УДК 81'36. 373. 423:811.93

Г. В. Карнаух

ДЕЯКІ АСПЕКТИ ЗНЯТТЯ ГРАМАТИЧНОЇ НЕОДНОЗНАЧНОСТІ СЛОВОФОРМ У ТЕКСТАХ ХУДОЖНЬОГО СТИЛЮ (НА МАТЕРІАЛІ ПРОЗОВИХ І ПОЕТИЧНИХ ТВОРІВ)

Карнаух Г. В. Деякі аспекти зняття граматичної неоднозначності словоформ у текстах художнього стилю (на матеріалі прозових та поетичних творів).

У статті розглядаються якісні та кількісні характеристики даних, отриманих у результаті застосування алгоритму для автоматичного зняття неоднозначності словоформ, які виступають як власні і загальні назви.

Ключові слова: граматична неоднозначність словоформ, дизамбігуація (зняття неоднозначності словоформ), алгоритм, власна назва, загальна назва.

Карнаух А. В. Некоторые аспекты снятия грамматической неоднозначности словоформ в текстах художественного стиля (на материале прозаических и поэтических произведений).

В статье рассматриваются качественные и количественные характеристики данных, полученных в результате использования алгоритма для автоматического снятия неоднозначности словоформ, выступающих в качестве имен собственных и нарицательных.

Ключевые слова: грамматическая неоднозначность словоформ, дизамбигуация (снятие неоднозначности словоформ), алгоритм, имя собственное, имя нарицательное.

Karnaukh G. V. Some aspects of elimination of grammatical ambiguity separate word forms in belles-lettres texts (based on prose and poetical compositions).

This article examines qualitative and quantitative data obtained in the result of algorithm using for automatic elimination of grammatical ambiguity of separate word forms which appear in the text as proper and common names.

Key words: grammatical ambiguity, disambiguation (elimination of ambiguity of separate word forms), algorithm, proper name, common name.

Однією з унікальних властивостей мови як системи знаків є те, що більш ємнісний за обсягом і багатовимірний за структурною організацією план змісту не має однозначного відповідника плану вираження. Словесний знак є перш за все таким, що «прикладається» до низки однотипних, однорідних або ж аналогічних явищ [4, с. 13], тоді як у повністю формалізованій мові логіки питання про багатозначність відсутнє, оскільки в ідеальній моделі мови їй взагалі немає місця.

Асиметрія мовного знака є головною умовою його функціонування [10, с. 78]. Означувані й означувальні постійно рухаються у площині реальності. «Кожне виходить з меж, призначених йому партнером: означувальне прагне оволодіти інакшими функціями, аніж його власні; означуване прагне до того, щоб виразити себе інакшими засобами, аніж його власний знак. Вони асиметричні, будучи парними, опиняються у стані нестійкої рівноваги. Саме завдяки цьому асиметричному дуалізму структури знаків лінгвістична система може еволюціонувати: «адекватна» позиція знака постійно переміщується внаслідок пристосування до потреб конкретної ситуації» [3, с. 90].

Мова – складний, динамічний живий організм і не є тією ідеальною системою знаків, у якій всі елементи характеризуються однозначною відповідністю. Крім того, мові притаманний ряд властивостей і рис, що суперечать ідеальній моделі. Водночас, як інструмент для виконання складних функцій вона постійно розвивається і вдосконалюється. А феномен неоднозначності пояснюється перш за все тим, що мовний знак «завжди є акт мислення, а мислення безмежне вже через те, що воно є відображенням дійсності, також завжди безмежної» [8, с.123].

У зв'язку з цим Г. В. Колшанський наголошує на необхідності визнати, що багатозначність як у сфері лексики, так і у сфері граматики є необхідною якістю мови, зумовленою самою сутністю її матеріальної будови, самою природою людської мови: узагальнюючим характером мовного знака – слова і абстрактною сутністю граматичних категорій [4, с. 13]. Проте саме явище неоднозначності[1] є суттєвою перешкодою при автоматичній індексації, машинному перекладі, автокорекції і т. ін. Тому створення й удосконалення систем автоматичної обробки текстів, насамперед, передбачає автоматичне зняття неоднозначності словоформ, зокрема граматичної омонімії. В Українському мовно-інформаційному

[1] Неоднозначність, або ж багатозначність, трактується як родове поняття, що поєднує полісемію й омонімію [6, с. 93].

фонді НАН України (УМІФ) «активно ведеться розробка систем розуміння мови та природомовного людино-машинного діалогу, засобів міжмовної адаптації, систем семантичного аналізу тексту, а також інших інтелектуальних артефактів, орієнтованих на природну мову» [5, с. 3]. Запропонована в УМІФі комп'ютерна програма базується на застосуванні статистичних підходів і надає змогу виконувати граматичну розмітку текстів, написаних українською мовою, аналіз статистичних параметрів текстів (триграм)[2] і граматичну дизамбігуацію[3].

Розроблення алгоритмів автоматичного усунення граматичної неоднозначності словоформ передбачає вивчення поведінки граматичних омонімів у мові та мовленні на достатньо репрезентативному мовному матеріалі, адже «урахування кількісно-якісних характеристик явища сприятиме не тільки побудові оптимального алгоритму автоматичного зняття морфологічної неоднозначності текстових словоформ, а й вивченню закономірностей функціонування граматичної омонімії у процесі зберігання та передання інформації» [11, с.71].

Метою пропонованої статті є апробація алгоритму для лінгвістичного процесора[4], що сприятиме підвищенню кількості правильно ідентифікованих програмою неоднозначних словоформ, які виступають як власні і загальні назви, на матеріалі прозових та поетичних творів, а також детальний опис отриманих результатів у зіставному аспекті.

Аналіз даних ідентифікації текстових словоформ зазначеною програмою продемонстрував, що певна кількість омонімів зумовлена лексико-граматичною категорією відношення власна / загальна назва (*потік* – іменник, чоловічий рід, називний і знахідний відмінки, однина; *Потік* – іменник, чоловічий рід, називний і знахідний відмінки, однина, власна назва (річка в Україні); *Потік* – іменник, чоловічий рід, називний і знахідний відмінки, однина, власна назва (населений пункт в Україні); *потік* (від *потекти*) – дієслово док. виду, мин. час, чол. рід, одн.).

У писемному варіанті мови цей тип граматично неоднозначних словоформ розмежовується на графічному рівні, що дозволяє побудувати відповідний алгоритм, метою якого є ідентифікація неоднозначних словоформ, що виступають як власні і загальні назви[5]. Апробація алгоритму проводилася на матеріалі текстів художнього стилю, зокрема

[2] Кожне речення в тексті поділяється на трійки словоформ (триграми), які складаються з одиниці, що досліджується, і двох сусідніх (попередньої та наступної). Отже, при статистичному аналізі будь-якої словоформи враховуються показники всіх трьох складників триграми.

[3] Зняття неоднозначності, розв'язання неоднозначності, використання лінгвістичних та екстралінгвістичних чинників для уточнення неоднозначного слова в конкретному вживанні [1, с. 178].

[4] Сукупність штучних моделей природної мови, алгоритмів і програм, що описують будову та функціонування цих моделей, і технічні засоби, що реалізують цю модель [9, с. 10].

[5] Див. про це докладніше: Карнаух Г. В. Велика і мала літери як засіб розрізнення неоднозначних словоформ при автоматичній дизамбігуації / Г. В. Карнаух // Проблеми граматики і лексикології української мови : зб. наук. праць. – К. : НПУ ім. М. П. Драгоманова, 2011. – Вип. 8. – С. 26 – 36.

новели «Спогад про океан» Олеся Гончара, поетичних творів «Тінь Сізіфа» Ліни Костенко та «Кирпатий барометр» Василя Симоненка.

Для об'єктивної оцінки роботи запропонованого алгоритму дослідження виконано на різних за якісно-кількісними показниками робочих текстах[6], які, у свою чергу, належать до одного й того ж художнього твору. Так, на матеріалі новели було створено 3 робочих тексти (різняться за кількістю словоформ): речення (10 словоформ), уривок (174 словоформи), твір (20 208 словоформ). В основу формування робочих текстів на матеріалі поетичних творів покладено формальну ознаку: оформлення рядків у строфі (написання початкової літери першого слова кожного рядка). Відтак, Варіант 1 відповідає оригінальному (авторському) оформленню твору, у Варіанті 2 спеціально змінено початкову літеру першого слова в рядку на малу / велику відповідно до синтаксичних вимог оформлення поезій.

Отже, у поезії «Тінь Сізіфа» (Варіант 1) рядок починається зі слова, написаного з великої літери, якщо воно є початком речення, і з малої – якщо є продовженням попереднього. У поетичному творі «Тінь Сізіфа» (Варіант 2) кожен рядок починається словом, написаним із великої літери.

У робочому тексті «Кирпатий барометр» (Варіант 1) усі рядки починаються словом, написаним з великої літери; текст «Кирпатий барометр» (Варіант 2) оформлений за аналогією до поезії «Тінь Сізіфа» (Варіант 1).

Дослідження, присвячене апробації зазначеного алгоритму, проводилося в кілька етапів:

1. Маркування прозового твору
   - 1.1. «Спогад про океан»
     - 1.1.1. Речення
     - 1.1.2. Уривок
     - 1.1.3. Новела
   - 1.2. Опис та зіставлення отриманих результатів
2. Маркування поетичного твору
   - 2.1. «Тінь сізіфа»
     - 2.1.1. Варіант 1
     - 2.1.2. Варіант 2
   - 2.2. «Кирпатий барометр»
     - 2.2.1. Варіант 1
     - 2.2.2. Варіант 2
   - 2.3. Опис та зіставлення отриманих результатів

[6] Робочі тексти – ті, на матеріалі яких проведено апробацію алгоритму. Тексти взято з художніх творів «Спогад про океан» Олеся Гончара (Гончар Олесь. Твори : у 12 т. / О. Т. Гончар ; [редкол. : М. Г. Жулинський (голова) та ін.]. – К. : Наук. думка, 2001. – Т. 7 : Твоя зоря ; Далекі вогнища ; Спогад про океан; Коментарі / [упорядкув. та комент. С. А. Гальченка]. – 560 с.); «Тінь Сізіфа» Ліни Костенко (Костенко Л. В. Вибране. – К. : Дніпро, 1989. – 559 с.); «Кирпатий барометр» Василя Симоненка (Симоненко В. А. Земне тяжіння. Поезії. / В. А. Симоненко. – К. : Молодь, 1964. – 120 с.).

3. Загальний опис та зіставлення даних, отриманих у результаті маркування прозових і поетичних творів.

Для полегшення та прискорення процесу обробки результатів було створено допоміжну програму, що дозволяє зіставляти статистичні портрети[7] досліджуваних текстів і порівнювати отримані дані.

1. Маркування прозового твору. При аналізі результатів практичного застосування алгоритму в реченні та в уривку не було виявлено жодної помилки (неправильно ідентифікованої неоднозначної словоформи, що входить у досліджувану групу, яку умовно можна назвати «власна / загальна назва», тобто таку, що містить множини неоднозначних словоформ, які позначають і власну, і загальну назви); у новелі 150 словоформ програма ідентифікувала неправильно (усі належать до таких, що мають декілька варіантів граматичних характеристик для словоформ на позначення загальної назви).

Усі помилки, отримані внаслідок застосування алгоритму, умовно можна поділити на два типи:

1. Неправильно визначено відмінок. Приклад: *Яворницький, звичайно, розумів, чому він саме таким постає в очах людей, ...йому лестило, що ...люди ніби на вічне збереження несуть йому дари своєї пам'яті, найчистіші творіння душі, несуть* ***<u>пісні</u>****, ще ніким не записані...*

Пісні – іменник, жіночий рід, знахідний відмінок, множина (промарковано: *пісні* – іменник, жіночий рід, називний відмінок, множина).

2. Неправильно встановлено частиномовну належність. Приклад: *Адже, з погляду мого егоїзму і моїх нетерпимостей, не ви, а лиш я* ***<u>маю</u>*** *рацію у визначенні ходу подій.*

Маю – дієслово, недоконаний вид, теперішній час, перша особа, однина (промарковано: *маю* (*май* (*зелень*)) – іменник, чоловічий рід, родовий відмінок, однина).

Зміни в кількості ідентифікованих словоформ, які відбулися після застосування алгоритму, представлено у вигляді діаграми 1 (див. рис. 1).

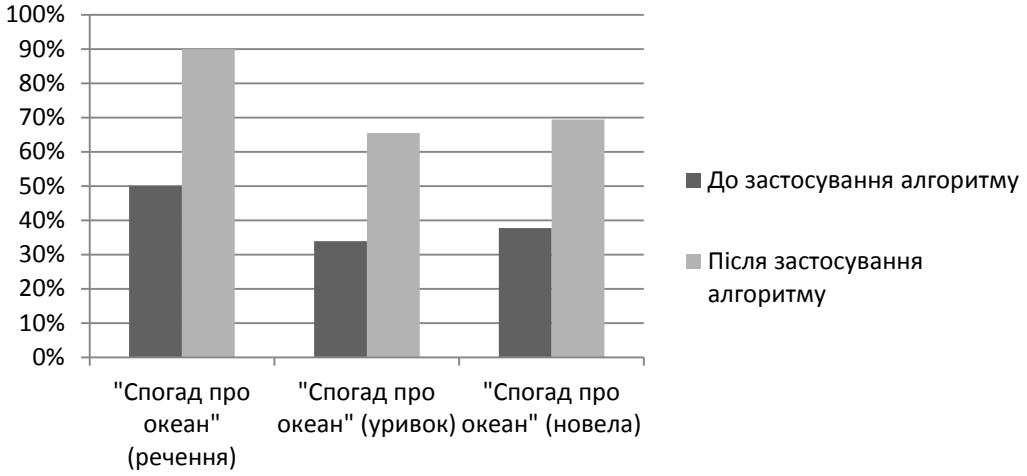
2. Маркування поетичних творів. Аналіз отриманих даних показав, що зміни в написанні початкової літери першого слова в рядку поезії «Тінь Сізіфа» майже не вплинули на результати маркування (у Варіанті 2, на відміну від Варіанта 1, неідентифікованою залишилася словоформа *потік*).

Проте зроблена нами зміна в написанні перших слів рядка поетичного твору «Кирпатий барометр» викликала збільшення кількості правильно розпізнаних неоднозначних словоформ, які входять до групи «власна/загальна назва» (Варіант 1 – 22 словоформи; Варіант 2 – 28 словоформ).

Також було виявлено словоформи, при ідентифікації яких програма припустилася помилки (неправильно визначений відмінок).

[7] Певним чином сформована сукупність статистичних ознак тексту, отриманих шляхом його статистичної обробки за певними заздалегідь встановленими принципами [7, с. 75].

Рис. 1



Приклад 1: *Над народами, над віками*
*Встало **горе**, мов чорний гном.*

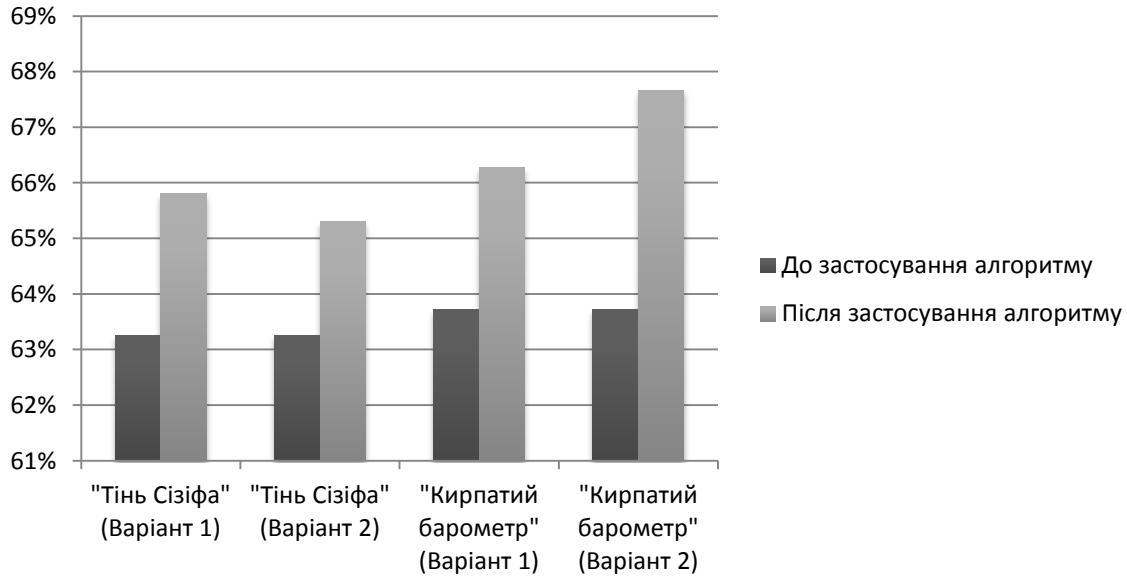
Горе – іменник, називний відмінок, середній рід, однина (промарковано: *горе* – іменник, знахідний відмінок, середній рід, однина).

Приклад 2: *Ми винуваті, що міліють ріки*
*І лисинами світять **береги**...*

Береги – іменник, називний відмінок, чоловічий рід, множина (промарковано: *береги* – іменник, знахідний відмінок, чоловічий рід, множина).

Помилок у виявленні частиномовної належності неоднозначної словоформи не виявлено. Кількісні зміни щодо ідентифікації словоформ, пов'язані з використанням алгоритму, відображає діаграма 2 (див. рис. 2).

Рис. 2



3. Зіставлення даних, отриманих у результаті маркування прозових і поетичних творів продемонструвало, що кількість неоднозначних словоформ, які належать до групи «власна/загальна назва», коливається у межах від 7 до 13 %. Кількість неправильно ідентифікованих словоформ зростає прямо пропорційно збільшенню кількості словоформ у тексті («Тінь Сізіфа» (Варіант 1): кількість словоформ (група «власна/загальна назва») – 14, неправильно розпізнані словоформи – 1 (7,14 %) і «Кирпатий барометр» (Варіант 1) – 60 і 9 (15 %) відповідно; «Спогад про океан» (уривок) – 11 і 0 та «Спогад про океан» (новела) – 20 208 і 150 (8,96 %) відповідно). Отримані результати подано у вигляді *Таблиці 1*.

Таблиця 1

Текст	Кількість словоформ у тексті				Кількість словоформ (група «власна/загальна назва»)			Промарковано словоформ		Нерозпізнано словоформ	
	Загальна	Омонімічні словоформи	Неомонімічні словоформи	Власна/загальна назва	Промарковано		Нерозпізнано омонімічних словоформ	До застосування алгоритму	Після застосування алгоритму	До застосування алгоритму	Після застосування алгоритму
					Правильно	Неправильно					
«Тінь Сізіфа» (Варіант 1)	196	118 60,20%	78 39,80%	14 7,14%	5 35,71% (від 14) 2,55% (від 196)	1 7,14% (від 14) 0,51% (від 196)	8 57,14% (від 14) 4,08% (від 196)	124 63,26%	129 65,81%	72 36,73%	67 34,18%
«Тінь Сізіфа» (Варіант 2)	196	118 60,20%	78 39,80%	14 7,14%	4 28,57% (від 14) 2,04% (від 196)	1 7,14% (від 14) 0,51% (від 196)	9 64,28% (від 14) 4,59% (від 196)	124 63,26%	128 65,30%	72 36,73%	68 34,69%
«Кирпатий барометр» (Варіант 1)	430	281 65,35%	149 34,65%	60 13,95%	22 36,66% (від 60) 5,11% (від 430)	9 15% (від 60 2,09% (від 430)	29 48,33% (від 60) 6,74% (від 430)	274 63,72%	285 66,27%	156 36,27%	145 33,72%
«Кирпатий барометр» (Варіант 2)	430	281 65,35%	149 34,65%	60 13,95%	28 46,66% (від 60) 6,51% (від 430)	9 15% (від 60) 2,09% (від 430)	23 38,33% (від 60) 5,34% (від 430)	274 63,72%	291 67,67%	156 36,27%	139 32,32%
«Спогад про океан» (речення)	10	5 50%	5 50%	1 10%	1 10%	0	1 10%	5 50%	9 90%	5 50%	1 10%
«Спогад про океан» (уривок)	174	115 66,09%	59 33,9%	11 6,32 %	6 54,54% (від 11) 3,44% (від 174)	0	5 45,45% (від 11) 2,87% (від 174)	59 33,91%	114 65,51%	115 66,09%	60 34,48%
«Спогад про океан» (новела)	20208	12582 62,26%	7626 37,74%	1673 8,29%	870 52,00% (від 1676) 4,30% (від 20208)	150 8,96% (від 1676) 0,74% (від 20208)	653 39,03% (від 1676) 3,23% (від 20208)	7626 37,74%	14033 69,44%	12582 62,26%	6175 30,55%

Отже, аналіз отриманих результатів показав, що використання запропонованого алгоритму сприяє збільшенню кількості ідентифікованих словоформ загалом, зокрема тих, які виступають як власні та загальні

назви, а також дає змогу розраховувати на підвищення відсотка точності автоматичної дизамбігуації словоформ у тексті. Виявлені при маркуванні неоднозначних одиниць порушення не пов'язані з роботою алгоритму, адже помилку допущено після того, як алгоритм «зняв» неоднозначність, зумовлену існуванням власних і загальних назв (за умови наявності декількох варіантів граматичних значень для словоформ на позначення загальної назви). Тому можна зробити висновок, що застосування алгоритму на практиці надало змогу виявити неточності в роботі програми загалом. Що ж до кількості неправильно ідентифікованих словоформ, то вона зростає прямо пропорційно збільшенню кількості словоформ у тексті. Написання початкової літери першого слова в рядку поетичного твору (велика / мала) впливає на кількість розпізнаних одиниць (зокрема, написання першого слова кожного рядка строфи з великої літери значно зменшує кількісні показники автоматичної дизамбігуації).

Література

1. Англо-русский словарь по лингвистике и семиотике / [ред.-упоряд. А. Н. Баранов, Д. О. Добровольский]. – Т. 1. – М. : Помовский и партнеры, 1996. – 641 с.

2. Карнаух Г. В. Велика і мала літери як засіб розрізнення неоднозначних словоформ при автоматичній дизамбігуації / Г. В. Карнаух // Проблеми граматики і лексикології української мови : зб. наук. праць. – К. : НПУ ім. М. П. Драгоманова, 2011. – Вип. 8. – С. 26 – 36.

3. Карцевский С. Об асимметричном дуализме лингвистического знака / С. Карцевский // История языкознания XIX – XX веков в очерках и извлечениях / [под ред. В. А. Звегинцева]. – Ч. II. – М. : Просвещение, 1965. – С. 85 – 90.

4. Колшанский Г. В. Контекстная семантика / Г. В. Колшанский. – М. : Наука, 1980. – 151 с.

5. Корпусна лінгвістика / [В. А. Широков, О. В. Бугаков, Т. О. Грязнухіна та ін.] ; під ред. В. А. Широкова. – К. : Довіра, 2005. – 407 с.

6. Кочерган М. П. Слово і контекст / М. П. Кочерган. – Львів : Вища школа, 1980. – 182 с.

7. Крыгин М. Ю. Текст на естественном языке как объект статистического анализа / М. Ю. Крыгин // Біоніка інтелекту : наук.-техн. журнал. – 2000. – № 1 (72). – С. 75 – 82.

8. Лосев А. Ф. Знак, символ, миф / А. Ф. Лосев. – М. : Изд-во Моск. ун-та, 1982. – 478 с.

9. Пиотровский Р. Г. Инженерная лингвистика и теория языка / Р. Г. Пиотровский. – Ленинград : Наука, 1979. – 112с.

10. Уфимцева А. А. Лексическое значение. Принципы семиологического описания лексики / А. А. Уфимцева. – М. : Наука, 1986. – 240 с.

Шипнівська О. О. Функціонування міжчастиномовних морфологічних омонімів в українських текстах / О. О. Шипнівська // Мовознавство. – 2005. – № 6 (233). – С. 70 – 80.