УДК 004.827

С.А. Катренко, О.М. Демська-Кульчицька*
Національний університет "Львівська політехніка",
кафедра прикладної лінгвістики
*Інститут української мови НАН України

УСУНЕННЯ НЕОДНОЗНАЧНОСТІ В ТЕКСТОВИХ ІНФОРМАЦІЙНИХ РЕСУРСАХ НА МОРФОЛОГО-СИНТАКСИЧНОМУ РІВНІ

Морфолого-синтаксичний аналіз є одним з рівнів аналізу тексту. Як було зазначено багатьма дослідниками у цій галузі, цей аналіз не є тривіальним, особливо для флективних мов. Ця стаття подає підхід, що застосовувався для зняття неоднозначності на морфолого-синтаксичному рівні для української мови. Окрім цього, також описано існуючі підходи для здійснення морфолого-синтаксичного аналізу.

The parts-of-speech tagging (PoS) is one of the levels of text analysis. As it has been pointed out by the researchers in this field, the PoS tagging is not a trivial task. It is especially true when considering the languages with rich morphology. This paper thus presents the approach taken to perform parts-of-speech disambiguation for Ukrainian. Besides this, the variety of approaches to PoS disambiguation proposed so far is also discussed.

Постановка проблеми у загальному вигляді

Опрацювання природномовного тексту складається зазвичай з кількох етапів, починаючи з токенізації, себто виявлення ймовірних словоформ, далі – морфолого-синтаксичного, синтаксичного аналізів та закінчуючи семантичним та прагматичним аналізом. Хоча існує взаємодія між усіма рівнями, процес аналізу тексту є послідовним. Токенізація є доволі тривіальним завданням для таких мов як українська, але її значно важче здійснювати для, скажімо, китайської [12]. Морфолого-синтаксичний аналіз виконується після виявлення словоформ і полягає у присвоєнні кожній з них позначки (теґу), який містить інформацію про частину мови та граматичні категорії. Проте здійснення цього аналізу ускладнюється наявністю таких явищ, як лексико-граматична омонімія та внутрішньопарадигматична неоднозначність (пр.1).

Приклад 1
Мила мати йде до хати.
Мила (а) Іменник, род. відм. (мило)
(б) Іменник, наз. відм. (мила)
(в) Прикметник, жін. рід, наз. відм. (мила)
(г) Дієслово, мин. час, 3 ос. одн. (мити)
Мати (а) Іменник, наз. відм. (мати)
(б) Дієслово, неозначена форма (мати)

Як показують попередні дослідження, усунення неоднозначностей на морфолого-синтаксичному рівні підвищує ефективність багатьох застосувань, для прикладу, використання лексико-граматичної інформації зумовило вищу точність під час видобування інформації. Ще однією мотивацією для роботи над усуненням неоднозначності є те, що це уможливлює подальше опрацювання тексту, насамперед синтаксичний аналіз.

У цій статті описано підходи, що застосовуються для усунення неоднозначності на морфолого-синтаксичному рівні, а також описано структуру позначок та застосування евристичного методу усунення неоднозначності на морфолого-синтаксичному рівні для української мови.

Аналіз останніх досліджень

Підходи

Серед існуючих методів усунення неоднозначності на морфолого-синтаксичному рівні можна виділити два напрями – статистичні методи та евристики[1], написані експертами. До досліджень першого типу належать [11, 16], а до останніх — [15, 17, 21, 22]. Вибір методу зумовлений такими факторами, як розмір анотованого корпусу текстів, затрати часу тощо. Хоча спільним для усіх цих підходів є те, що вони є фактично навчанням "з вчителем". Тобто, для оцінки ефективності методів (точності та повноти) необхідно мати дані, у яких неоднозначність усунена, – "золотий стандарт". Ці дані містяться в анотованих корпусах текстів — репрезентативній вибірці текстів у електронному вигляді. Аналогом репрезентативності є поняття репрезентативної вибірки у статистиці, а саме, вибірка повинна адекватно та якомога повніше відображати характеристики генеральної сукупності. Отже, репрезентативна вибірка текстів мала б охоплювати різні мовні феномени. Для кількох мов, зокрема й української [18], проводилися дослідження з автоматичним отриманням позначок для словоформ, використовуючи машинне навчання "без вчителя". Дослідження продемонстрували, що кластеризаційні методи ефективно групують слова за частинами мови (для української мови – й за відмінками), однак неможливо отримати інформацію про інші граматичні категорії (рід, число), використовуючи дистрибутивний аналіз.

Підхід, що базується на правилах (евристиках), є трудомістким і вимагає залучення експертів, проте його перевагою є можливість охопити небезпосередні (неконтактні) залежності, як у реченні "Цікаву я прочитав книжку", де прикметник *цікаву*, що узгоджується з іменником *книжку* в роді, числі та відмінку, знаходиться на певній відстані від іменника.

Евристику можна подати так:

$$[x,y] = |LC| \text{ у } |RC|,$$

де [x,y] — можливі інтерпретації (теґи), у — обрана інтерпретація, LC та RC — контексти ліворуч та праворуч обраної інтерпретації. Можливими є ситуації, коли один з контекстів не задано.

Типовим прикладом евристики для англійської мови є така :

За артиклем не може знаходитися дієслово.

If tag_{w-1}=det then $tag_w \neq$verb або [verb,noun] = |det| noun

Статистичні методи, що теж використовуються для здійснення морфолого-синтаксичного аналізу, є простими у застосуванні, однак передумовою їх використання є наявність великого за розміром корпусу текстів. Недоліком є також те, що простір параметрів є звикло великим, для прикладу, кількість параметрів у випадку триграмної моделі та кількості теґів, що дорівнює 700, сягає $700^3 = 343$ млн. N-грамна модель описується за допомогою ланцюгів Маркова таким чином [10]:

нехай $W = (w_1,...,w_L)$ позначає послідовність словоформ тексту, де L — кількість словоформ, а $T = (t_1,...,t_L)$ — послідовність позначок (теґів), які відповідають словоформам. Ймовірність цієї послідовності теґів T при послідовності слів W $p(T|W)$ визначається з врахуванням правила Байєса як

$$p(T|W) = \arg\max_T p(W|T)p(T).$$

У свою чергу, $p(T)$ та $p(W|T)$ обчислюються за формулами

$$p(T) = \prod_{i=1}^{L} p(t_i | t_{i-1},...,t_1)$$

$$p(W|T) = \prod_{i=1}^{L} p(w_i | T, w_{i-1},...,w_1)$$

Ці формули можна наблизити:

$$p(t_i | t_{i-1},...,t_1) \approx p(t_i | t_{i-1},...,t_{i-n}),$$

$$p(w_i | T, w_{i-1,...,} w_1) \approx p(w_i | t_i)$$

[1] Тут евристики слід інтерпретувати як "заснований на власному досвіді принцип, правило, тактичний прийом,...,який допомагає досягти необхідного результату в пошуку або розв'язанні задачі" [2, с. 266]

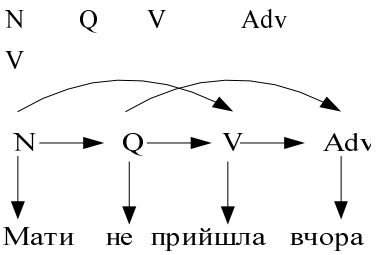
Модель n-го ряду називають (n+1)-грамною моделлю. Відповідно, найуживаніша триграмна модель подається як

$$\tau(w_1,...,w_L) = \arg\max_T \prod_{i=1}^{L} p(t_i \mid t_{i-2}, t_{i-1}) p(w_i \mid t_i).$$

Отже, теґер є моделлю Маркова, де станами є позначки, а ймовірностями переходу є ймовірності присвоєння теґа, використовуючи інформацію про попередній теґ. Стаціонарними ймовірностями є ймовірності використання даного слова за умови присвоєння даного теґа.

Приклад 2

Мати не прийшла вчора.



$$p = p(N)p(Q \mid \$, N)p(V \mid N, Q)p(Adv \mid Q, V)\times$$
$$\times p(мати \mid N)p(не \mid Q)p(прийшла \mid V)p(вчора \mid Adv)$$

Ще одним напрямком у теґуванні є трансформаційно-орієнтоване навчання [9]. Особливістю цього підходу є те, що правила для усунення неоднозначностей продукуються автоматично. Для цього необхідно мати корпус текстів, де неоднозначність є повністю усунута, а також лексикон з усіма можливими позначками для кожного елемента лексикону. Як наслідок, трансформаційні правила формуються на основі частотності усіх можливих теґів у певному контексті і зазвичай мають таку форму:

Змінити теґ обраного слова з χ на Y у контексті C, де $Y \in \chi$, для цього для кожного теґу $Z \in \chi, Z \neq Y$, обчисливши відношення частотності появи слова з однозначним теґом Y $F(Y)$ до частотності слів з однозначним теґом Z $F(Z)$, помноживши його на кількість разів $Q(Z,C)$, де слово однозначно позначене через Z виникає у контексті C, тобто

$$F(Y)/F(Z)*Q(Z,C)$$

Результати морфолого-синтаксичного аналізу для різних мов подані у табл. 1. Варто зауважити, що для слов'янських мов такої інформації не існує (окрім як для чеської, польської та болгарської), що насамперед пов'язане з тим, що такі дослідження почали проводитися порівняно недавно.

Таблиця 1

Морфологосинтаксичний аналіз для різних мов

Мова	**Розмір корпусу**	**Кількість позначок**	**Метод**	**Точність**
Чеська [16]	29972 токенів – для тестування			96,13 %
Польська [11]	590К токенів – для тренування 4200 токенів –для тестування	не вказано	Триграмна модель	90,6%[2]
Англійська [20]	1120К токенів–для тренування 50К токенів – для тестування	45	Дерева рішень	97,29%
Німецька [17]	5732 токени	не вказано	Евристики	77,08%
Турецька [21]	2000 речень (для тренування) 270 речень (для тестування)	не вказано	Правила, написані експертом + статистична інформація	93,70%
Румунська [24]	>60000 слів (для тренування) 20109 слів (для тестування)	615 та 92	Поєднання класифікацій-них моделей	98,78%
Іспанська [20]	70К токенів – для тренування 25К токенів –для тестування	62	Дерева рішень у поєднанні з алгоритмом релаксації	97,4%

[2] У таблиці подано найкращі результати, що були отримані для цих мов. Повний опис значень параметрів та відповідної точності наведено у [11,16,20,21,24].

Основний матеріал

Створення та опис набору позначок для українського корпусу

Форма і семантика морфологічного теґа і, відповідно, теґсет для конкретної мови залежить від системи лексико-граматичних класів слів та морфологічних значень, релевантних для кожного із цих класів, поданих експліцитно.

Анотаційно зорієнтована диференціація слів за частинами мови не завжди відповідає традиційному для слов'янських мов лексико-граматичному поділові. Так, порівняємо набір лексико-граматичних розрядів слів польської, чеської, російської та української мов, які покладено в основу побудови морфологічних теґсетів.

Морфологічні теґсети для слов'янських мов

Для корпусу польської мови IPI PAN [25] виділено 21 лексико-граматичний клас слів:

1. Іменник (rzeczownik);
2. Прикметник (przymiotnik);
3. Степеневий прислівник (przysłówek stopniowany);
4. Прийменниковий 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 (rozkaznik);
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 дієприслівник (imiesłów przymoitnikowy czynny);
18. Пасивний дієприслівник (imiesłów przymoitnikowy bierny);
19. Прийменник (przyimek);
20. Сполучник (spójnik);
21. Частко-прислівник (partykuło-przysłówek).

Для теґсета Чеського національного корпусу виділено 10 т. зв., базових частин мови з двома окремими мітками для нез'ясованих / невідомих / некласифікованих одиниць і пунктуаційних символів [13]: прикметник (Adjective), числівник (Numeral), прислівник (Adverb), вигук (Interjection), сполучник (Conjunction), іменник (Noun), займенник (Pronoun), дієслово (Verb), прийменник (Preposition), частка (Particle).. Кожна з основних частин мови деталізована і мітки деталізації об'єднано у субчастини мови (subpos), наприклад: A – adjective, general; C – adjective, nominal (short, participial) form rád, schopen, ...; G – adjective derived from present transgressive form of a verb.

У морфологічному стандарті Національного корпусу російської мови розподіл за частинами мови є таким:

- S – существительное (яблоня).
- S-PRO – местоимение-существительное (она).
- A – прилагательное (коричневый).
- A-PRO – местоимение-прилагательное (который, твой).
- NUM – числительное (четыре).
- A-NUM – числительное-прилагательное (один).
- PRAEDIC – предикатив (жаль).
- A-PRAEDIC – местоимение-предикатив (некого, нечего).
- V – глагол (пользоваться).

• ADV – наречие (сгоряча, очень).
• ADV-PRO – местоименное наречие (где).
• PR – предлог (под).
• CONJ – союз (и, чтобы).
• PART – частица (бы).
• INTJ – междометие (увы).
• PARENTH – вводное слово (кстати).

Тобто виділено 16 частин мови, серед яких фіксуємо традиційні: іменник, прикметник, числівник, дієслово, прислівник, прийменник, сполучник, частка і вигук. А також: займенник-іменник, займенник-прикметник, числівник-прикметник, предикатив, займенник-предикатив, займенниковий прислівник та вставне слово. На наш погляд, в основі такого розподілу лежить синтаксично-функціональний та формальний аспекти розподілу слів російської мови на лексико-граматичні розряди.

Анотаційно зорієнтований репертуар морфологічних значень аналогічно до частиномовного розподілу також є індивідуальним для кожної конкретної мови. Узагальнені морфологічні та синтаксичні значення прийнято розглядати в морфології як граматичні категорії, серед яких 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" [4]. З названих далеко не всі морфологічні значення послідовно відображають у структурі теґів (табл. 2).

Таблиця 2

Параметри, виокремлені у теґсетах чеської, польської та російської мов

Параметр	Чеський	Російський	Польський
1	2	3	4
число	S (однина), P (множина), D (двоїна), X both, or special combinations	sg — однина (яблоко), pl — множина (яблоки)	sg однина (*oko*), pl множина (*oczy*)
відмінок	1 називний, 2 називний, 3 давальний, 4 знахідний, 5 кличний, 6 місцевий, 7 орудний, X будь-який	nom називний (голова), gen родовий (головы), acc знахідний (голову), dat давальний (голове), loc місцевий([о] голове), ins орудний (головой), gen2 другий родовий (чашка чаю), acc2 другий знахідний (по два человека), loc2 другий місцевий (на оси), voc клична форма (Господи)	nom називний (*woda*), gen родовий (*wody*), dat давальний (*wodzie*), acc знахідний (*wodę*), inst орудний (*wodą*), loc місцевий (*wodzie*), voc кличний (*wodo*);
рід	M чол. істота, I чол. неістота, F жіночий, N середній, X будь-який, Y -M або I (чол.), H -F або N (не чол.), Q-F або N (не чол.), у спец. випадках, T-I або F (чол. неістота або жіночий), Z-M, I або N (не жіночий), W-I or N (не чол. істота, не жіночий)	m — чоловічий (стол), f — жіночий (работница), m-f — «общий род» (задира, пьяница), n — середній (озеро)	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*)

Продовж. табл. 2

1	2	3	4
ступені порівняння	1 нульовий 2 вищий 3 найвищий	comp — вищий (глубже) comp2 — форма ‘по+ вищий’ (поглубже) supr —найвищий (глубочайший)	pos нульовий (*cudny*), comp вищий (*cudniejszy*), sup найвищий (*najcudniejszy*);
заперечення	A незанегована N заперечна форма	Не виділено	aff незанегована (*pisanie*), neg заперечна форма (*niepisanie*)
особа	1 1st 2 2nd 3 3rd	1p — перша (украшаю) 2p — друга (украшаешь) 3p — третя (украшает)	pri перша (pierwsza: *bredzę*), sec друга (druga: *bredzisz*), ter третя (trzecia: *bredzi*)
стан	A активний, P пасивний	act — активний (разрушил) pass — пасивний (разрушенный) med — “медиальный, или средний залог” (форми на -ся: разрушился)	
вид		pf — доконаний (пошёл), ipf — недоконаний (ходил)	imperf — недоконаний (*iść*), perf — доконаний (*zajść*);
час	P теперішній, R минулий, U майбутній, H минулий або теперішній, X будь-який	praet — минулий (украшали) praes — теперішній (украшаем) fut — майбутній (украсим)	
спосіб	N дійсний, R наказовий, C умовний (лише деякі дієслова)	indic — дійсний (украшаю), imper —наказовий (украшай)	
перехідність	Не виділено	intr неперехідність (вариться), tran перехідність (вести)	Не виділено
коротка форма		brev коротка (высок), plen повна (высокий)	
форма дієслова		inf інфінітив (украшать) partcp причастие (украшенный) ger дієприслівник (украшая)	
депреціативність	Не виділено		ndepr недеперціативна форма (niedeprecjatywna: *chłopi*), depr деперціативна (deprecjatywna: *chłopy*)
акцентованість			akc акцентована форма (akcentowana: *jego, niego*), nakc ненаголошена (nieakcentowana: *go, -ń*)
післяприйменниковість			praep: післяприйменникова форма (*niego, -ń*), npraep:непіслясприйменникова форма (*jego, go*)
акомодативність			узгодження (uzgadniająca = congr: *dwaj*), керування (rządząca = rec: *dwóch, dwu*)
аглютинатиність			nagl неаглютинативна форма (*niósł, dlaczego*), agl аглютинативна (*niósł-*)
вокаліч-ність			wok вокалічна форма (*-em*), nwok невокалічна (*-m, -ś, z*)

Морфологічний теґсет для української мови

Для морфологічного теґсета в межах проекту Національного корпусу української мови передбачено розрізнення таких граматичних класів слів:

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).

Разом шістнадцять граматичних класів.

Схема морфологічної анотації українського корпусу 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, є) спосіб, ж) ступінь порівняння, кожне з яких може набувати лише таких 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;
- ступінь порівняння: нульовий↔вищий↔найвищий

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

- чоловічий = m (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 = p (locative або prepositional);
- кличний = v (vocative);
- особа: 1↔2↔3 – позначена відновно цифрами 1,2,3;
- теперішній = p (present);
- минулий = t (past);
- майбутній = f (future);
- давноминулий = c (pluperfect);
- доконаний = p (perfect);
- недоконаний = i (imperfect);
- двовидова форма = d (double form);
- активний = a (active);
- пасивний = p (passive)

Отже, зведена таблиця анотаційно орієнтованого частиномовного набору в порівнюваних мовах з доповненням морфологічних значень матиме вигляд (де X— наявність певного парметра для певної частини мови) (табл. 3):

Таблиця 3

Параметри, що виділені для частин мови

Частина мови	рід	число	відмінок	час	особа	аспект	спосіб	ступінь порівняння	стан
N (іменник)	X	X	X						
A (прикметник)	X	X	X					X	
P (займенник)	X	X	X						
V (дієслово)	X	X		X	X	X	X		X
F (інфінітив)									
D (предикатив)									
T (дієприкметник)	X	X	X	X		X			X
B (дієприслівник)				X		X			
L (порядковий / прикметниковий числівник)	X	X	X						
M (власне числівник)			X						
MN (числові назви)			X						
R (прислівник)								X	
C (сполучник)									
S (прийменник)									
Q (частка)									
I (вигук)									

Евристики

Евристики ґрунтуються на знаннях експерта з української мови. Як вже зазначалося, застосування евристик якнайкраще описує неконтактні залежності. Тому при створенні евристик розглядався не лише фіксований контекст одиниці, що має кілька інтерпретацій на морфолого-синтаксичному рівні, а й контекст речення.

Складністю морфолого-синтаксичного аналізу є наявність внутрішньопарадигматичної та лексико-граматичної омонімії. За визначенням, внутрішньопарадигматичною омонімією є збіг

словоформ в межах однієї парадигми, лексико-граматичними омонімами є слова, що належать до різних частин мови.

Слід наголосити, що у цьому випадку не враховуються лексичні омоніми — слова, що належать до однієї частини мови, однакові за звучанням або написанням, але які мають різне значення (мул — свійська тварина та відклади на дні водоймищ) [1]. Насамперед, неврахування цього типу омонімії пояснюється тим, що різні значення слів не наведено у лексиконі. Тобто лексикон, що використовується для морфолого-синтаксичної анотації містить лише інформацію про частини мови та граматичні категорії, але аж ніяк не лексичні значення слів. За нашою інформацією, у жодному з досліджень у цій галузі не використовували таку інформацію та не розрізняли лексичні омоніми.

Усі евристики можна умовно поділити на такі:

- ті, що використовують узгодження (для зняття внутрішньопарадигматичної неоднозначності);
- ті, що використовують прийменникове керування;
- ті, що використовують інформацію про поєднання лексичних одиниць у тексті (для зняття лексико-граматичної омонімії).

За даними, котрі використовувалися у цьому дослідженні, в українській мові припадає 1,9 можливих інтерпретацій (теґів) на кожну словоформу. Для порівняння, для румунської мови цей показник становить 1,6 інтерпретацій/словоформу, в англійському корпусі WSJ на одне слово припадає 1,52 теґи, а для німецького тексту – більше ніж 3. В останньому випадку такий високий рівень неоднозначності зумовлений іншим типом мови [19]. Водночас, за нашими підрахунками, є близько 60 різних типів внутрішньопарадигматичної омонімії для української мови (для іменників, прикметників, займенників та числівників). Приклади міжчастиномовної омонімії подано у [3].

Окрім цього, близько 47% словоформ мають більш ніж одну позначку. Для порівняння – у англійському корпусі WSJ цей показник становить 36,5%, а для іспаномовного корпусу – лише 39,26% слів у корпусі є неоднозначними [20].

Реалізація

Для проведення дослідження з усунення неоднозначності на морфолого-синтаксичному рівні було взято статті з видання “Дзеркало тижня”[3]. Такий вибір був частково зумовлений тим, що це видання має тематичний рубрикатор, тому обрані дані можна використовувати і для інших задач, скажімо, для тематичної класифікації або для кластеризації документів. Для анотування українського тексту було використано інверсний словник, тобто процес анотування здійснювався семіавтоматично — словоформи були відсортовані з кінця, чим спростився вибір позначок для більшості словоформ.

Усі статті було подано у форматі XML [26], внаслідок чого можна було створити необхідне визначення типового документа (DTD) та використати бажані теґи для анотації. Використання формату XML надає багато переваг, серед яких можливість переформатування документа засобами XSLT.

Фрагмент тексту, анотованого на морфолого-синтаксичному рівні, наведено на рис. 1. Теґ aa відповідає усім можливим позначкам для слова, а ta — правильній позначці, тобто для слова “Україні” двома можливими позначками є Nfsd та Nfsp (іменник, жін. рід, однина, давальний відмінок та іменник, жін. рід, однина, місцевий відмінок), а правильною позначкою у контексті є Nfsp. Теґи s та pt позначають речення та пунктуаційні знаки (включно з пробілами). Імовірні словоформи у тексті визначено із застосуванням каскаду регулярних граматик [17], внаслідок чого спершу виокремлювалися літери та знаки пунктуації, а потім визначалися слова (для української мови ними вважалися будь-які послідовності літер поєднано з апострофом або дефісом).

[3] www.zn.kiev.ua

```
<?xml version="1.0" encoding="windows-1251"?>
<!DOCTYPE newspaper SYSTEM "Final_ZN.dtd">
<newspaper><number>42(467)</number><date>1–7/11/2003</date>
<article>
<section>Гроші</section><subsection>Соціальний захист</subsection>
<title>ПЕНСІЙНА ЗРІВНЯЛІВКА ЛІКВІДУЄТЬСЯ. З ПЕРШОГО СІЧНЯ</title>
<author>Наталія ЯЦЕНКО</author>
<text><s>
<tok type="cyrw">Для<aa tag="S"/></tok><pt></pt>
<tok type="cyrw">фахівців<aa tag="N0pg;N0pa;Amsn0;Amsa0;Amsv0"/></tok><pt></pt>
<tok type="cyrw">не<aa tag="Q"/></tok><pt></pt>
<tok type="cyrw">таємниця<aa tag="Nfsn"/></tok><pt>,</pt><pt></pt>
<tok type="cyrw">що<aa tag="Q;C;P00n;P00a"/></tok><pt></pt>
<tok type="cyrw">нове<aa tag="Ansn0;Ansa0;Ansv0"/></tok><pt></pt>
<tok type="cyrw">пенсійне<aa tag="Ansn0;Ansa0;Ansv0"/></tok><pt></pt>
<tok type="cyrw">законодавство<aa tag="Nnsn;Nnsa;Nnsv"/></tok><pt></pt>
<tok type="cyrw">ухвалювалося<aa tag="Vais3sn"/></tok><pt></pt>
<tok type="cyrw">в<aa tag="S"/></tok><pt></pt>
<tok type="cyrw">Україні<aa tag="Nfsd;Nfsp"/></tok><pt>,</pt><pt></pt>
<tok type="cyrw">так<aa tag="C;Q;R0"/></tok><pt></pt>
<tok type="cyrw">би<aa tag="Q"/></tok><pt></pt>
<tok type="cyrw">мовити<aa tag="Fi"/></tok><pt>,</pt><pt></pt>
<tok type="cyrw">на<aa tag="S"/></tok><pt></pt>
<tok type="cyrw">виріст<aa tag="Nmsn;Nmsa"/></tok><pt>.</pt>
</s></text></article></newspaper>
```

Рис. 1. Фрагмент анотованого тексту

```
<xsl:transform version="2.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
<xsl:template match="tok">
 <xsl:for-each select="aa">
   <xsl:variable name="allt" select="@tag"/>
<xsl:choose>
<xsl:when test="../preceding-sibling::tok[text()='Для']">
<tok><xsl:value-of select="parent::tok[text()]"></xsl:value-of>
   <xsl:analyze-string select="$allt"
      regex="\.*,*((N|A|P)..g[0-2]?)">
   <xsl:matching-substring>
      <aa><xsl:value-of select="$allt"/></aa><ta><xsl:value-of select="regex-group(1)"/></ta>
   </xsl:matching-substring>
  </xsl:analyze-string>
</tok>
</xsl:when>
<xsl:otherwise><tok>
     <xsl:value-of select="parent::tok[text()]"></xsl:value-of>
     <aa><xsl:value-of select="$allt"/></aa><ta><xsl:value-of select="$allt"/></ta></tok>
     </xsl:otherwise>
</xsl:choose>
 </xsl:for-each>
</xsl:template>
</xsl:transform>
```

Рис. 2. Фрагмент XSL

Евристики були реалізовані за допомогою трансформації вхідного анотованого документа. Оскільки регулярні вирази не підтримуються у версії XSLT 1.0, а лише подані у рекомендаціях консорціуму W3C, для даного дослідження використовувалося програмне забезпечення Saxon 8.1. Цей програмний продукт дозволяє використовувати регулярні вирази, а також відповідає іншим рекомендаціям W3C у межах XSLT 2.0. Важливість використання регулярних виразів обгрунтовується на прикладі правил узгодження. Тобто, для узгодження іменника та прикметника лише у відмінку необхідно було б створювати окреме правило для кожного з відмінків, що більше – для усіх інших можливих категорій (роду, числа), що є ознакою надлишковості правил. Натомість із застосуванням регулярних виразів таке узгодження описується лише одним правилом. Приклад правила наведено на рис. 2 цієї статті, а фрагмент тексту, у якому повністю була знята неоднозначність для тексту на рис. 1, наведено на рис. 3.

```
<text>
<s><tok type="cyrw">Для<aa>S</aa><ta>S</ta></tok><pt></pt>
<tok type="cyrw">фахівців<aa> N0pg;N0pa;Amsn0;Amsa0;Amsv0</aa>
<ta>N0pg</ta></tok><pt></pt>
<tok type="cyrw">не<aa>Q</aa><ta>Q</ta></tok><pt></pt>
<tok type="cyrw">таємниця<aa>Nfsn</aa><ta>Nfsn</ta></tok><pt>,</pt><pt></pt>
<tok type="cyrw">що<aa>Q;C;P00n;P00a</aa><ta>C</ta></tok><pt></pt>
<tok type="cyrw">нове<aa>Ansn0;Ansa0;Ansv0</aa><ta>Ansn0</ta></tok><pt></pt>
<tok type="cyrw">пенсійне<aa>Ansn0;Ansa0;Ansv0</aa><ta>Ansn0</ta></tok><pt></pt>
<tok type="cyrw">законодавство<aa>Nnsn;Nnsa;Nnsv</aa><ta>Nnsn</ta></tok><pt></pt>
<tok type="cyrw">ухвалювалося<aa>Vais3sn</aa><ta>Vais3sn</ta></tok><pt></pt>
<tok type="cyrw">в<aa>S</aa><ta>S</ta></tok><pt></pt>
<tok type="cyrw">Україні<aa>Nfsd;Nfsp</aa><ta>Nfsp</ta></tok><pt>,</pt><pt></pt>
<tok type="cyrw">так<aa>C;Q;R0</aa><ta>Q</ta></tok><pt></pt>
<tok type="cyrw">би<aa>Q</aa><ta>Q</ta></tok><pt></pt>
<tok type="cyrw">мовити<aa>Fi</aa><ta>Fi</ta></tok><pt>,</pt><pt></pt>
<tok type="cyrw">на<aa>S</aa><ta>S</ta></tok><pt></pt>
<tok type="cyrw">виріст<aa>Nmsn;Nmsa</aa><ta>Nmsa</ta></tok><pt>.</pt></s></article>
</text>
```

Рис. 3. Приклад тексту з усуненою неоднозначністю

Висновки

Сформульовано проблему усунення неоднозначності на морфолого-синтаксичному рівні. Для зняття неоднозначності запропоновано три підходи, два з яких мінімізують роботу експерта (застосування ланцюгів Маркова та трансформаційного навчання), а третій орієнтований на створення правил/евристик. Оскільки не існує достатньо великих корпусів текстів для української мови для того, щоби застосовувати статистичні методи, остаточно було запропоновано евристики для зняття внутрішньопарадигматичної та лексико-граматичної омонімії. Варто також наголосити, що у дослідженні використовувався набір позначок (теґсет), запропонований для створення українського національного корпусу. Також було порівняно цей теґсет з наборами позначок для інших слов'янських мов. Попри незначний розмір корпусу текстів, що використовувався у цьому дослідженні, ми плануємо з ростом розміру анотованого корпусу застосовувати статистичні методи для усунення неоднозначності на морфолого-синтаксичному рівні.

1. Єрмоленко С. Я., Бибик С. П., Тодор О. Г. Українська мова: Короткий тлумачний словник лінгвістичних термінів. – К., Либідь. 2001. 2. Карамишева Н.В. Логіка. Пізнання. Евристика. — Львів: Астролябія, 2002. 3. Кушлик О.П. Омонімія незмінних класів слів. Дис. ... канд. філолог. наук

зі спеціальності 10.02.01 – українська мова. Дрогобицький державний педагогічний університет імені І. Франка. – Дрогобич, 2000. 4. Лопатин В.В. Грамматическая категория // Лингвистический энциклопедический словарь. – М., 1990. 5. Мельчук И.А. Два оператора установления соответствия (для автоматического синтаксического анализа) . — М.: АН СССР, Сектор структурной и прикладной лингвистики, 1961. 6. Перебийнос В.И., Грязнухина Т.А. и др. Морфологический анализ научного текста на ЭВМ / АН УССР, Ин-т языковедения им. А.А. Потебни. — Киев: Наук. думка, 1989. — 264 с. 7. Сучасна українська мова / За ред. О.Д. Пономарева. – К., Либідь. 2001. 8. Th. Brants, O. Plaehn. Interactive Corpus Annotation. In Second International Conference on Language Resources and Evaluation LREC-2000. Athens, Greece. 2000. 9. E. Brill. Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging. 3rd Workshop on Very Large Corpora, 1995. 10. Ł. Dębowski. Tagowanie i disambiguacja morfosyntaksyczna. Przegląd metod i oprogramowania. Nr.934. Warszawa, listopad 2001. http://www.ipipan.waw.pl/~ldebowsk/raporty/ kropka934.pdf. 11. Ł. Dębowski. Trigram morphosyntactic tagger for Polish. 2004. http://www.ipipan. waw.pl/~ldebowsk/artykuly/ iipwm2004_debowski.pdf. 12. G. Greffenstette, P. Tapanainen. What is a word, what is a sentence? Problems of Tokenization. In Proceedings of the 3rd Conference on Computational Lexicography and Text Research (COMPLEX'94). Budapest, April 22, 1994. 13. Hajic J. Positional Tags: Quick Reference (Czech "HM" Morphology). – http://ucnk.ff.cuni.cz, 2000. 14. Hajič J., Hladká B. Tagging Inflective Languages: Prediction of Morphological Categoriesfor a Rich, Structured Tagset. 15. Hinrichs E.W., Trushkina J.S. Forging Agreement: Morphological Disambiguation of Noun Phrases. Proceedings of The First Workshop on Treebanks and Linguistic Theories (TLT'2002). Sozopol, Bulgaria, 2002. 16. Hladká B. Czech language tagging. Doctoral thesis. Institute of Formal and Applied Linguistics. Faculty of Mathematics and Physics. Charles University, Prague. 2000. 17. Ivanova K., Doikoff D. Cascaded regular grammars and constraints over morphologically annotated data for ambiguity resolution. Proceedings of The First Workshop on Treebanks and Linguistic Theories (TLT'2002). Sozopol, Bulgaria, 2002. 18. Katrenko S. Towards Unsupervised Learning of Morphology Applied to Ukrainian. The 16th European Summer School in Logic, Language and Information (ESSLLI). Nancy, France, pp. 138-149, 2004. 19. Katrenko S. Some peculiarities of lexical-grammatical ambiguity in Ukrainian and German. In CADSM'2003. 20. Márquez L. et al. A Machine Learning Approach to POS Tagging. Machine Learning Journal, Vol.39, n.1, April 2000. 21. Oflazer K., Tür G. Using Multiple Sources of Information for Constraint-based Morphological Disambiguation. 22. Oliva K., Petkevič V. Morphological and syntactic tagging of Slavonic languages. Handouts within the school „Empirical Linguistics and Natural Language Processing". Sozopol, Bulgaria, 2002. 23. Skut W., Krenn B., Brants Th., Uszkoreit H. An Annotation Scheme for Free Word Order Languages. In Proceedings of the Fifth Conference on Applied Natural Language Processing. Washington, U.S., 1997. 24. Dan Tufiş. Tiered Tagging and Combined Language Models Classifiers. In TSD'99, pp. 28-33, 1999. 25. Woliński M., Przepiórkowski A. – Projekt anotaciji morfosyntaksycznej korpusu języka polskiego. – Warszawa: IPI PAN, 2001. 26. XML. Extensible Markup Language (XML) 1.0 (Second Edition). W3C Recommendation. http://www.w3.org/TR/REC-xml, 2000.