Морфосинтаксичне тагування польсько-українського паралельного корпусу (PolUKR)

Наталія Коциба

Варшавський університет (Польща)

The paper deals with theoretical grammatical issues of morphological annotation for PolUKR[1] – Polish-Ukrainian parallel corpus. Contemporary academic grammars' approaches for the respective languages, as well as existing tagsets for monolingual corpora and the present standard for multilingual common morphological tagset, created within the MULTEXT-East international project, are compared, and a proposal of a common Polish-Ukrainian tagset is presented. In particular, selected grammatical categories like participles, pronouns, predicatives, adjectives, are considered. Types of formal presentation of morphological annotation in corpora are discussed, and reasons for the preferred type are given.

У статті йдеться про особливості спільного формату граматичного знакування для польсько-українського паралельного корпусу (PolUKR). Розглядаються та порівнюються підходи до граматичного опису обох мов як у сучасних академічних граматиках, так і у відповідних одномовних корпусах та у багатомовному паралельному проекті MULTEXT-East, спільний тагсет (набір граматичний позначок) котрого вважається претендентом на міжнародний стандарт. Запропоновано коротку класифікацію тагсетів залежно від способу представлення, рівня ґрануляції та загніжденості граматичної інформації. Розглянуто, зокрема, категорії дієприслівників, займенників, предикативів та прикметників. Представлено пропозицію спільного польсько-українського тагсету.

Ключові слова: паралельний корпус, PolUKR, тагсет, таг, морфосинтаксичне знакування, тагування, польська мова, українська мова.

Key words: tag, tagset, paralel corpus, PolUKR, morphosyntactic tagging, Polish, Ukrainian.

1. Вступ

Польсько-український паралельний корпус може бути застосований у вигляді бази для машинного перекладу, як матеріал для вивчення української та польської мов, для поповнення польсько-українського словника, матеріал для навчання обидвох мов та для розроблення контрастивної граматики цих мов. Для більшості з цих завдань граматична інформація про слова у текстах є необхідною. Оскільки корпус двомовний, а кожна з цих мов має свої традиції граматичного запису, як у традиційній, так і в корпусній лінгвістиці, то з'являється питання обсягу та формату запису граматичної інформації у паралельному корпусі. Звичайно, можна було б вибрати простіший шлях і

[1] Дана, розширена версія корпусу створюється за підтримки гранту Міністерства Науки і Вищої Освіти Польщі (MNiSW) NN 104 0403 33 у 2007-2009 роках.

розставити позначки (теги) у текстах кожної з мов у форматі, який використовується у провідних мовних проектах кожної з країн.[2] Такий спосіб вибрано для багатомовного паралельного корпусу слов'янських мов, відомого як Parasol або Ренсбурський. Використання більшої кількості мов надзвичайно ускладнює завдання, тому такий підхід є виправданим для багатомовного корпусу. Однак те ж саме завдання для двох мов виконати легше, та й використання спільного граматичного поняттєвого апарату має незаперечні переваги, оскільки дає можливість фактично порівнювати дві мови. Це набуває особливо важливого значення при контрастивних дослідженнях граматики. Тому ми ставимо за мету проаналізувати системи запису граматичної інформації для обидвох мов та окреслити спільний набір і формат запису граматичних позначок для використання у паралельному корпусі.

2. Джерела морфосинтаксичної інформації для польської мови

Польські корпусні традиції сягають 1960-тих років минулого століття, коли з'явився перший, т.зв. частотний корпус польської мови. У той же час велася робота над систематичним описом граматичної інформації на рівні окремих словоформ. Перші спроби системного запису граматичної інформації словоформ пов'язані з іменем Яна Токарського, котрому належить концепція словника *a tergo* для польської мови, та котрий зібрав основну частину матеріалу для цього словника. Словоформи згруповані в ньому за закінченнями, до кожної з груп додано всі можливі морфосинтаксичні інтерпретації. Це дає можливість як лематизації словоформ у текстах, так і додавання до них граматичної інформації. Паперова версія словника, відредагована, допрацьована та вперше опублікована Зигмунтом Салоні у 1996 році, стала підставою для комп'ютерного морфологічного аналізатора польської мови Марціна Воліньського, що є також основним джерелом граматичної інформації у пакеті TaKIPI для тагування польських текстів. Програмний пакет включає модулі структурного поділу текстів, лематизації, морфологічного аналізу та дезамбіґуації (граматичного уоднозначнення). Словник тагера налічує біля 200 тисяч слів. Першим великим проектом, для якого використовувався цей тагер, був корпус польської мови IPIPAN. Тагсет, чи набір граматичних позначок, набув теперішнього вигляду саме під час праць над цим корпусом у 2000-2004 роках. Організація граматичної інформації у вихідному форматі TaKIPI спирається не тільки на польську граматичну традицію, але також на практичні потреби та вимоги морфологічного аналізу слів із перспективою синтаксичного аналізу (саме тому автори корпусу і тагсету вживають терміни *морфосинтаксичний аналіз* та *морфосинтаксичний тагсет*). Ще однією особливістю польського тагсету є те, що він базується не на категорії частини мови у її традиційному розумінні, як вихідній, а на категорії флексеми, що визначається строгим набором спільної граматичної характеристики слів, що належать до неї. Нижче подаємо детальніший опис поняття флексеми.

---

[2] Для польської мови найбільшим публічно доступним є Korpus Języka Polskiego IPI PAN (далі вживатимемо скорочення Корпус IPI), http://korpus.pl, для української мови це корпус, що розробляється Українським мовно-інформаційним фондом НАНУ, далі Корпус УМІФ. У цій статті ми керуємося описом тагсету з [Широков, 2005].

3. Джерела морфосинтаксичної інформації для української мови

Основним джерелом морфосинтаксичної інформації для українських текстів є граматичний словник української мови (далі скорочено УГС) [Шевченко та ін., 2005, Шевченко 1996], створений Ігорем Шевченком в Українському мовно-інформаційному фонді НАНУ в 1990-тих роках минулого століття. Словник налічує біля 250 тис. слів. Записана в ньому інформація дозволяє проводити лематизацію та морфологічний аналіз українських текстів, але можливості морфологічного уоднозначнення як наразі є досить малі і ця функція вимагає додаткової праці. На відміну від польського розв'язку, УГС загалом віддзеркалює українську граматичну традицію. Слова у ньому погруповано у парадигматичні класи, що характеризуються спільним набором параметрів опису: починаючи від частини мови та типу основи і відмінювання, закінчуючи особливостями форм відмінків, типу денотата чи акцентуації та дефектності парадигми (всього 17 параметрів). Іншими словами, найменше відхилення у наборі значень параметрів веде до виділення нового парадигматичного класу, що пояснює їх велику кількість – таких класів виділено понад 2500. Теоретично ці класи можуть групуватися у ієрархії навколо одного або кількох параметрів, але практично ця властивість не є реалізована в УГС на формальному рівні. Деякий рівень оптималізації та умовності запису, що веде за собою відхід від традиційного філологічного трактування категорій, можна спостерігати також і в УГС. Наприклад, створено штучні частини мови «іменники жіночого роду», «прізвища»; прикметники та прислівники вищого та найвищого ступенів трактуються як окремі лексеми, що нівелює категорію ступеня як граматичну, і ін., але виведення граматичної інформації відповідає традиційному описові. Так само, як і в ситуації з польською мовою, УГС включає окрім морфологічної ще й частково синтаксичну інформацію. Нотується інформація про керування відмінками для прийменників, та й сам поділ на службові частини мови базується на синтаксичних відмінностях. Тому термін *морфосинтаксичний* можемо вживати в стосунку до обох частин паралельного корпусу.

Джерела граматичної інформації для обох мов використовують концептуально різні підходи до екстракції цієї інформації та її організації і запису. Спостерігається також різний ступінь ґрануляції інформації. Як наслідок неоднакового роздріблення, притаманні обидвом мовам явища описані по-різному. У кожному випадку такого розходження потрібно було вирішити, чи варто кодувати явища для обох мов, чи ні. Другий варіант небезпечний втратою можливо цінної інформації. У першому ж випадку інформацію, котрої бракує для однієї з мов, треба було вводити додатково. Часто застосування такої оккамової бритви для паралельного тагсету обґрунтовувалося не теоретичною доцільністю, а практичними потребами і можливостями.

4. Види тагсетів залежно від форми презентації граматичної інформації

Варто також згадати про способи кодування морфосинтаксичної інформації. Вони відрізняються за рядом параметрів, таких як: рівень кодування (словоформа чи значення притаманних їй категорій), фіксованість порядку кодових символів та

пов’язана з цим вимога їх унікальності, принципи та глибина категоризації граматичної інформації. Розглянемо кожен з цих параметрів по порядку.

4.1 Рівень кодування

Залежно від **рівня кодування** тагсети можна поділити на **символьні**, у котрих граматична характеристика цілої словоформи передається одним кодом, та **ланцюжкові**, в котрих кожна категорія, її атрибут і значення цього атрибута мають свій унікальний код, а морфосинтаксична характеристика кожної словоформи представлена послідовністю кодів значень. Прикладами символьного кодування є тагсети Британського національного корпусу та Корпусу української мови УМІФ. Тагсет УМІФ складається з 384 двосимвольних кодів, перша літера яких вказує на частину мови або підкатегорію (різними є, наприклад перші символи кодів для дієслів без або з постфіксом *-ся*, умовно позначених як активний та пасивний стан).

Приклади граматичних кодів, що використовуються в Корпусі УМІФ:

Граматичне значення	**Код**	**Приклад**
Дієслово, інфінітив, доконаний вид, активний стан	VA	*прочитати*
Дієприкметник, чоловічий рід, однина, називний відмінок, доконаний вид, минулий час, активний	BA	*зрослий*
Невідмінюваний прикметник	AZ	*ультра*
Іменник загальний, жіночий рід, однина, давальний відмінок	FC	*квітці*
Предикатив (присудкове слово)	X0	*слід*

Приклад українського нетагованого тексту:

«Львів розташований на етнічних українських землях і є одним з головних нервових вузлів українського народу, найважливішим клапаном його серця, вічним збудником честолюбства, гордості й потягу до волі.»

Цей же текст з усіма можливими морфосинтаксичними інтерпретаціями та лемами:

Львів<JDJAJIJK><Львів 0|Львів 0|Лев 1|Лев 1|> **розташований<BDBAV?>**<розташований 0|розташований 0|розташувати 0|> **на<N0N0Z0PF>**<на 4|на 3|на 2|на 1|> **етнічних<AVATAX>**<етнічний 0|етнічний 0|етнічний 0|> **українських<AVATAXJIGIJKGKJMGM>**<український 0|український

0|український 0|Український 0|Український 0|Український 0|> **землях<FM>**<земля 2|> **і<SSSCN0Z0>**<і 1|і 3|і 2|> **є<UPUOUNUKUMUL>**<бути 0|бути 0|бути 0|бути 0|бути 0|бути 0|> **одним<HUHQHERQRERU>**<один 0|один 0|один 0|оден 0|оден 0|оден 0|> **з<PE>**<з 0|> **головних<AVATAX>**<головний 0|головний 0|головний 0|> **нервових<AVATAX>**<нервовий 0|нервовий 0|нервовий 0|> **вузлів<MIMI>**<вузол 2|вузол 1|> **українського<ANADABJDJBKB>**<український 0|український 0|український 0|Український 0|Український 0|Українське 0|> **народу<MBMCMBMC>**<народ 0|народ 0|нарід 0|нарід 0|>, **найважливішим<AQAEAU>**<найважливіший 0|найважливіший 0|найважливіший 0|> **клапаном<ME>**<клапан 0|> **його<FGODOBODOB>**<йога 0|воно 0|воно 0|він 0|він 0|> **серця<NKNHNBNN>**<серце 0|серце 0|серце 0|серце 0|>, **вічним<AQAEAU>**<вічний 0|вічний 0|вічний 0|> **збудником<MEME>**<збудник 1|збудник 2|> **честолюбства<NB>**<честолюбство 0|>, **гордості<FCFBFF>**<гордість 0|гордість 0|гордість 0|> **й<SSSCZ0>**<й 1|й 2|> **потягу<MFMCMGMBMCMFMGFDGD>**<потяг 2|потяг 2|потяг 2|потяг 1|потяг 1|потяг 1|потяг 1|потяга 0|Потяга 0|> **до<NGNFNENDNCNBNANHNINJNKNLNMNNPB>**<до 2|до 2|до 2|до 2|до 2|до 2|до 2|до 2|до 2|до 2|до 2|до 2|до 2|до 2|до 2|до 2|до 2|до 2|до 1|> **волі<UOFCFBFFGCGBGFGFGCGBGHGNNFMF>**<воліти 0|воля 0|воля 0|воля 0|Воля 2|Воля 2|Воля 2|Воля 1|Воля 1|Воля 1|Воля 1|Воля 1|воло 0|віл 0|>.

Словоформи з тексту зазначено жирним шрифтом, за ними у трикутних дужках ідуть набори можливих граматичних інтерпретацій, що складаються з двосимвольних кодів, далі подано леми до кожної з інтерпретацій із зазначенням номера омоніма, якого стосується ця інтерпретація.

Безперечною перевагою символьного запису є його компактність. Доцільність саме такого кодування була пов’язана свого часу із обмеженнями, які накладалися можливостями комп’ютерної техніки у минулому столітті. Символьний запис вимагає мало машинної пам’яті, але не є зручним для користування людиною.

Ланцюжкові таги загалом користуються більшою популярністю, ніж символьні. Використовуються в таких проектах, як MULTEXT/MULTEXT-East, Корпус польської мови IPI PAN, Чеський національний корпус, Національний корпус російської мови.

Приклад польського нетагованого тексту:

«W dzisiejszym posiedzeniu komisji uczestniczy ekspert komisji pan profesor Jan Gajewski.»

Фрагмент файлу ХМЛ, де подано всі можливі інтерпретації перших чотирьох слів вказаного польського тексту:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE cesAna SYSTEM "xcesAnaIPI.dtd">
…
<group id="a11" rule="Uncertain: Dobre PrepNG (z postmodyfikatorem dopełniaczowym) na koncu zdania lub nawiasu, lub przed czasownikiem" synh="a1" semh="a3" type="PrepNG">
<tok id="a1">
<orth>W</orth>
<lex disamb_sh="0"><base>w</base><ctag>prep:acc:nwok</ctag></lex>
<lex disamb="1"><base>w</base><ctag>prep:loc:nwok</ctag></lex>
</tok>
<tok id="a2">
<orth>dzisiejszym</orth>
<lex disamb_sh="0"><base>dzisiejszy</base><ctag>adj:sg:inst:m1:pos</ctag></lex>
<lex disamb_sh="0"><base>dzisiejszy</base><ctag>adj:sg:inst:m2:pos</ctag></lex>
<lex disamb_sh="0"><base>dzisiejszy</base><ctag>adj:sg:inst:m3:pos</ctag></lex>
```

<lex disamb_sh="0"><base>dzisiejszy</base><ctag>adj:sg:inst:n:pos</ctag></lex>
<lex disamb_sh="0"><base>dzisiejszy</base><ctag>adj:sg:loc:m1:pos</ctag></lex>
<lex disamb_sh="0"><base>dzisiejszy</base><ctag>adj:sg:loc:m2:pos</ctag></lex>
<lex disamb_sh="0"><base>dzisiejszy</base><ctag>adj:sg:loc:m3:pos</ctag></lex>
<lex disamb="1"><base>dzisiejszy</base><ctag>adj:sg:loc:n:pos</ctag></lex>
<lex disamb_sh="0"><base>dzisiejszy</base><ctag>adj:pl:dat:m1:pos</ctag></lex>
<lex disamb_sh="0"><base>dzisiejszy</base><ctag>adj:pl:dat:m2:pos</ctag></lex>
<lex disamb_sh="0"><base>dzisiejszy</base><ctag>adj:pl:dat:m3:pos</ctag></lex>
<lex disamb_sh="0"><base>dzisiejszy</base><ctag>adj:pl:dat:f:pos</ctag></lex>
<lex disamb_sh="0"><base>dzisiejszy</base><ctag>adj:pl:dat:n:pos</ctag></lex>
</tok>
<tok id="a3">
<orth>posiedzeniu</orth>
<lex disamb_sh="0"><base>posiedzenie</base><ctag>subst:sg:dat:n</ctag></lex>
<lex disamb="1"><base>posiedzenie</base><ctag>subst:sg:loc:n</ctag></lex>
<lex disamb_sh="0"><base>posiedzieć</base><ctag>ger:sg:dat:n:perf:aff</ctag></lex>
<lex disamb="1"><base>posiedzieć</base><ctag>ger:sg:loc:n:perf:aff</ctag></lex>
</tok>
<tok id="a4">
<orth>komisji</orth>
<lex disamb="1"><base>komisja</base><ctag>subst:sg:gen:f</ctag></lex>
<lex disamb_sh="0"><base>komisja</base><ctag>subst:sg:dat:f</ctag></lex>
<lex disamb_sh="0"><base>komisja</base><ctag>subst:sg:loc:f</ctag></lex>
<lex><base>komisja</base><ctag>subst:pl:gen:f</ctag></lex>
</tok>
</group>

Словоформи з тексту позначені тагом <orth/>, леми – <base/>, граматична інтерпретація – тагом <ctag/>, вибрана як неправильна інтерпретація – <lex disamb_sh="0"/>, як правильна – <lex disamb="1"/>.
Незважаючи на, здавалося б, громіздкий запис, ланцюжкові таги можуть бути економніші при пошуку, ніж символи, напр. коли обмежуємо пошук на значення атрибуту, що властивий кільком частинам мови (відмінок, рід для іменника і прикметника). Перевагою цього формату є прозорість та мнемонічність. З ним легше працювати дослідникам-лінгвістам, зникає потреба написання спеціальних програм-конверторів.

4.2 Фіксованість позиції елементів у ланцюжковому тазі

Проблема фіксованості позиції граматичної інформації у коді торкається тільки ланцюжкових тагсетів. Відносно цього параметру тагсети можна поділити на **строго позиційні**, де кожна категорія (виражена її вартістю в тазі) має фіксовану позицію у ланцюжку, тоді вартості різних категорій можуть мати ту саму нотацію (t,n,y,p, 1, 2, 3) та **універсальні**, де вартості кожного атрибуту мають унікальний код, що не перекривається з іншими у цьому ж тагсеті. Прикладом позиційного є тагсет, що використовується у Чеському національному корпусі та у міжнародному багатомовному проекті MULTEXT/MULTEX-East.

Приклад позиційних тагів та їх інтерпретації з Чеського національного корпусу:

volen: VsYS---XX-AP--- verb, passive participle, masculine, singular, any person, any tense, positive, passive

hraniční: AAIS4----1A---- standard adjective, masc. inanimate, singular, accusative, positive

Щоб задати пошук по одному з атрибутів, треба записати цілий ланцюжок. На противагу тагам зі строго фіксованим порядком, універсальні таги є більш гнучкими, пошук за параметром родового відмінку незалежно від частини мови у Корпусі IPI задаємо фрагментом ланцюжка: „.*gen.*”, „case=„gen””.

4.3 Принципи та глибина категоризації граматичної інформації

Передовсім тут мова йде про кореневу категоризацію слів та словоформ. Звісно, що всі словоформи зводяться спочатку до лем, котрим приписується певна базова категорія. Найчастіше це є частина мови у її традиційному, арістотелівському розумінні. Навіть це базове розрізнення не є таким однозначним, як могло би здаватися. Залежно від провідної традиції країни можуть виділятися додаткові частини мови, як, наприклад, предикативи; різним є частиномовний (у корпусному розумінні) статус скорочень, пунктуації та немовних або змішаних знаків, як то різні способи запису дат. Окрім того, частиномовна категоризація є занадто загальною і не віддзеркалює спільності граматичної поведінки слів. З одного боку, до різних частин мови потрапляють слова з однаковою або дуже наближеною граматичною поведінкою (порядкові числівники, дієприкметники, прикметники, і т.д.), з іншого ж, у межах однієї частини мови трапляються різні підкласи, і це не тільки сумнозвісна займенникова проблематика. Щоб провести чіткішу демаркаційну лінію між лексико-граматичними класами, укладачі корпусів вдаються до різних методів. У тому ж чеському корпусі проблема вирішена за допомогою введення категорії підчастини мови. Так, якщо «технічних» частин мови там налічується 15, то підчастин – уже 75, зокрема за рахунок деталізованого класоподілу займенників. Внаслідок додаткової категоризації чеський тагсет є дуже розбудованим і нараховує біля 4 тисяч можливих комбінацій кодів в окремих тагах. Такий спосіб категоризації у тагсеті називаємо **гніздовим**. Підчастинам мови притаманна однакова граматична поведінка, що спрощує завдання обробки інформації. Звісно, що наслідком такого підходу є певний надмір інформації.

Інакше до проблеми підійшли розробники польського Корпусу IPI. Кореневі лексико-граматичні класи там представлені **флексемами**, під якими розуміють множини слів зі спільними наборами змінних граматичних атрибутів. Поняття та термін ввів Януш Бєнь у 1960-тих, застосовуючи його при роботі з частотним корпусом польської мови. Таких класів для польської мови налічується 29, причому деякі з них, як то *siebie, winien, będę*, складаються з одної-двох лексичних одиниць, а навіть і окремих фрагментів парадигм.[3] Більшість флексем групується в рамках однієї частини мови, напр. іменники, дерогативні іменники та герундії можна об’єднати у групі іменників, але деякі з них охоплюють різні частини мови, напр. *partykuło-przysłówki* (частко-прислівники) поєднують частки і прислівники, а власне іменники включають також неособові іменникові займенники. Флексема прикметників охоплює також порядкові числівники

[3] Пор. трактування особових закінчень дієслова у теперішньому часі, напр. *-(e)m, -(e)ś,* як форм дієслова *być* «бути». Такий відхід від загальноприйнятої категоризації пов’язаний з потребою економності запису та механізмом аглютинації, який використовується для цього.

та присвійні займенники, що мають подібну граматичну характеристику. Для практичної роботи мовознавців з корпусним матеріалом цей поділ на флексеми є дещо штучним і надто подрібненим, тому на рівні пошуку для групування деяких флексем та вартостей їх атрибутів використовують механізм альясів (від лат. *alias* «інакше»). Таким чином, порівняно з чеським способом запису, уникають надлишковості інформації в тагах.

Приклади альясів:

masc	m1 m2 m3
noun	subst depr ger xxs ppron12 ppron3 PPRON GNOUN PROPNOUN
pron	ppron12 ppron3 siebie PPRON
verb	fin praet aglt będzie inf imps impt **pact ppas pcon pant** ger winien PART(PPAST) FUT PRES

Для витримки логічності та послідовності структури граматичного опису введено ряд додаткових характеристик-атрибутів, які не трапляються в традиційних описах польської граматики. Зокрема, це категорії: заперечення (для герундія, дієприкметника та дієприслівника); акцентованості (для прийменника), післяприйменниковості (для прикметника), акомодації (для числівника), аглютинації (для т.зв. л-вого дієприкметника[4], для якого використовують термін *pseudoimiesłów*), вокалічності (для дієслівних закінчень теперішнього часу, що трактуються як форми дієслова *być*).

Враховуючи характеристику та перелічені переваги різних типів граматичного кодування, вирішено зупинитися на позиційно-гніздовому типі спільного польсько-українського тагсету, з більшою опорою на тагсет Корпусу IPI. У тагах записана інформація тільки про підчастину мови, натомість альяси підключатимуться на рівні пошуку. Формат запису тагованих текстів відповідає стандарту XCES.

Фрагмент українського тагованого тексту після конвертування (XCES):

```
<?xml version="1.0" encoding="UTF-8"?>
…
<tok><orth>Оопубліковані</orth>
<lex><base>опублікований</base><ctag>adj:pl:nom</ctag></lex>
<lex><base>опублікований</base><ctag>adj:pl:acc</ctag></lex>
</tok>
<tok><orth>наприкінці</orth>
<lex><base>наприкінці</base><ctag>adv</ctag></lex>
</tok>
<tok><orth>минулого</orth>
<lex><base>минулий</base><ctag>adj:sg:gen:masc</ctag></lex>
<lex><base>минулий</base><ctag>adj:sg:acc:masc</ctag></lex>
<lex><base>минулий</base><ctag>adj:sg:gen:n</ctag></lex>
</tok>
<tok><orth>тижня</orth>
```

[4] Основа дієслівної форми минулого часу.

<lex><base>тиждень</base><ctag>gnoun:sg:gen:masc</ctag></lex>
</tok>
<tok><orth>y</orth>
<lex><base>y</base><ctag>int</ctag></lex>
<lex><base>y</base><ctag>prep</ctag></lex>
<lex><base>y</base><ctag>prep</ctag></lex>
<lex><base>y</base><ctag>prep</ctag></lex>
</tok>
<tok><orth>Financial</orth>
<lex><base>Financial</base><ctag>ign</ctag></lex>
</tok>
<tok><orth>Time</orth>
<lex><base>Time</base><ctag>ign</ctag></lex>
</tok>
<tok><orth>звинувачення</orth>
<lex><base>звинувачення</base><ctag>gnoun:sg:nom:n</ctag></lex>
<lex><base>звинувачення</base><ctag>gnoun:sg:gen:n</ctag></lex>
<lex><base>звинувачення</base><ctag>gnoun:sg:acc:n</ctag></lex>
<lex><base>звинувачення</base><ctag>gnoun:sg:voc:n</ctag></lex>
<lex><base>звинувачення</base><ctag>gnoun:pl:nom:n</ctag></lex>
<lex><base>звинувачення</base><ctag>gnoun:pl:acc:n</ctag></lex>
<lex><base>звинувачення</base><ctag>gnoun:pl:voc:n</ctag></lex>
</tok>

5. Інтерпретація вибраних категорій у граматичних словниках і корпусах

При зіставленні мов на теоретичному рівні виникає багато питань щодо категоризації та значень атрибутів категорій. На практичному рівні, коли залучається масив мовного матеріалу, ці проблеми стають ще гострішими. Виникають питання розподілу мовних одиниць за категоріями, стикаємося з різним трактуванням атрибутів, навіть таких, здавалося б, однозначних, як відмінок та рід. Практично на рівні кожної категорії чи значень її атрибутів вимагається прийняття рішення на користь одного чи другого, або й незалежного, альтернативного розв'язку. З огляду на обсяг матеріалу у цьому розділі зможемо обговорити лише деякі з категорій, див. також [Kotsyba, Shypnivska, Turska, 2008].

5.1 Дієприслівники

На відміну від тагсету УМІФ, де дієприслівники розглядаються як дієслівні форми, у Корпусі IPI дієприслівники трактуються як окремі флексеми (відповідно adv.pres.prtcp. від „present adverbial participle" і adv.anter.prtcp. від „anterior adverbial participle"). Набір атрибутів для цієї категорії вичерпується самою дефініцією флексеми. Натомість у тагсеті УМІФ формі дієприслівника притаманний також вид і час. Для польських дієприслівників вид не позначають, тому що згідно із сучасною граматичною традицією дієприслівники минулого часу завжди є доконані, а теперішнього – недоконані. Категорія часу дієприслівників є досить умовною категорією, оскільки одночасність чи попередність, яку вона передає, є прив'язана не до моменту мовлення, а до інших подій, які можуть мати різну часову віднесеність. Цим теоретичним моментом пояснюється вживання саме термінів *współczesny* „одночасний" і *uprzedni* „попередній" у згаданих польських флексемах, а не „теперішній" та „минулий". Ситуація з українською мовою є дещо складніша, оскільки в ній функціонують

дієприслівники, які вживаються тільки у минулому часі (напр. „працювавши”)[5], що є аргументом для застосування відповідної позначки у їхньому граматичному описі, пор. опис у порядку: код, інтерпретація, приклад.

1. **VW** дієслово, дієприслівник, доконаний вид, ***минулий** час, активний стан, *прочитавши* (піду, прочитавши)

2. **UW** дієслово, дієприслівник, недоконаний вид, минулий час, активний стан, *читавши*.

3. **UQ** дієслово, дієприслівник, недоконаний вид, ***теперішній** час, активний стан, *читаючи* (робив/робитиме читаючи).[6]

Отже, до спільного тагсету PolUKR дієприслівники включено як дієслівні форми, що характеризуються часом та видом (відповідні таги: verb:part:perf , verb:part:imperf, verb:part:imperf:praet).

5.2 Займенники

Займенники є дуже різноманітною категорією, і останнім часом далеко не всі дослідники схильні їх визначати як окрему частину мови на самій лише підставі семантичного характеру змінності. У слов’янській корпусології їх вважають сумнозвісною проблемою – детальний опис 309 займенників дає 296 можливих тагів, пор. [Paskaleva 2007]. У Корпусі IPI, залежно від частини мови, котру «замінюють», займенники належать до тієї чи іншої категорії, причому без відрізнення від незайменникових одиниць. Окремо виділяються тільки ті займенники, котрі мають виразно розбіжну синтаксичну поведінку, а саме: займенники 1-ї та 2-ої особи; окремо займенники 3-ї особи; *siebie* є теж окремою флексемою. В українській спостерігаємо подібний підхід: є займенник-іменник (до якого належать також всі особові та зворотний), займенник-прикметник.[7] Спільний тагсет PolUKR використовує підхід УМІФ, з додатковим поділом Корпусу IPI на 1-2 і 3 особові займенники, тим самим зберігаючи суму грануляції.

5.3 Предикативи

Питання предикативів впродовж усієї історії їхнього існування залишалося контроверсійним. Проблему цієї категорії детально описуємо в [Derzhanski, Kotsyba 2008], де подається її інтерпретація в історії граматичної традиції для чотирьох мов: російської (для якої вони вперше були виділені), української, польської та болгарської. Щодо складу та критеріїв належності до цієї категорії повної згоди серед авторів немає.

[5] Корпусні дослідження показують, що у польській живій мові ця дещо архаїчна форма теж зустрічається, але оскільки вона не є запрограмована у тагсеті, тагер її не може правильно описати.

[6] Інші вживання «часів» можна вважати умовними, що позначаємо зірочкою.

[7] Інакше трактують займенники у різних версіях MTE, де вони поділені на традиційні семантичні розряди. При конвертуванні УГС до формату MTE ці групи було визначено та додатково закодовано.

Навіть погляди тих самих авторів змінювалися у часі. Цим пояснюється різниця представлення цих категорій в корпусах відповідних мов: для польської їх вирізняють 26, для української 176 (т. зв. присудкові слова), для російської більше 1200 (згідно зі словником Єфрємової [Ефремова 2006]), у болгарському та чеському корпусах їх не виділяють взагалі, трактуючи еквіваленти предикативів вищезгаданих мов у рамках інших частин мови. У спільному польсько-українському тагсеті категорія предикативів відсутня, а слова, що трактуються як предикативи у польській або українській мові, перекеровано до інших частин мови. Нижче подано приклади з обох мов (оригінальний предикатив зазначено курсивом):

займенник *to*, напр. *To* książka. (То книжка.) → займенник-іменник;

модальні слова: *można, trzeba, wolno, wiadomo, trza, niepodobna, podobna, dość, dosyć* →модальні прислівники (диференціація на семантичному рівні);

віддієслівні: *słychać, widać, stać, czuć, znać* і відіменникові: *szkoda, potrzeba, żal, wstyd, strach, pora, czas, brak, śmiech* → прислівники;

похідні від семельфактивів: *зирк, круть-верть* → вигуки[8];

відприслівникові: *зимно, безвітряно, безсніжно* → прислівники;

демінутивні дієслова: *спунькати, спатки, ходитоньки, їсточки, їстки* → неозначена форми дієслова (інфінітив).

5.4 Прикметники

Значні розбіжності між двома мовами спостерігаємо також в описі прикметників. В УГС УМІФ прикметники вищого і найвищого ступенів подано як окремі лексеми, ступінь як категорія не виділяється. У Корпусі IPI до прикметників зараховують деад'єктивні частини одиниць типу *po-polsku* „по-польськи", які трактуються як прислівники в українській. Складені прикметники типу *polsko-niemiecki* розглядаються як сполучення прикметників двох різних типів, де частина *polsko-* має додатний атрибут приприкметниковості. Підрахунок слів у текстах та словниках, відповідно, даватиме не до кінця співвимірні результати. Щоб вирівняти можливості пошуку в українських та польських текстах, було розроблено алгоритм визначення прикметників вищого та найвищого ступеня з відповідним записом граматичної інформації та їх релематизації. Польські післяприйменникові прикметники трактуються як частини прислівників.

6. Інші формати запису морфосинтаксичної інформації в PolUKR

Враховуючи реальну потребу і можливість у майбутньому розширення польсько-українського корпусу на інші слов'янські мови, а також інтегрування його матеріалів до інших розроблених паралельних корпусів, що охоплюють слов'янські мови,

[8] Цю групу, можливо, варто занотувати як дефектну форму дієслова. Але таке рішення вимагало би додаткового дослідження.

опрацьовано можливість конвертування граматичної інформації так, як вона закодована у PolUKR, до міжнародно визнаних форматів. Проектом зі спільним граматичним тагсетом, який охоплює найбільшу кількість слов'янських мов, є MULTEXT-East, тому саме його було вибрано як взірець. До MULTEXT-East у його перших трьох версіях не ввійшла ані польська мова, ані жодна зі східнослов'янських, тому тагсети для польської та української у цьому форматі опрацьовуються нами окремо, пор. також [Derzhanski, Kotsyba 2009]. Ці тагсети були максимально допасовані до решти слов'янських мов MTE-3, включно з російською, доданою дещо раніше до останньої на цей момент версії MTE-4, що все ще перебуває у стані розробки. Український варіант тагсету в форматі MTE також включено до версії MTE-4. Опрацьовано таблицю відповідностей кодування спільного польсько-українського тагсету та MTE-4. Таким чином, у PolUKR можна буде здійснювати граматичний пошук у двох форматах – розширеному, із максимальним збереженням інформації з граматичних словників обох мов, та скороченому, що «розпізнається» міжнародними стандартами.

Бібліографія

Bień, J.S. 1991. Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji, Rozprawy Uniwersytetu Warszawskiego, t. 383. Warszawa: Wydawnictwo Uniwersytetu Warszawskiego.

British National Corpus: http://www.natcorp.ox.ac.uk/

Dalewska-Greń, Hanna. *Języki słowiańskie*, *Warszawa*, PWN, 1997.

INTERA unified tagset project: http://www.elda.org/intera

Derzhanski, Ivan and Natalia Kotsyba. *The Category of Predicatives in the Light of Consistent Morphosyntactic Tagging*, "Lexicographic Tools and Techniques." Proceedings of MONDILEX First Open Workshop, Moscow, Russia, 3-4 October, 2008. Moscow 2008, p. 68-79.

Derzhanski, Ivan and Natalia Kotsyba. *Towards a Consistent Morphological Tagset for Slavic Languages: Extending MULTEXT-East for Polish, Ukrainian and Belarusian.* Metalanguage and Encoding Scheme Design for Digital Lexicography: MONDILEX Third Open Workshop, Bratislava, 15–16 April 2009.

Tomas Erjavec et al. *Multext-East specifications for Slavic languages*, Budapest, 2003.

Jan Hajič. Positional Tags: Quick Reference (Czech „HM" Morphology), 2000.

Kotsyba, Natalia, Olha Shypnivska and Magdalena Turska. *Linguistic principles of organizing a common morphological tagset for PolUKR (Polish-Ukrainian Parallel Corpus)*. Proceedings of the international conference "Intelligent Information Systems, 16-18 June 2008, Zakopane, Poland", Warszawa, 2008.

ParaSol: A Parallel Corpus of Slavic and other languages: http://www.uni-regensburg.de/Fakultaeten/phil_Fak_IV/Slavistik/RPC/

Paskaleva, Elena. *Balcan South-East Corpora Aligned to English*. In: *The Proceedings of the Workshop on Common Natural Language Processing Paradigm for Balkan Languages*, EACL 2007.

Polsko-ukraiński korpus równoległy (PolUKR): http://corpus.domeczek.pl/index.php?option=search

Przepiórkowski, Adam and Marcin Woliński. *A Flexemic Tagset for Polish*. In: *The Proceedings of the Workshop on Morphological Processing of Slavic Languages*, EACL 2003.

Tagger TaKIPI: http://nlp.ipipan.waw.pl/TaKIPI/

Tokarski, Jan. *Schematyczny indeks a tergo polskich form wyrazowych.* Opracowanie i redakcja Zygmunt Saloni. Wydanie drugie. Warszawa: Wydawnictwo Naukowe PWN, 2001.

Ефремова Татьяна Ф., *Современный толковый словарь русского языка*. Москва: Астрель, 2006.

Шевченко І.В. *Алгоритмічна словозмінна класифікація української лексики*. // Мовознавство. 1996. №4–5, с. 40–44.

Шевченко И.В., Широков В.А., Рабулець А.Г. *Электронный грамматический словарь украинского языка*. // Труды международной конференции «Megaling'2005. Прикладная лингвистика в поиске новых путей». 27 июня – 2 июля 2005 года. Меганом, Крым, Украина. С. 124–129.

Широков В.А, О.В.Бугаков, Т.О.Грязнухіна, О.М.Костишин, М.Ю.Кригін, Т.П.Любченко, О.Г.Рабулець, О.О.Сидоренко, Н.М.Сидорчук, І.В.Шевченко, О.О.Шипнівська, К.М.Якименко *Корпусна лінгвістика*. Київ: Довіра, 2005.