УДК 81'373.423

Г. Б. Буньо,
Національний університет «Львівська політехніка», м. Львів

СУЧАСНІ МЕТОДИ ВИРІШЕННЯ ПРОБЛЕМИ ГРАМАТИЧНОЇ ОМОНІМІЇ В ТЕКСТІ

У статті розглянуто явище граматичної омонімії, а саме її різновид – омонімію морфологічну, з позицій текстоцентричного підходу. Проаналізовано основні підходи, досвід та перспективи вирішення цієї проблеми у процесі автоматичного морфологічного аналізу тексту, зокрема для української та інших мов зі складною морфологією.

Ключові слова: *морфологічна омонімія, автоматичний морфологічний аналіз, словоформа, уоднозначнення, ймовірнісні методи, методи на основі правил.*

В статье рассматривается явление грамматической омонимии, а именно один из её подвидов – морфологическая омонимия, с позиции текстоцентрического подхода. Проанализированы основные подходы, опыт и перспективы решения проблемы грамматической неоднозначности в процессе автоматического морфологического анализа текста, в частности для украинского языка, а также других языков со сложной морфологией.

Ключевые слова: *морфологическая омонимия, автоматический морфологический анализ, словоформа, деомонимизация, вероятностные методы, методы на основе правил.*

The article studies the phenomenon of grammatical homonymy, namely the morphological homonymy, from the text-centered perspective. The main approaches, experience, and prospects for solving the issue of grammatical ambiguity in the process of automatic morphological analysis are considered, notably in terms of Ukrainian and other morphologically complex languages.

Key words: *morphological homonymy, automatic morphological analysis, word form, disambiguation, stochastic methods, rule-based methods.*

Вступ. Явище багатозначності властиве кожній мові та існує на всіх рівнях системи мови. Неоднозначність мовного знака розглядали здебільшого крізь призму словоцентричного (лексикографічного) сприйняття, натомість текстоцентричний (функційний) підхід залишався значно поза увагою. Скажімо, до загальномовних академічних словників української мови входять оморяди базових форм слів, тому «Словник омонімів української мови» О. Демської та І. Кульчицького, у якому матеріал дібрано зі згаданих словників, подає омоніми за тим самим принципом [3, с. 12]:

БІГУН І, á, ч., заст. 1. Полюс. 2. Щось цілком протилежне.

БІГУН ІІ, á, ч. Спаровані камені у дробильній машині.

БІГУН ІІІ, á, ч. Вісь у дверях, воротях.

Як зазначає Е. Аврам'юк, яка досліджувала польську міжпарадигматичну омонімію, її «важко описувати за допомогою традиційних лексикографічних підходів, бо прийнятий у них спосіб презентації мовного матеріалу подає лише омонімію назв лексем» [15, с. 9].

Із появою систем автоматичного опрацювання тексту, великих текстових корпусів та необхідності аналізу тексту на різних мовних рівнях функційний аспект набув особливої актуальності. Це зумовлено тим, що змінне слово у тексті репрезентовано тільки у вигляді певної словоформи [11, с. 318], тобто під час граматичного аналізу тексту ми фактично маємо справу із граматичними формами слів у контексті (словоживаннями), а не вихідними (лексикографічними, канонічними) формами слів. Вирішення проблеми неоднозначності лексико-граматичного трактування реалізованого в контексті слова є одним із найактуальніших завдань прикладної лінгвістики.

Сучасні системи опрацювання тексту передбачають такі етапи аналізу тексту, як попередній поділ на слова й речення, граматичний та семантичний аналіз. Граматичний аналіз складається з морфологічного й синтаксичного. Відповідно, у сфері автоматичного опрацювання тексту найбільшої уваги приділяють двом видам неоднозначності: неоднозначності слів (лексичної та граматичної) та неоднозначності синтаксичних конструкцій. Нездатність системи правильно граматично проінтерпретувати одиницю тексту на морфологічному рівні виявляється в неправильних результатах аналізу на вищих рівнях, і зрештою до некоректних результатів роботи систем, де використовується граматичний аналіз тексту, а саме: машинного перекладу, контент-аналізу, систем перевірки правопису, лексикографічних, корпусних досліджень тощо. Тому після морфологічного аналізу тексту в системах опрацювання природної мови потрібне граматичне уоднозначнення або деомонізація [2, с. 285], результатом чого будуть правильно визначені граматичні характеристики словоформ у тексті.

Власне ми розглядатимемо граматичну неоднозначність із позицій текстоцентричного підходу. Граматичними омонімами вважатимемо формально тотожні граматичні форми або конструкції, що мають різне граматичне значення [9, с. 9]. І відповідно, граматичну багатозначність словоформ у тексті називатимемо граматичною омонімією або вужче, на рівні словоформ – морфологічною омонімією (омоформією).

Ще в другій половині минулого століття вийшло друком «Типологічне дослідження морфологічної омонімії різних мов» чеського лінгвіста Ї. Крамського [22]. Граматичну омонімію розглядали у своїх працях чеська дослідниця М. Тєшітелова, польські лінгвісти Д. Бутлер, К. Вашакова, Е. Аврам'юк, У. Андреєвич, М. Маєвська. Серед українських мовознавців проблеми граматичної омонімії досліджували Л. Кіцила, І. Данилюк (синкретизм), А. Лучик (вигуків), О.Бугаков (прийменників), О.Кушлик (незмінних класів слів), Н. Борисенко (в сучасній іспанській мові), Н. Глібчук та О. Шипнівська (міжчастиномовну морфологічну омонімію).

Метою нашого дослідження є окреслення проблематики одного з видів мовної неоднозначності в тексті – морфологічної омонімії, огляд основних підходів до граматичного уоднозначнення у процесі автоматичного

морфологічного аналізу, досвіду їх застосування та дослідження можливостей вирішення цієї проблеми для української, польської, чеської та інших флективних мов.

Морфологічна омонімія у тексті

Під час аналізу тексту засобами опрацювання природної мови кожній текстовій одиниці присвоюють набір граматичних характеристик. Морфологічний аналіз передбачає визначення щонайменше частини мови, а також значень відповідних граматичних категорій (грамем) і базової словоформи (леми) [4, с. 43]. Якщо у процесі автоматичного морфологічного аналізу не вдається однозначно проінтерпретувати словоформу, пропонується низка варіантів можливих лем і граматичних характеристик.

Приклад 1: *Пихнув два **рази** димом, смачно затягся, прислухаючись, як у голові колобродить легкий туманець.* (В. Шкляр, «Залишенець»)

Таблиця 1.

Варіанти граматичної інтерпретації слова «рази» у реченні

Лема	Граматичні характеристики
раз	Іменник, загальна назва, чол. рід, множина, Н.в., неістота
раз	Іменник, загальна назва, чол. рід, множина, З.в., неістота
раз	Іменник, загальна назва, чол. рід, множина, К.в., неістота
разити	Дієсл., основне, недок. вид, наказ. форма, 2-а особа одн.

У перших трьох варіантах відрізняється значення відмінка, усі три форми належать до парадигми однієї лексеми і творять т.зв. внутрішньопарадигматичну (внутрішньочастиномовну) омонімію. Дієслівна форма творить з ними міжпарадигматичну міжчастиномовну омонімію.

Приклад 2: *Високий статистичний показник рівня забезпеченості навчальних закладів комп'ютерами не відображає реального **стану** речей.*

Таблиця 2.

Варіанти граматичної інтерпретації слова «стану» у реченні

Лема	Граматичні характеристики
стан	Іменник, загальна назва, чол. рід, однина, Д.в., неістота
стан	Іменник, загальна назва, чол. рід, однина, Р.в., неістота
стан	Іменник, загальна назва, чол. рід, однина, М.в., неістота
станути	Дієсл., основне, док. вид, дійсна форма, майб. час, 1-а ос. одн.
стати	Дієсл., основне, док. вид, дійсна форма, майб. час, 1-а ос. одн.

Перші три та два останні слововживання мають однакову частиномовну належність, тому є внутрішньочастиномовними омонімами. В іменникових словоформ відрізняються лише грамеми категорії відмінка, отже, це внутрішньопарадигматичні омоніми. У двох останніх словоформ лексико-граматичне значення збігається, але відрізняються леми – це омоніми внутрішньочастиномовні міжпарадигматичні. Розрізнення лексем у випадку дієслів відбувається на рівні лематизації – визначення базової словникової форми слова [17, с. 381]. Проте цьому передує визначення відповідного набору граматичних характеристик.

Позаяк система морфологічного аналізу розглядає кожне слово ізольовано, вирішити таку неоднозначність на цьому рівні неможливо, треба звернутися до контексту [23, с. 744]. Це завдання можна вирішити вручну, бо зазвичай людина легко визначає граматичне значення слів у контексті. У системі морфологічного аналізу текстів українською мовою UGTag паралельного польсько-українського корпусу PolUKR передбачено таке ручне уоднозначнення [20]. Однак якщо брати до уваги обсяги сучасних корпусів текстів, якими послуговуються сучасні мовознавчі дослідження (малі – до 1 млн. слововживань, великі – понад 100 млн.) [5], усунення такої неоднозначності вручну – завдання надто трудомістке, тому з'являється все більше напрацювань щодо автоматизації цього процесу.

Частково завдання морфологічного уоднозначнення для української мови вирішував О. Бугаков під час дослідження функціонування прийменників у тексті. Було створено алгоритм встановлення текстових умов зняття функціональної омонімії, коли одним із компонентів омоніма є прийменник; за допомогою дистрибутивного методу знято граматичну омонімію прийменників з іншими граматичними класами [1, с. 62–68, 203–214].

О. Шипнівська в межах свого дослідження сформувала лексикографічні бази даних міжчастиномовних омонімів та лінгвістичну базу даних для дослідження контекстів, де актуалізуються ті чи інші значення омонімічних одиниць; це стало «основою для розробки правил автоматичного усунення міжчастиномовної морфологічної омонімії» в українській мові [12, с. 4]. Обидві праці – на базі Українського національного лінгвістичного корпусу, створеного в Українському мовно-інформаційному фонді НАНУ [10]. В обмежено доступній версії цього корпусу морфологічне уоднозначнення практично відсутнє, як і для українських текстів польсько-українського паралельного корпусу (PolUKR) [7; 8]. У доступному в мережі «Корпусі Української Мови» граматичне уоднозначнення потребує значного доопрацювання [7; 21].

Огляд підходів до граматичного уоднозначнення

Відповідно до способу отримання контекстної інформації, на основі якої відбувається розрізнення слів, виокремлюють такі основні підходи до граматичного уоднозначнення у процесі автоматичного морфологічного аналізу [29, с. 108; 17, с. 381; 16, с. 95]:

– ймовірнісні (англ. «stochastic», «probabilistic»; також: статистичні, на основі машинного навчання):

– контрольованого навчання (супроводжуваного людиною, англ. «supervised);

– машинного самонавчання (без супроводу людини, англ. «unsupervised»);

– на основі правил, розроблених вручну (надалі: «на основі правил», англ. «rule-based»; не плутати з правилами, сформованими на основі машинного навчання).

На думку М. Тєшітелової, «морфологічну омонімію слід досліджувати як у системі мови, так і в контексті, використовуючи статистичні методи, оскільки йдеться про омонімію як фактор економії, який необхідно характеризувати квантитативно» [6, с. 81].

Статистичні методи дають змогу обрати найвірогіднішу граматичну інтерпретацію словоформи в контексті на основі статистичних даних.

Система машинного навчання аналізує певний текст, тренується на ньому, розпізнає деякі закономірності і на їх основі може робити певні узагальнення. Згідно з набутою інформацію про закономірні властивості словоформ у всіх контекстах, які система проаналізувала, вона може робити прогнози щодо найвірогіднішої граматичної інтерпретації словоформи у нових текстах.

Методи контрольованого машинного навчання роблять ймовірнісний прогноз після тренування на корпусі текстів повністю або частково розміченого інформацією про словоформи, а методи машинного самонавчання дозволяють працювати з нерозміченим корпусом [31].

Системи морфологічного аналізу на основі цього методу створено для англійської мови (Brill tagger, RDRPOSTagger) та адаптовано до деяких флективних мов [14; 28]. Ці аналізатори використовують методику трансформаційних правил, виведених в результаті машинного навчання. Загалом точність уоднозначнення в корпусах текстів англійської мови ймовірнісними аналізаторами перевищує 97% [19, с. 2949].

Застосування ймовірнісних методів не потребує великих зусиль, якщо дослідник знається на методології. Залучення великих за обсягом ресурсів робить їх досить ефективними, принаймні для англійської мови. Вони дозволяють зменшити відсоток неоднозначності, але не вирішують проблеми повністю, особливо для мов з відносно вільним порядком слів у реченні.

Методи на основі правил полягають у використанні вручну розроблених та формалізовано представлених правил, обмежувальної граматики, у якій кожне правило на основі контекстного оточення неоднозначної словоформи дозволяє або унеможливлює присвоєння їй певної грамеми. Правила можуть застосовуватися циклічно багато разів до максимального зменшення кількості варіантів граматичної інтерпретації словоформи. Перший великий аналізатор на основі контекстних правил TAGGIT містив 3300 правил і досягав уоднозначнення у Браунівському корпусі точністю 77%.

Методи на основі правил не поступаються статистичним, і навіть краще відповідають мовам із вільним порядком слів у реченні [25, p. 26–44; 17, с. 381; 24]. Щоби скористатися цим методом, потрібна інформація щодо типів і моделей морфологічних омонімів та особливості їх функціонування в тексті. Варто зважати на те, що:

– цей метод вимагає компетенції і часто значних зусиль від розробника;

– можливе замкнене коло, коли для морфологічного уоднозначнення потрібна інформація синтаксична чи лексична, а водночас для синтаксичного чи лексичного уоднозначнення потрібна морфологічна;

– характер проблеми й обсяг завдання залежить також від ефективності роботи самого граматичного аналізатора [27]; якщо існує кілька аналізаторів для мови, то варто спробувати їх спільне використання, принаймні для польської мови такий досвід був успішний [19].

Отже, добре написані правила уоднозначнення можуть значно покращити результати деомонізації [25]. Але позаяк цілу граматику часто розробити складно, цей метод доцільно поєднувати з ймовірнісним [17, с. 384]: скажімо, спробувати отримати правила уоднозначнення за допомогою статистичних методів і розробити правила для випадків, яких не було враховано за допомогою статистичних методів [26, с. 152]. Експериментально доведено, що результати комбінації методів граматичного уоднозначнення кращі, ніж використання лише одного з них, зокрема, статистичного. [18; 30; 26].

У системі морфологічного аналізу UGTag для української мови передбачено прості правила для автоматичного уоднозначнення прийменників на основі статистичного аналізу. Надалі розробники планують поєднати правила і статистичний аналіз уоднозначнених даних [20].

Актуальним залишається питання, які типи текстової інформації можуть бути стрижневими для вирішення граматичної неоднозначності, і наскільки близько в тексті до цільового слова вони мають бути. О. Бугаков та О. Шипнівська довели, що для контекстуального уоднозначнення важливо визначити типи та моделі омонімії та їх функціонування в контексті. У дослідженні О. Шипнівської розпізнавальними для усунення міжчастиномовної омонімії були граматичний і лексико-граматичний контексти [13, с. 42]. О. Бугаков виявив, що стиль тексту може враховуватися при укладанні правил визначення значення омографа в контексті [1, с. 53].

Відкритим є питання вибору методів вирішення лексичної багатозначності та їхнього оцінювання, зокрема вибору оптимального варіанту для української мови в умовах, коли практичних розробок дослідників немає у відкритому доступі, або їх не впроваджено в загальнодоступні ресурси, де передбачено автоматичне опрацювання тексту. Перспективними з огляду на позитивний досвід використання для флективних мов є методи на основі правил та поєднання підходів до граматичного уоднозначнення словоформ.

Висновки. Рівень складності завдання граматичного уоднозначнення текстових словоформ та вибір методології уоднозначнення значно залежить від типологічних ознак мови й наявних мовно-інформаційних та людських ресурсів. Для високофлективних мов завдання ускладнюється через багату словозміну, а також відносно вільний порядок слів у реченні. Підтверджено важливість подальшого дослідження морфологічної парадигматики та її зв'язку з іншими мовними рівнями задля простеження закономірностей, які допоможуть з вибором способу уоднозначнення. На нашу думку, варто застосовувати методи граматичного уоднозначнення на основі правил, а також комбінувати ці методів зі статистичними.

Література:

1. Бугаков О. В. Функціонування прийменників в українському тексті: морфологічний та семантико-синтаксичний аспекти.: дис. … канд. філолог. наук : 10.02.01 / О. В. Бугаков. – Київ, 2005. – 234 с.

2. Великий тлумачний словник сучасної української мови (з дод. і допов.) / Укл. і гол. ред. В. Т. Бусел. – К. ; Ірпінь : ВТФ «Перун», 2005. – 1728 с.

3. Демська О. Словник омонімів української мови / О. Демська, І. Кульчицький. – Л. : Фенікс, 1996. – 224 с.

4. Демська-Кульчицька О. М. Базові поняття корпусної лінгвістики / О. М. Демська-Кульчицька // Українська мова. – 2003. – №1 (6). – С. 40–45.

5. Демська-Кульчицька О. М. Корпуси мов і один із можливих підходів до проектування корпусу текстів української мови / О. М. Демська-Кульчицька // Лінгвістичні студії: Збірник наукових праць. – Вип. 13. – Д. : ДонНу, 2005. – С. 8–16.

6. Кіцила Л. Чеські та словацькі мовознавці про омонімію / Лідія Кіцила // Проблеми слов'янознавства. – Л., 1999. – Вип. 50. – С. 77–85.

7. Корпус української мови Лабораторії комп'ютерної лінгвістики КНУ ім. Т. Шевченка [Електронний ресурс]. – Режим доступу : http://www.mova.info/corpus.aspx?l1=209

8. Коциба Н. Морфосинтаксичне тагування польсько-українського паралельного корпусу (PolUKR) [Електронний ресурс] / Н. Коциба // Proceedings of the International Conference «MegaLing'2008. Horizons of Applied Linguistics and Linguistic Technologies. – Kyiv, 2009. – Режим доступу : http://www.domeczek.pl/~natko/papers/megaling2008.pdf

9. Смірнова Є. С. Англійська мова : навч.-метод. посіб. / Є. С. Смірнова, Г. А. Чередніченко. – К. : НУХТ, 2011. – 180 с.

10. Український національний лінгвістичний корпус УМІФ НАНУ [Електронний ресурс]. – Режим доступу : http://lcorp.ulif.org.ua/virt_unlc/

11. Цоуфал Л. С. Лінгвістичні засади навчання морфології в загальноосвітній школі / Л. С. Цоуфал // Філологічні студії: наук. вісн. Криворіз. нац. ун-ту : зб. наук. пр. – Кривий Ріг, 2012. – Вип. 7, Ч. 2 – С. 315–322.

12. Шипнівська О. О. Структурно-семантичні та функціональні характеристики міжчастиномовної морфологічної омонімії сучасної української мови : автореф. дис. … канд. філол. наук : спец. 10.02.01 «Українська мова» / О. О. Шипнівська – Київ, 2007. – 19 с.

13. Шипнівська О. О. Структурно-семантичні та функціональні характеристики міжчастиномовної морфологічної омонімії сучасної української мови : дис. … канд. філол. наук : спец. 10.02.01 / О. О. Шипнівська. – Київ, 2007. – 236 с.

14. Acedański S. A Morphosyntactic Brill Tagger for Inflectional Languages / Szymon Acedański // Advances in Natural Language Processing. – Berlin : Springer Berlin Heidelberg, 2010. – P. 3–14.

15. Awramiuk E. Systemowość polskiej homonimii międzyparadygmatycznej. – Białystok : Wydaw. Uniwersytetu w Białymstoku, 1999. – 242 s.

16. Buitelaar P. Linguistic Annotation for the Semantic Web [Electronic Resource] / Paul Buitelaar, Thierry Declerck // Annotation for the Semantic Web : 96 of Frontiers in Artificial Intelligence and Applications. – 2003. – Mode of access : http://books.google.com.ua/books?id=JMw8Y897c7MC

17. Ezeiza N. Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages / N. Ezeiza, I. Alegria, J.M. Arriola, R. Urizar, I. Aduriz // ACL '98 Proceedings of the 36th Annual Meeting of the Association for Comput. Linguistics and 17th International Conference on Comput. Linguistics. – 1998. – Vol. 2. – P. 379–384.

18. Hulden M. Boosting Statistical Tagger Accuracy with Simple Rule-Based Grammars / M. Hulden, J. Francom // Proceedings of the LREC 2012. – Istanbul, 2012. – P. 2114–2117.

19. Kobyliński Ł. PoliTa: A multitagger for Polish / Łukasz Kobyliński // Proceedings of the LREC 2014. – Reykjavík : ELRA, 2014. – P. 2949–2954.

20. Kotsyba N. UGTag: morphological analyzer and tagger for the Ukrainian language [Electronic Resource] / N. Kotsyba, A. Mykulyak, I. Shevchenko // Explorations across languages and corpora: PALC 2009. – Lodz, Łódź studies in language. – v. 24. – 2009. – Mode of access : http://www.domeczek.pl/~natko/papers/PALC-2009_UGTag.pdf

21. Kotsyba, N. Praktyczny przewodnik po korpusach języka ukraińskiego [Electronic Resource] / N. Kotsyba // Praktyczny przewodnik po korpusach języków słowiańskich. – Warsaw, 2013. – Mode of access : http://www.domeczek.pl/~natko/papers/przewodnik-korp-ukr2013.pdf

22. Krámský J. A typological study of morphological homonymy in languages // Papers in General Linguistics. – The Hague: Mouton, 1974. – P. 156–180.

23. Lezius W. A Freely Available Morphological Analyzer, Disambiguator and Context Sensitive Lemmatizer for German / W. Lezius, R. Rapp, M. Wettler // P. of the 36th Annual Meeting of the Assoc. for Comp. Ling. and 17th International Conf. on Comp. Ling. – 1998. – Vol. 2. – P. 743–748.

24. Oliva K. The Linguistic Basis of a Rule-Based Tagger of Czech / K. Oliva, M. Hnátková, V. Petkevic, P. Kveton // Text, Speech and Dialogue: Third International Workshop, TSD 2000. – Brno: Springer, 2000. – P. 3–8.

25. Petkevič V. Reliable Morphological Disambiguation of Czech: a Rule-Based Approach is Necessary / V. Petkevič // Insight into the Slovak and Czech Corpus Linguistics (Šimková M. ed.). – Veda, Bratislava, 2006. – P. 26–44.

26. Piasecki M. Polish Tagger TaKIPI: Rule Based Construction and Optimisation / Maciej Piasecki // Proceedings of the SIIS'11. – Berlin, Heidelberg: Springer-Verlag, 2012. – P. 359–369.

27. Radziszewski A. Taggers gonna tag: an argument against evaluating disambiguation capacities of morphosyntactic taggers [Electronic Resource] / Adam Radziszewski; Szymon Acedański // Proceedings of TSD 2012. – Mode of access : http://nlp.pwr.wroc.pl/ltg/files/publications/taggereval.pdf

28. RDRPOSTagger: A Ripple Down Rules-based PoS Tagging Toolkit [Electronic Resource]. – Mode of access : http://rdrpostagger.sourceforge.net/

29. Sak H. Morphological Disambiguation of Turkish Text with Perceptron Algorithm / H. Sak, T. Güngör, M. Saraçlar // CICLing 2007 Conference. – Mexico City, Mexico, February 18–24, 2007. – P. 107–118.

30. Spoustová D. Combining Statistical and Rule-Based Approaches to Morphological Tagging of Czech Texts / D. Spoustová // The Prague Bulletin of Mathematical Linguistics № 89 – 2008. – P. 23–40.

31. Yatbaz M. Unsupervised Morphological Disambiguation Using Statistical Language Models [Electronic Resource] / M. A. Yatbaz, D. Yuret // NIPS 2009. – Whistler, 2009. – Mode of access : www0.cs.ucl.ac.uk/staff/rmartin/grll09/yat1.pdf