням вихідних принципів розподілу слів на частини мови формальним принципом і йдеться про наявність / відсутність експліцитних формальних ознак у слові, які відрізняють конкретну форму слова від інших його форм. Наприклад: формальною ознакою розрізнення особових дієслівних форм та інфінітива в сучасній українській мові є ідентифікація / неідентифікація у наборі літерних символів від відступу до відступу квазіморфеми *-ти/-ть,* а для порядкового і кількісного числівника - ідентифікація/неідентифікація квазіморфем *-ий, -а, -е,* у лексикографічній формі тощо.

Найбільшої модифікації у нашому підході зазнало дієслово. Так, у межах дієслова окремо виділено: а) інфінітив, б) особові форми, чи так зване лексичне дієслово, в) предикатив, г) дієприкметник і ґ) дієприслівник. Першопричиною такого виділення були формальні ознаки в межах буквеного ряду від відступу до відступу: а) для інфінітива квазіморфема *ти / -ть: писати, говорить, бути, думать*; б) для предикатива - *-ано, -єно: роблено, писано;* в) для дієприкметника - флективна парадигма прикметникового типу: *будований, копійованого, мальованій*; г) для дієприслівника - *-ши, -чи: писавши, думаючи, страшачи* і ґ) усі решта - особові форми: *сміятимемося, скажемо, дивлюся.* Крім того, ідентифіковано допоміжні дієслова: *бути, могти, хотіти;* морфему-формант наказового способу: *хай, нехай;* показник умовного способу: *б, би.*

Для числівника здійснено таку диференціацію: розмежовано порядкові та непорядкові числівники на підставі формальної флективної відмінності. Формально порядковий числівник або прикметниковий числівник щодо непорядкового має, по-перше, інший набір морфологічних значень, і, по-друге, відмінну флективну парадигму, яка, по суті, є прикметниковою, наприклад: *веселий* -

перший, зеленої - десятої, квадратне - п'ятнадцяте. Додатково виділено й власне числівник: *сім, сто дев'яносто;* числові назви: *багато, кільканадцять* і слова, за допомогою яких утворені форми дробових числівників: *ціла, соті, тисячних.*

Таким чином, для морфологічної анотаційної схеми Українського національного корпусу маємо шістнадцять лексико-граматичних розрядів слів у сучасній українській мові, для яких передбачено таке символьне позначення у тегах:

1. Іменник = N (noun);
2. Прикметник = A (adjective);
3. Порядковий / прикметниковий числівник = L (ordinal numeral);
4. Числівник: власне числівник = M (numeral), числові назви = MN (numeral name);
5. Займенник = P (pronoun);
6. Дієслово = V (verb);
7. Інфінітив = F (infinitive);
8. Предикатив = D (predicative);
9. Дієприкметник = T (adjectival participle);
10. Дієприслівник = B (adverbial participle);
11. Прислівник = R (adverb);
12. Прийменник = S (preposition);
13. Сполучник = C (conjunction);
14. Частка = Q (particle);
15. Вигук = I (interjection).

У розглянутих підходах формалізоване лексико-граматичне значення інтерпретовано як найважливіший параметр і символ на його позначення стоїть на першій позиції у тегах. Наприклад, в уривку з Біблії:

/ На початку Бог створив небо та землю.

2 А земля була пуста та порожня, і темрява була над безоднею, і Дух Божий ширяв над поверхнею води.

у тегах, приписаних до слів: *початок, Бог, небо, землю, темрява, безоднею, Дух Божий, поверхнею* і *води* першим символом повинна бути позначка N = noun, тобто іменник.

1 На початку_N Бог_N створив иебо_N та землю .

2 А земля_N була пуста та порожня, і темрява_N була над безоднею_N, і Дух_Божий_N ширяв над поверхнею_N води_N.

І аналогічно до дієслів V = verb, прикметників A = adjective:

1 На початку Бог_N створив_V небо_N та землю_N.

2 А земля_N була_V пуста_A та порожня_A, і темрява_N була_V над безоднею_N, і Дух_Божий_N ширяв_V над поверхнеюN води_N.

Наступним кроком після здійснення лексико-граматичного перерозподілу стала верифікація кожної із виділених, як академічних, так і додаткових, частин мови щодо репертуару морфологічних значень.

Загалом, морфологічні та синтаксичні значення прийнято розглядати в морфології як граматичні категорії, які поділяють / не поділяють на морфологічні та синтаксичні. Серед морфологічних категорій виділяють, наприклад, граматичну категорію виду, стану, часу, способу, особи, роду, числа, відмінка, а „послідовність вираження цих категорій характеризують цілі граматичні класи слів, тобто частини мови" [5].

Схема морфологічної анотації українського корпусного тексту детермінувала відбір лише морфологічних значень, і лише тих, які, по-перше, мають формальну експлікацію у морфемній будові слова, і, по-друге, є релевантними для машинної моделі української мови. Такими морфологічними значеннями стали: а) рід, б) число, в) відмінок, г) особа, ґ) час, д) аспект, е) стан і є) спосіб, кожне з яких може набувати лише наступних значень:

рід: чоловічий жіночий середній;

число: однина <-> множина <-> pluralia tantum <->двоїна;

відмінок: називний родовий <-> давальний знахідний орудний місцевий кличний;

особа: 1 <->2 <->3;

час: теперішній <-> минулий майбутній <-> давноминулий;

аспект: доконаний недоконаний <-> двовидова форма;

стан: активний пасивний;

спосіб: дійсний умовний <-> наказовий.

Згідно з методикою побудови морфологічних тегів, яку ми обрали, кожне із детермінованих значень забезпечено відповідною символом-міткою:

чоловічий = **ш** (masculine);
жіночий = **f** (feminine);
середній = n (neutral);
однина = **s** (singular);
множина = **p** (plural);
pluralia tantum = **t;**
двоїна = **d** (dual);
називний = n (nominative);
родовий = **g** (genitive);
давальний = **d** (dative);
знахідний = **a** (accusative);
орудний = i (instrumental);
місцевий = 1 (locative);
кличний = **v** (vocative);
особа: 1<->2<->3 - позначена відповідно цифрами 1,2,3;
теперішній = p (present);
минулий = **t** (past);
майбутній = **u** (future);
давноминулий = c (pluperfect);
доконаний = r (perfect);

недоконаний = **i** (imperfect);
двовидова форма = **d** (double form);
активний = **a** (active);
пасивний = **v** (passive);
дійсний = **a** (indicative);
умовний = **j** (conditional);
наказовий = **m** (imperative).

Об'єднавши символи визначених розрядів слів української мови та їхніх релевантних морфологічних значень, було сформовано конкретні морфологічні теґи та їхню семантику і, відповідно, весь кодовий набір, чи теґсет, для української мови, засобами якого можна здійснити морфологічну анотацію українських корпусних текстів. Наприклад:

для іменника:

NMSN - „іменник чоловічого роду однини, у формі Н. відмінка" *вельможа, суддя, темп, батько, дощ, лікар, вчитель, звичай, вітрище, дідисько;*

NMSG - „іменник чоловічого роду однини, у формі Р. відмінка" *вельможі, судді, темпу, батька, дощу, лікаря, вчителя, звичаю, вітрища, дідиська;*

NMSD - „іменник чоловічого роду однини, у формі Д. відмінка" *вельможі, судді, темпові, батькові, дощу, лікареві, вчителю, звичаю, вітрищу, дідиськові;*

NMSA - „іменник чоловічого роду однини, у формі З. відмінка" *вельможу, суддю, темп, батька, дощ, лікаря, вчителя, звичай, вітрище, дідиська;*

для прикметника:

AMSD - „прикметник чоловічого роду однини, в Д. відмінку, нульовий ступінь порівняння" *веселому, народному, батьковому, радому, синьому, колишньому;*

AMSA - „прикметник чоловічого роду однини, в З. відмінку, нульовий ступінь порівняння" *веселого, народний, батьків, радого, синього, колишній;*

для порядкового числівника:

LFSD - „порядковий числівник жіночого роду, однини в Д. відмінку" *десятій;*

LFSA - „порядковий числівник жіночого роду, однини в З. відмінку" *десяту,*

LFSI - „порядковий числівник жіночого роду, однини в О. відмінку" *десятою;*

MKA+WMF+MKA+WMF - „дробовий числівник у З. відмінку зі словами *цілі і десяті" десять цілих сім десятих;*

для займенника:

PMSL - „займенник чоловічого роду однини в М. відмінку" ... *моєму, всякому, всьому;*

PFSN - „займенник жіночого роду однини в Н. відмінку" *ця, абияка;*

PFSG - „займенник жіночого роду однини в Р. відмінку" *цієї, абиякої;*

для дієслова:

VOTOSM - „допоміжне дієслово, минулого часу, без експлікації особового значення, однина, чоловічого роду" *був;*

VOTOSF - „допоміжне дієслово, минулого часу, без експлікації особового значення, однина, жіночого роду" *була;*

VARU1S0 - „дієслово дійсного способу доконаного виду, майбутнього часу, першої особи однини, без родового значення" *спечу, збудуєш;*

VARU2S0 - „дієслово дійсного способу доконаного виду, майбутнього часу, другої особи однини, без родового значення" *спечеш, збудуєш;*

для дієприкметника:

TVTRFLS - „дієприкметник, пасивного стану, минулого часу, доконаного виду, жіночого роду, в М. відмінку однини" ... *закритій, ... виготовленій;*

TVTRNNS - „дієприкметник, пасивного стану, минулого часу, доконаного виду, середнього роду, в Н. відмінку однини" *закрите, виготовлене;*

TVTRNGS - „дієприкметник, пасивного стану, минулого часу, доконаного виду, середнього роду, в Р. відмінку однини" *закритого, виготовленого.*

Отже, сучасні корпуси текстів природних мов, щоб слугувати реальною базою для лінгвістичних досліджень із найрізноманітнішими рівнями пошуку та екстрагування мовного матеріалу, повинні бути лінгвістично анотованими. Найпоширенішим типом лінгвістичної анотації є морфологічна анотація, що, відповідно, поставило вимогу, за створення корпусу текстів української мови, розв'язати завдання морфологічного анотування корпусу текстів української мови, починаючи із побудови морфологічних тегів та їхньої семантизації. А найважливішими параметрами якісних морфологічних тегів є експліцитність, аналітичність та однозначність, чого дотримано у морфологічному тегсеті для морфологічної анотації Українського національного корпусу.

1. Автоматизация анализа научного текста. – К., 1984
2. *Безпояско О. К., Городенська К. Г., Русанівський В. М.* Граматика української мови: Морфологія. – К., 1993.
3. Сучасна українська літературна мова: Морфологія / Білодід І.К. – К., 1969.
4. *Лопатин В.В.* Грамматическая категория // Лингвистический энциклопедический словарь. – Москва, 1990.

5. *Шаров С. А.* Большой Корпус русского языка. –www.bokrcorpora.narod.ru, 2002; *Kennedy G.* Introduction to Corpus Linguistics. – London – New-York: Longman, 1998.
6. *Грязнухіна Т., Нікуліна М.* Система автоматичного морфологічного аналізу української мови // Проблеми українізації комп'ютерів: Матеріали 2-ї міжнародної конференції. – Львів – К., 1992.
7. *Белоногов Г.Г., Новоселов А.П.* Автоматизация процессов накопления, поиска и обобщения информации. – М., 1979.
8. *Bien J.S.* Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji // Rozprawy Uniwersytetu Warszawskiego: – Warszawa, 1991.
9. *Church K.W., Hanks P.* Word association norms, mutual information and lexicography // Computational Linguistics. – 1990. – Vol. 16.
10. *Lu X.* POS Tagging: An Overview. – Oxford: Oxford University Press, 2002.
11. *Woliński M., Przepiórkowski A.* Projekt anotaciji morfosyntaksycznej korpusu języka polskiego. – Warszawa: IPI PAN. 2001; EAGLES: Recommendations for the Morphosyntactic Annotation of Corpora. – www.ilc.pi.cnr.it/EAGLES 96, 1996.

Мовна мозаїка

КІЛЬКА СЛІВ ЩОДО БУРЯКА ТА БУРЯКІВ

У деяких словниках, зокрема орфографічних, розрізнено дві флексії родового відмінка однини в іменникові **буряк**. Флексія **-а** (*буряка́*) стосується тут одного кореня цієї рослини, а флексія *-у* (*буряку́*) виражає значення збірності. Чи правильне таке розрізнення?

Ні, не правильне. Закінчення **-у** іменникові **буряк** не властиве. Він уживаний в однині тільки із закінченням **-а** й означає один корінь цієї городньої рослини, наприклад: *Коло кожного рядка молодиць стояв "наставник". Він пантрував, щоб копальниці не лінувались та добре обгортали землею кожний* **буряк** (М.Коцюбинський); *Іван вирвав* **буряк** *і, пожбурнувши ним, влучив її* [Олександру] *прямо в хустку, аж очіпок з'їхав трохи набік* (М.Коцюбинський); *Молодиця почервоніла, наче* **буряк**, *гнівні іскорки заграли у запалих очах* (Панас Мирний).

Складніші значеннєві вияви має іменник **буряк** у множині. Множина може стосуватися двох, трьох, чотирьох і більше коренів рослини і співвідноситься з одниною цього слова (*буря́к – бурякú*), наприклад: *В їхніх торбах знайшлося навіть кілька цукрових* **буряків.**. (О.Гончар). А інша форма множини в іменнику **буряки** передає значення збірності: *Перед хатиною, на маленьких грядочках, росла картопля,* **буряки** *та цибуля* (І.Нечуй-Левицький); *Треба* **буряків**, *вівса, сіна, висівок, а де його візьмеш?* (Г.Тютюнник); *А мати із сапою день при дні На* **буряках** .. (М.Вінграновський); *У руці тримала чорну господарську сумку, з якої виглядає морква й* **буряки** (Є.Гуцало). Отже, в цьому іменникові збірне значення передає форма множини – **бурякú**, а не однини – **буряку́**.

Іван Вихованець