Ольга Шипнівська, Сергій Стариков*
Український мовно-інформаційний фонд НАН України (Київ)
УДК 81'322.5

ФОРМУВАННЯ ЛІНГВІСТИЧНОЇ БАЗИ ДАНИХ ГРАМАТИЧНИХ ОМОНІМІВ НА МАТЕРІАЛІ УКРАЇНСЬКОГО НАЦІОНАЛЬНОГО ЛІНГВІСТИЧНОГО КОРПУСУ

У статті розглядаються принципи побудови лінгвістичної бази даних граматичних омонімів на основі Українського національного корпусу текстів. Головна увага зосереджена на описі структури та принципів функціонування бази.

При розробці систем автоматичного аналізу тексту особливого значення набуває проблема автоматичного усунення омонімії. Поставлена ще в минулому столітті, ця проблема потребує нагального розв'язання на сучасному етапі розвитку корпусної лінгвістики [1; 2; 3; 4; 5; 6]. Успішність вивчення прикладного аспекту граматичної омонімії значною мірою залежить від наявності лінгвістичних ресурсів, зокрема лінгвістичних баз даних. Виникає необхідність розробки ефективних методик, які б ураховували сучасний стан лінгвістичних знань та комп'ютерних технологій.

У межах корпусної лінгвістики можна виділити два підходи щодо автоматизованого зняття морфологічної неоднозначності. Використовуючи для прогнозування рішень метод статистико-дистрибутивного аналізу, розробники цих підходів послуговуються двома типами спеціальних лінгвістичних текстових корпусів, що формуються на загальному корпусі текстів: тестовим корпусом зі знятою вручну омонімією та тестовим корпусом із потенційною омонімією [6].

Розробка алгоритмів автоматичного зняття морфологічної неоднозначності для лінгвістичного процесора передбачає накопичення максимально повної інформації про потенційні можливості словникового складу мови стосовно прояву омонімії та мовленнєвої характеристики явища. Так, поява інтернет-ресурсу Ж. Г. Аношкіної [3], у якому була здійснена спроба зібрати всі типи граматичних омонімів для російської мови на матеріалі граматичного словника А. А. Залізняка, стимулювала низку прикладних та теоретичних досліджень у напрямку граматичної омонімії [5; 6].

Створення лінгвістичної бази даних для дослідження морфологічної омонімії сучасної української мови, що здійснюється в Українському мовно-інформаційному фонді НАН України, пріоритетним робить залучення значних масивів лінгвістичного матеріалу як словникового підкорпусу УНЛК, так і текстового підкорпусу [1]. Створені на цьому матеріалі ЛБД морфологічних омонімів практично реалізують ідею інтегрального опису мови, що ґрунтується на врахуванні зв'язку між словником, граматикою і текстом.

Такий підхід опису явища омонімії передбачає врахування специфіки мови як абстрактної системи мовних знаків та мовлення як конкретної реалізації цієї системи. Як власне мовно-мовленнєве співвідношення у нашій роботі розглядаємо протиставлення потенції та її реалізації. Мовним аспектом явища морфологічної омонімії вважається система потенційних можливостей типів морфологічних омонімів, виявлених на основі граматичного словника. Мовленнєвий аспект досліджуваного явища полягає в дослідженні реалізації цих можливостей в текстах сучасної української мови.

Лінгвістична база даних у контексті проблем морфологічної омонімії повинна: представляти реєстр морфологічно неоднозначних словоформ сучасної української мови; подавати моделі та типи морфологічних омонімів, яким відповідають конкретні

словоформи; демонструвати частотні характеристики типів, моделей, конкретних омонімів у стилістично диференційованих текстах; містити відповідні кожній із моделей контексти її реалізації в тексті. Побудовані ЛБД слугують дослідникам як інформаційні системи, що розраховані на широке коло запитів стосовно явища граматичної омонімії української мови.

Автоматична побудова таких баз стала можливою завдяки використанню: програми автоматичного морфологічного аналізу; словника словоформ, створеного на основі електронного граматичного словника української мови, розроблених в Українському мовно-інформаційному фонді НАН України; стилістично диференційованого підкорпусу текстів, сформованого на основі Українського національного лінгвістичного корпусу.

Для дослідження граматичної омонімії було створено три бази даних: база даних потенційних типів та моделей омонімів української мови на основі словника словоформ, база даних виділених на основі корпусу текстів типів та моделей, та база даних текстових контекстів для кожного з цих типів.

Потенційні типи морфологічних омонімів української мови, що задаються її граматичною системою, кількісні та структурні характеристики цих типів визначалися на матеріалі словника словоформ, який був сформований на основі розробленого в УМІФі граматичного словника[1]. У результаті автоматичної морфологічної розмітки з 3.094.174 (171.620 лексем) словоформ словника 1.554.372 словоформам було поставлено у відповідність омонімічні коди за допомогою програми автоматичного морфологічного аналізу. На основі цього списку було виділено 610 612 омонімічних рядів (ОР).

У списку морфологічно неоднозначних одиниць словника (610 612 омографів) 98 % (600 397) становлять омографи в середині словозмінної парадигми лексико-граматичного класу, решта 2 % (10 215) – це морфологічні омографи між різними лексико-граматичними класам.

Список моделей морфологічних омонімів, представлених у словнику словоформ, формувався за списком попередньо виділених омонімічних рядів шляхом зведення до одного тих ОР, що мають однакові омонімічні коди. У результаті було одержано список з 1 486 різних структурних моделей граматичних омографів української мови.

Після розгляду отриманих моделей з погляду представлення ними типів морфологічних омонімів на рівні граматичних класів було виявлено, що 820 моделями (з 1486) описується міжчастиномовна омонімія. Дані моделі охоплюють 10 215 ОР і представляють 81 тип морфологічних омонімів[2]. Решта моделей представляє омонімію в середині одного лексико-граматичного класу.

На рис. 1. подано фрагмент таблиці морфологічних омонімів, що представляє типи міжчастиномовних морфологічних омонімів (ядро міжчастиномовних морфологічних омонімів).

Для аналізу функціонування морфологічних омонімів в текстах сучасної української мови в межах УНЛК було сформовано текстові вибірки наукового, публіцистичного, художнього стилів, кожна з яких містить по 1 000 000 словоформ. Принцип репрезентативності в цьому разі реалізується завдяки тому, що, по-перше, обрані для аналізу три стилі у своїй сукупності представляють сучасну українську мову і, по-друге,

[1] Див. про це докладно у роботі: Шипнівська О. О.Формування реєстру омографів сучасної української мови // Мовні і концептуальні картини світу: Зб. наукових праць. – Випуск 14. Книга 2. – К.: Вид. Дім Дмитра Бураго, 2004. – С. 255–259.

[2] Див. про це у роботі: Шипнівська О. О. Кількісна та якісна характеристика типів морфологічних омонімів сучасної української мови // Актуальні проблеми української лінгвістики: теорія і практика: Зб. Наукових праць. – К.: Видавничо-поліграфічний центр „Київський університет", 2005. – С. 36–42.

як показав попередній аналіз матеріалу, у сукупній вибірці у 3 000 000 словоформ він представляє 95,06 % усіх типів морфологічних омонімів, властивих українській мові.

Рис. 1. Таблиця типів міжчастиномовних морфологічних омонімів (фрагмент)[1]

Тип	Кількість омографів	Приклад
АІ	5682	вартові, виборні
ДІ	2406	коти, лови
АД	588	бадьорим, білим
ДТ	389	значимо, мислимо
ІТ	342	візаві, ажур
АТ	272	балакуче, захоплююче

У досліджуваних текстах сучасної української мови реалізовано 11,86 % (72 437) усіх можливих для української мови омографів, 58,15 % (5 950) міжчастиномовних омографів та 11,07 % омографів у середині однієї частини мови, 51,48 % (1 486) усіх моделей, 93,29 % (765) моделей міжчастиномовних омонімів та 71,17 % (666) моделей морфологічних омонімів усередині однієї частини мови.

У результаті аналізу одержаних вибірок з погляду співвідношення в них однозначних і омонімічних одиниць було визначено, що в текстах наукового стилю морфологічно неоднозначні словоформи становлять 67,01 % (тобто 656 303 словоформи) від загальної кількості словоформ вибірки. Для публіцистичного стилю цей показник становить 60,13 % (618 017 словоформ), для художнього – 55,11 % (560 632 словоформ).

На наступному етапі аналізу для кожної вибірок було визначено загальну кількість омонімічних рядів. Найбільша кількість ОР представлена в текстах публіцистичного стилю – 37 774. У науковому стилі їх 32 983, у художньому – 33 639.

Аналіз сукупностей ОР, виділених в аналізованих вибірках, з погляду представлення їх структурними моделями показав, що найбільшу кількість зі списку визначених потенційних моделей омонімів української мови реалізують художні тексти – 981, найменшу – наукові тексти – 752. Омонімічні ряди публіцистичного стилю представлені 869 моделями.

До баз морфологічних омонімів (МО) сучасної української мови, створених на основі граматичного словника та стилістично диференційованого корпусу текстів, було інтегровано лінгвістичну базу текстових контекстів міжчастиномовних омонімів. Остання в повному обсязі представляє контекстні оточення всіх моделей та типів морфологічних омонімів. Вихідним матеріалом для формування бази контекстів стали морфологічно розмічені корпуси текстів досліджуваних стилів. Під контекстом тут розуміємо лексико-граматичне оточення омографа в межах речення.

Довжина контексту визначалася сімома одиницями ліворуч та праворуч від омографа. При цьому враховувалися результати структурної розмітки, за якою розділові знаки розглядаються як окремі структурні елементи.

[1] У роботі використано граматичну класифікацію і принципи кодування граматичної інформації, прийняті в системі АМА УМІФу, де І – іменник, А – ад'єктив, Д – дієслово, Т – прислівник.

База даних лінгвістичних контекстів МО складається із 2 таблиць: таблиця лексико-граматичних контекстів міжчастиномовних морфологічних омонімів (табл. 1) і таблиця діагностуючих контекстів (табл. 2). Таблицю лексико-граматичних контекстів міжчастиномовних морфологічних омонімів структуровано за такими полями: "Омограф" (поле 1), "Модель омографа" (поле 10), "Лексико-граматичний контекст" (поле 2). Компоненти граматичного контексту представлено морфологічними кодами кожного із слів з лівого чи правого контекстів. Лівий контекст – поля 3–9, правий – 11–17. Поля таблиці заповнювалися автоматично.

На рис. 2 подано фрагмент таблиці лексико-граматичних контекстів міжчастиномовних морфологічних омонімів для моделі типу дієслово / іменник <UBMHMKMN> (дієсл., недок. виду, наказ. сп., активн. ст., 2 особа одн. / ім., заг., чол. р., Н. в., мн. / ім., заг., чол. р., З. в. мн. / ім., заг., чол. р., К. в. мн.).

Така форма репрезентації даних дозволяє за будь-яким параметром, відповідно до виділених полів, або за їх комбінаціями автоматично формувати необхідні субкорпуси контекстів.

На рис. 3 подано фрагмент таблиці діагностуючих контекстів.

Рис. 3. Таблиця діагностуючих контекстів (табл. 2)

Поле 1	Поле 2	Поле 3	Поле 4	Поле 5	Поле 6
Модель омографа	Діагностуючий контекст	Актуальне граматичне значення омографа	Довжина ДК ліворуч	Довжина ДК праворуч	Частота (%)
UBMHMKMN	<ASAV[1]> [[<UBMHMKMN>]]	Іменник	1	0	20,6
UBMHMKMN	<UA> [[<UBMHMKMN>]]	Іменник	1	0	0,6
UBMHMKMN	[[<UBMHMKMN>]] <FHFCFFFBFKFN>	Іменник	0	1	0,6

Для визначення правил, необхідних для побудови алгоритму усунення морфологічної омонімії в результаті опрацювання таблиці лексико-граматичних контекстів міжчастиномовних морфологічних омонімів, формувалася таблиця діагностуючих контекстів (ДК). Дослідником визначалися розміри контексту, який може бути достатнім для зняття омонімії, позиція цього компоненту стосовно омографа (лівого, правого, і лівого, і правого), конкретне граматичне значення ДК, а також актуальне граматичне значення омографа в даному контексті. Після такої лінгвістичної експертизи автоматично заповнювалася таблиця 2, яка складається з полів: "Модель омографа" (поле 1), "Діагностуючий контекст" (поле 2), "Актуальне граматичне значення омографа" (поле 3), "Довжина ДК ліворуч" (поле 4), "Довжина ДК праворуч" (поле 5), "Частота" (поле 6). У полі "Діагностуючий контекст" подається граматичний контекст, який є достатнім для зняття омонімії. Довжина контексту в полях "Довжина ДК ліворуч" та "Довжина ДК праворуч", визначається кількістю текстових одиниць у межах ДК омографа з урахуванням лівої та правої позиції. Актуальне граматичне значення, яке реалізується омографом, у цьому контексті записується в полі "Актуальне граматичне значення омографа". Відносна частота діагностуючого граматичного контексту в межах вибірки подано полі "Частота".

[1] Значення двосимвольних кодів у наведеному прикладі: AS – ад'єкт., мн., Н.в.; AV – ад'єкт., мн., Д.в..; UA –дієсл., інфін., недок. в., акт. стан; FH – ім. заг., жін.р., мн., Н.в.; FC – ім. заг., жін. р., одн., Д.в.; FF – ім. заг., жін. р., одн., К.в.; FB – ім. заг., жін. р., одн., Р.в.; FK – ім. заг., жін. р., мн., З.в. FN – ім. заг., жін. р., мн., К.в.

Література

1. Корпусна лінгвістика / За ред. В. А. Широкова. – К.: Довіра, 2005. – 471 с.
2. Морфологический анализ научного текста на ЭВМ. – К.: Наукова думка, 1989. – 264 с.
3. Аношкина Ж. Г. Словарь омонимичных словоформ русского языка. – М., 2001 http: //www.irlras-cfrl.rema.ru/homoforms/
4. Зинькина Ю.В., Пяткин Н.В., Невзорова О.А. Разрешение функциональной омонимии в русском языке на основе контекстных правил // http: // www.dialog-21
5. Кобзарева Т.Ю., Афанасьев Р.Н. Универсальный модуль предсинтаксического анализа омонимии частей речи в РЯ на основе словаря диагностических ситуацій // www.dialog-21
6. Ляшевская О. Н., Плунгян В. А., Сичинава Д. В. О морфологическом стандарте Корпуса современного русского языка // Научно-техническая информация. – Сер. 2. Информационные процессы и системы. – 2005. – № 6. – С. 2–9.