Віддієслівні похідні, утворюючись суфіксальним, префіксальним та постфіксальним способом, репрезентують інтегральні компоненти значення ʻметалʻ, ʻступінь корозійної стійкостіʻ, ʻпридатний для обробленняʻ твірного деривата *нікелювати*.

Отже, СГ із вершиною *нікел′-* характеризується наявністю афіксальних дериватів та композитних похідних, які, об'єднуючись у словотвірні пари і словотвірні ланцюжки, формують віяльно-ланцюжковий структурний тип гнізда. Найбільший словотвірний потенціал виявляє дериват *нікелювати*, репрезентований похідними II і III ступенів словотворення. Формально-семаничну організацію аналізованого СГ зумовлено нерозгалуженою семантичною структурою слова-вершини.

ЛІТЕРАТУРА

1. Василевич А.Я. Словообразовательные гнезда корней слов со значение ммышления в современном украинском языке: Дис. канд. филол. наук: 10.02.02 / А.Я. Василевич. – Запорожье, 1984. – 201 с.
2. Глинка Н.Л. Общая химия / Под. ред. В.А. Рабиновича. – Ленинград: Химия, 1979. – 720 с.
3. Русская граматика. –Ч. І. – М.: Наука, 1980.
4. Тихонов А.Н. Словообразовательный словарь русского языка: У 2-х тт. / А.Н. Тихонов – М.: Русский язык, 1985. – Т.1. – С. 36-50.
5. Украинская грамматика / Русановский В.М., Жовтобрюх М.А., Городенская Е.Г., Грищенко А.А. – К.: Наук.думка, 1986. – 360 с.
6. Карпіловська Є.А. Кореневий гніздовий словник української мови: Гнізда слів з вершинами – омографічними коренями / Є.А. Карпіловська– К.: Укр. енциклопедія, 2002. – 912 с.
7. Словник української мови: У 11-ти тт. – К.: Наук. думка, 1970 – 1980.
8. Українська мова. Енциклопедія. – К.: Укр. енцикл., 2000. – 752 с.
9. Український радянський енциклопедичний словник: У 3-х тт. – К.: Головна редакція УРЕ, 1986 – 1987.
10. Шкільний словотвірний словник сучасної української мови: 15600 слів у складі 127 гнізд / Уклад.: Н.Ф. Клименко та ін. – К.: Наук. думка, 2005. – 264 с.

В статье анализируется формальная и семантическая структура слов с корнем *нікел′-*. Установлены структурно-семантические отношения между словом-вершиной и производными единицами в словообразовательных парадигмах, взаимосвязи дериватов в словообразовательных цепочках. Определена продуктивность корня *нікел′-*.

Ключевые слова: словообразовательное гнездо, словообразовательная парадигма, словообразовательная цепочка, словообразовательное значение, слово-вершина, производящая основа, производное слово.

The article analyses the formal and semantic structure of word with root*нікел′-*. It is characterised the structural-semantic relations between the top and the units of word-formative paradigms, interrelations of derivatives in word-formative chains. It is also defined the degrees of root productivity*нікел′-*.

Key words: a word-formative jack,a word-formative paradigm,a word-formative chain, word-formative value, the word-top, the making basis, a derivative word.

УДК 801. 314.2 : 003. 32 : 81 – 139

Г. В. Карнаух

ВЕЛИКА І МАЛА ЛІТЕРИ ЯК ЗАСІБ РОЗРІЗНЕННЯ НЕОДНОЗНАЧНИХ СЛОВОФОРМ ПРИ АВТОМАТИЧНІЙ ДИЗАМБІГУАЦІЇ

У статті порушено питання зняття граматичної неоднозначності словоформ в системах автоматичної обробки текстів. Запропоновано алгоритм ідентифікації словоформ, які можуть позначати і власні, і загальні назви. Виявлено специфіку роботи алгоритму в текстах художнього стилю.

Ключові слова: граматична неоднозначність словоформ, омоніми, дизамбігуація (зняття неоднозначності словоформ), алгоритм, власна назва, загальна назва, граматичне значення, морфологічний аналіз, лінгвістичний процесор.

Бурхливий розвиток інформаційного суспільства зумовлює необхідність якісно нових засобів опрацювання інформації, які розвиваються, насамперед, у напрямі інтелектуалізації. Принципово нові можливості комп'ютерів і мереж вимагають розробки нових або вдосконалення вже наявних систем автоматичного аналізу тексту. Особливого значення при цьому набуває проблема автоматичного усунення граматичної неоднозначності словоформ, яка на сучасному етапі розвитку корпусної лінгвістики потребує нагального розв'язання, оскільки пов'язана з питанням автоматизованого маркування корпусів. Виникає необхідність створення ефективних методик у контексті сучасного стану лінгвістичних знань та комп'ютерних технологій [3].

Необхідною передумовою побудови відповідного лінгвістичного процесора[1] є розроблення формальної моделі розуміння текстів з урахуванням морфологічного, синтаксичного та семантичного аналізів. При цьому формалізований опис природної мови вимагає від дослідників акцентуації уваги на такій властивості мовного знака, як неоднозначність[2] (здатність в одному означальному поєднувати декілька означуваних) [12, с.70].

Метою пропонованої статті є розроблення та апробація алгоритму для лінгвістичного процесора, який сприятиме підвищенню кількості правильно ідентифікованих програмою неоднозначних словоформ.

Проблеми, пов‘язані з вивченням й усуненням граматичної неоднозначності словоформ, порушено багатьма лінгвістами і на теоретичному, і на практичному рівнях. Теоретичний аспект висвітлено в працях О. Потебні, В. Виноградова, Ю. Апресяна, О. Ахманової, О. Смирницького, Г. Уфімцевої, Д. Шмельова, В. Русанівського, Н. Клименко, М. Кочергана, О. Тараненка, М. Муравицької, Л. Авксентьєва, Г. Гнатюк, Л. Лисиченко, А. Лучик, В. Широкова, Т. Грязнухіної та інших вчених здебільшого в контексті загальних проблем мовної неоднозначності, омонімії та полісемії, зумовлених потребами лексикографії, лексикології та граматики. Практичний аспект граматичної неоднозначності представлено у роботах Р. Піотровського, Ю. Марчука, Т. Аполлонської, С. Білокриницької, Т. Молошної, Т. Грязнухіної, О. Невзорової, Ю. Орєхова, Т. Любченко. Ці дослідження присвячені автоматичній ідентифікації текстових словоформ (машинний переклад, автоматичне коригування тексту тощо). При цьому головна увага зосереджується на розгляді принципу контекстної зумовленості конкретного граматичного значення слова в тексті [11, с. 5].

В Українському мовно-інформаційному фонді НАН України (УМІФ) розробляють методи автоматичного визначення граматичних характеристик словоформ у тексті, які слугують маркерами граматичних неоднозначностей (зокрема, граматичної омонімії), на основі застосування статистичних підходів [6]. Запропоновано комп‘ютерну програму, що базується на застосуванні статистичних підходів і надає змогу виконувати граматичну розмітку текстів, написаних українською мовою, аналіз статистичних параметрів текстів (триграм)[3] і граматичну дизамбігуацію[4] (зняття неоднозначності словоформ).

Оскільки розробка алгоритмів автоматичної дизамбігуації (у межах лінгвістичного процесора) передбачає накопичення максимально повної інформації про потенційні можливості словникового складу мови стосовно прояву граматичної неоднозначності, джерелом відповідних даних для проведення морфологічного аналізу й дизамбігуації є веб-сервіс розподіленої програмної лексикографічної системи «Граматичний словник української мови». Він надає необхідні граматичні параметри для лексем тексту, що обробляється, і доступний у локальній мережі, а також у мережі Інтернет [10].

Перед застосуванням тексти підлягають попередньому аналізу (вони мають повністю збігатися з оригіналом, відповідати вимогам чинного правопису української мови (не містити помилок, пропусків)), а також розмітці відповідно до цілей дослідження.

Модель тексту, в якому здійснюється зняття граматичної неоднозначності словоформ, відображає формула:

$$T_i=\{(w_1)r_1(w_2)r_2(w_3)\ldots(w_N)\},$$

де w_i – словоформи, r_i – роздільники між словоформами – знаки пунктуації, пробіли та ін., N – кількість словоформ у тексті T_i.

Кожне слово тексту T у свою чергу задано параметрами $w_i = (v_i, g_i)$, де v_i відображає частину мови словоформи, а g_i – граматичне значення, тобто текст має вигляд: $M(T)=\{(v_1, g_1)\ (v_2, g_2)\ (v_3, g_3)\ldots(v_N, g_N)\}$. Перетворення M: $T \rightarrow M(T)$ будемо називати розміткою тексту T [5, с. 517].

Приклад: T = *Своєю граціозністю й мальовничістю журавель привертає до себе загальну увагу.*

---

[1] Сукупність штучних моделей природної мови, алгоритмів і програм, що описують будову та функціонування цих моделей, та технічні засоби, що реалізують цю модель [7, с. 10].

[2] Неоднозначність або ж багатозначність трактується як родове поняття, що поєднує полісемію та омонімію [4, с. 93].

[3] 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.

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

M(*T*)={(v_1=*займенник, g_1=орудний відмінок, жіночий рід, однина*) (v_2=*іменник, g_2=орудний відмінок, жіночий рід, однина*) (v_3=*сполучник, g_3=незмінне службове слово*) (v_4=*іменник, g_4=орудний відмінок, жіночий рід, однина*) (v_5=*іменник, g_5=називний відмінок, чоловічий рід, істота, однина*) (v_6=*дієслово, g_6=недоконаний вид, теперішній час, третя особа, однина*) (v_7=*прийменник, g_7= незмінне службове слово*) (v_8=*займенник, g_8=родовий відмінок*) (v_9=*прикметник, g_9=знахідний відмінок, жіночий рід, однина*) (v_{10}=*іменник, g_{10}=знахідний відмінок, жіночий рід, однина*)} (див. табл. 1).

Таблиця 1

w_1= своєю	w_2 =граціозністю	w_3=й	w_4=мальовничістю	w_5=журавель	w_6= привертає	w_7= до	w_8= себе	w_9= загальну	w_{10}= увагу
v_1=займенник, g_1=орудний відмінок, жіночий рід, однина	*v_2=іменник, g_2=орудний відмінок, жіночий рід, однина*	*v_3=сполучник, g_3= незмінне службове слово*	*v_4=іменник, g_4=орудний відмінок, жіночий рід, однина*	*v_5=іменник, g_5=називний відмінок, чоловічий рід,істота,однина*	*v_6=дієслово, g_6=недоконаний вид, теперішній час, третя особа, однина*	*v_7=прийменник, g_7= незмінне службове слово*	*v_8=займенник, g_8= родовий відмінок*	*v_9=прикметник, g_9=знахідний відмінок, жіночий рід, однина*	*v_{10}=іменник, g_{10}=знахідний відмінок, жіночий рід, однина*

Зняття граматичної неоднозначності статистичними методами передбачає морфологічний аналіз тексту алгоритмом лематизації [10], у результаті кожне слово отримує так званий набір граматичних характеристик, який внаслідок неоднозначності словоформ може бути досить чисельним, тому загалом кожне слово *w* характеризується вектором граматичних станів (*v, g*) (див. табл. 2).

Морфологічний аналіз тексту є відображенням M': $T \rightarrow M'(T)$. Завданням лінгвістичного процесора є максимально можливе наближення розмітки $M'(T)$ до $M(T)$, адже «розмітка $M'(T)$ є неоднозначною й містить інформацію про всі теоретично можливі граматичні значення словоформи» [5, с. 518].

Автори статті, присвяченої опису зазначеної програми, наголошують, що «параметри розмітки тексту $M(T)$ утворюють n-зв‘язний ланцюг Маркова (ЛМ), елементами якого виступають граматичні стани словоформ (v, g). Для вивчення поведінки цього ланцюга та знаходження статистичних закономірностей створюються так звані навчальні вибірки з текстів різних стилів» [5, с.518]

Такий підхід до опису явища неоднозначності передбачає врахування специфіки мови як абстрактної системи мовних знаків та тексту як конкретної реалізації цієї системи, тобто співвідношення мова/текст розглядають як протиставлення потенції та її реалізації.

Таблиця 2

w_1=своєю	w_2 =граціозністю	w_3=й	w_4=мальовничістю	w_5=журавель	w_6=привертає	w_7=до	w_8=себе	w_9=загальну	w_{10}=увагу
$(v_1, g_1)=\{(v_1^1, g_1^1)\}$	$(v_2, g_2)=\{(v_2^1, g_2^1)\}$	$(v_3, g_3)=\{(v_3^1, g_3^1), (v_3^2, g_3^1)\}$	$(v_4, g_4)=\{(v_4^1, g_4^1)\}$	$(v_5, g_5)=\{(v_5^1, g_5^1), (v_5^1, g_5^2), (v_5^1, g_5^3)\}$	$(v_6, g_6)=\{(v_6^1, g_6^1)\}$	$(v_7, g_7)=\{(v_7^1, g_7^1), (v_7^2, g_7^1), (v_7^2, g_7^2), (v_7^2, g_7^3), (v_7^2, g_7^4), (v_7^2, g_7^5), (v_7^2, g_7^6), (v_7^2, g_7^7), (v_7^2, g_7^8), (v_7^2, g_7^9), (v_7^2, g_7^{10}), (v_7^2, g_7^{11}), (v_7^2, g_7^{12}), (v_7^2, g_7^{13}), (v_7^2, g_7^{14})\}$	$(v_8, g_8)=\{(v_8^1 g_8^1), (v_8^1 g_8^1)\}$	$(v_9, g_9)=\{(v_9^1, g_9^1)\}$	$(v_{10}, g_{10})=\{(v_{10}^1, g_{10}^1)\}$
v_1^1=займенник, *g_1^1=орудний відмінок, жіночий рід, однина*	*v_2^1=іменник,* *g_2^1=орудний відмінок, жіночий рід, однина*	*v_3^1=сполучник,* *g_3^1=незмінне службове слово* *v_3^2=частка,* *g_3^1=незмінне службове слово*	*v_4^1=іменник,* *g_4^1=орудний відмінок, жіночий рід, однина*	*v_5^1=іменник,* *g_5^1=називний відмінок,чоловічий рід,істота,однина,* *g_5^2=називний відмінок,чоловічий рід,однина,* *g_5^3=знахідний відмінок, чоловічий рід, однина*	*v_6^1=дієслово,* *g_6^1=недоконаний вид, теперішній час, третя особа, однина*	*v_7^1=прийменник,* *g_7^1=незмінне службове слово* *v_7^2=іменник,* *g_7^1=називний відмінок, середній рід, однина,* *g_7^2=родовий відмінок, середній рід, однина,* *g_7^3= давальний*	*v_8^1=займенник,* *g_8^1=родовий відмінок,* *g_8^2=знахідний відмінок*	*v_9^1=прикметник,* *g_9^1=знахідний відмінок,жіночий рід, однина*	*v_{10}^1=іменник,* *g_{10}^1=знахідний відмінок, жіночий рід, однина*

						відмінок, середній рід, однина, g_7^4*= знахідний відмінок, середній рід, однина,* g_7^5*=орудний відмінок, середній рід, однина,* g_7^6*=місцевий відмінок, середній рід, однина,* g_7^7*=кличний відмінок, середній рід, однина,* g_7^8*=називний відмінок, середній рід, множина,* g_7^9*=родовий відмінок, середній рід, множина,* g_7^{10}*=давальний відмінок, середній рід, множина,* g_7^{11}*=знахідний відмінок, середній рід, множина,* g_7^{12}*=орудний відмінок, середній рід, множина,* g_7^{13}*=місцевий відмінок, середній рід, множина,* g_7^{14}*=кличний відмінок, середній рід, множина*			

Суттєвою перешкодою в процесі наближення розмітки $M'(T)$ до $M(T)$ є омоніми[5], зумовлені лексико-граматичною категорією відношення власна/загальна назва (*журавель* – іменник чоловічого роду, істота, називний відмінок, однина, загальна назва; *Журавель* – іменник чоловічого роду, неістота, називний та знахідний відмінки, однина, власна назва (річка в Україні)). В. Виноградов писав: «Жодна мова не була б спроможна виражати кожну конкретну ідею самостійним словом або кореневим елементом. Конкретність досвіду безмежна, ресурси ж найбагатшої мови суворо обмежені» [2, с. 18]. А оскільки творення власних назв тісно пов‘язане з назвами індивідуальних, унікальних, одиничних (за своєю онтологічною природою) предметів (у широкому витлумаченні), явищ, понять навколишнього світу і зумовлене набуттям ними ономастичної функції, кількість омонімів такого типу в досліджуваних текстах становить у середньому 8 %. Щодо зняття граматичної неоднозначності словоформ статистичними методами цей показник є суттєвим, адже кожна словоформа може як «допомогти», так і «завадити» процесу ідентифікації сусідніх словоформ (принцип триграм).

У писемному варіанті мови власне омоніми, що збігаються за звучанням, але належать до різних груп, виділених за принципом відношення власна/загальна назва, розрізняються на графічному рівні. За правописом української мови власні назви пишуться з великої літери (у складних і складених власних назвах з великої літери можуть писатися всі або деякі слова (частини)), загальні назви прийнято писати з малої літери (крім тих випадків, коли слово на позначення загальної назви стоїть на початку речення, рядка (в поезії), рубрики, прямої мови, цитат (у певних випадках)).

Запропоновані в УМІФі методи статистичного аналізу текстів, які використовуються при автоматичній дизамбігуації, враховують декілька рівнів мови: знаковий, фонетичний, лексичний, граматичний і т. ін. [6, с. 75], тобто існує можливість розрізнення графеми за написанням (велика/мала літера). Це дає змогу побудувати алгоритм зняття неоднозначності словоформ, що базуватиметься на принципі відношення власна/загальна назва.

Аналіз правил вживання великої/малої літери в українській мові показав, що слово, написане з малої літери, завжди позначає загальну назву; навіть тоді, коли воно є складовою складних і складених власних назв, позначаючи родове поняття, зберігає всі граматичні характеристики, що й словоформа, яка кваліфікується як загальна назва (див. табл. 3).

Приклад: *«Київська газета»* – назва періодичного видання й *київська газета* – газета, що видається в Києві.

Таблиця 3

Відмінки	Загальна назва		Власна назва	
	однина	множина	однина	множина
Називний	київська газета	київські газети	«Київська газета»	«Київські газети»
Родовий	київської газети	київських газет	«Київської газети»	«Київських газет»
Давальний	київській газеті	київським газетам	«Київській газеті»	«Київським газетам»
Знахідний	київську газету	київські газети	«Київську газету»	«Київські газети»
Орудний	київською газетою	київськими газетами	«Київською газетою»	«Київськими газетами»
Місцевий на (у)	київській газеті	київських газетах	«Київській газеті»	«Київських газетах»
Кличний	київська газето	київські газети	«Київська газето»	«Київські газети»

Враховуючи вищесказане, можна вважати правильним твердження, що словоформу, написану з малої літери, яка займає в реченні будь-яку позицію, крім початкової, можна ідентифікувати як таку, що позначає загальну назву.

Це твердження є основою побудови алгоритму для лінгвістичного процесора, що здійснює граматичну дизамбігуацію.

Робота алгоритму складається з декількох етапів:

1. Вибір словоформи.
2. Виявлення позиції словоформи в реченні (позиція 1 відповідає першому слову в реченні, позиція 2 – другому і т.д.).
3. Якщо лінгвістичний процесор фіксує позицію 1, неоднозначна словоформа залишається без змін (неоднозначність не знімається), словоформа, що займає позицію 2,3,...n

[5] О.О. Тараненко висловлює думку про неправомірність віднесення до омофонів омонімії загальної та власної назв (*роман і Роман*), адже збіг відбувається не тільки на фонетичному, а й на фонемному рівні. Такі одиниці він вважає власне омонімами [8, с. 403].

маркується відповідним чином (за умови, що в граматичному словнику міститься інформація про аналізовану словоформу як таку, що позначає і власну, і загальну назви).

4. Графемний аналіз словоформи: якщо словоформа написана з великої літери, неоднозначність не знімається, якщо з малої літери – словофрма ідентифікується як така, що вживається на позначення загальної назви (у випадках наявності лише одного варіанта граматичних характеристик, що відповідають загальній назві, неоднозначність словоформи знімається повністю).

Останній етап роботи алгоритму представлений на малюнку 1.

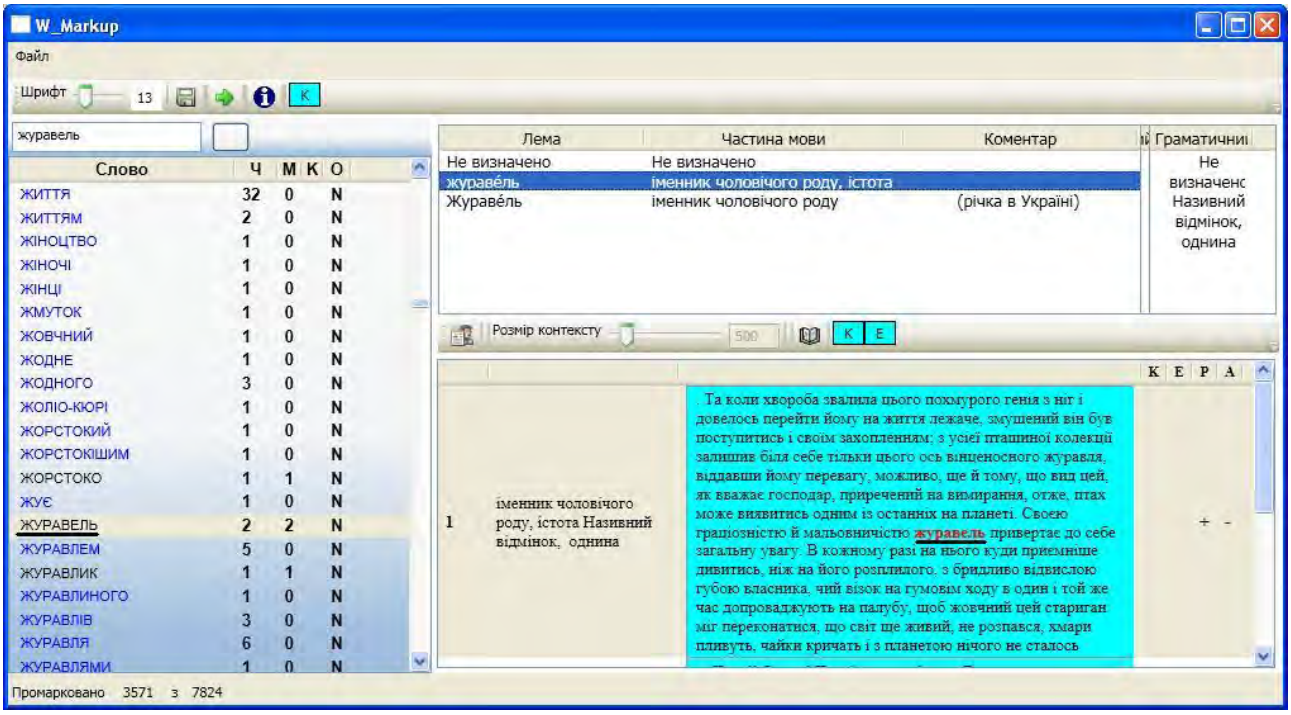
Схематично принцип дії алгоритму зображений на малюнку 2.

Апробацію алгоритму проведено на матеріалі тексту художнього стилю, написаного прозовою мовою («Спогад про океан» Олеся Гончара), що налічує 20 208 словоформ.

Для об‘єктивної оцінки роботи запропонованого алгоритму дослідження виконано на різних за розміром робочих текстах[6] (речення (10 словоформ), уривок (174 словоформи), твір (20 208 словоформ)).

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 ту, що містить множини неоднозначних словоформ, які позначають і власну, і загальну назви).

Мал. 1



У реченні лінгвістичний процесор з 10 словоформ ідентифікував 9. Нерозпізнаною залишилася словоформа *увага* (*зосередженість, турбота*), яка є повним лексичним омонімом словоформи *увага* (*зауваження, доповнення*)[7]. Для розрізнення цих словоформ необхідний семантичний аналіз досліджуваних одиниць, який цією програмою не передбачений.

В уривку з 11 словоформ, що належать до досліджуваної групи, програма правильно ідентифікувала 6. Нерозпізнаними залишилися 5 (серед них: 1 – займає у реченні позицію 1; 1 – жіноче ім‘я (власна назва); 3 – мають декілька варіантів граматичних характеристик для загальної назви). Усі три типи неідентифікованих словоформ демонструють випадки, коли досліджувані одиниці мають характеристики, які програма, за умовами побудови алгоритму, визначає як такі, за яких словоформа залишається нерозпізнаною лінгвістичним процесором.

У повісті з 1673 одиниць, які мають омонімічні словоформи на позначення власної та загальної назви, лінгвістичний процесор правильно ідентифікував 870. Також було виявлено 653

[6] Робочі тексти – ті, на матеріалі яких проведено апробацію алгоритму. Усі тексти взято з художнього твору «Спогад про океан» Олеся Гончара (Гончар Олесь. Твори : в 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.)

[7] Граматичні характеристики подано в таблиці 2.

словоформи, при ідентифікації яких було допущено помилку (серед неправильно розпізнаних неоднозначних одиниць усі належать до таких, що мають декілька варіантів граматичних характеристик для словоформ на позначення загальної назви). Зазначені словоформи умовно можна поділити на такі групи:

1. Неправильно визначений відмінок.

Приклад: *Із року в рік іде, гайдабурить степами, і привітом зустрічають натруджені* ***села*** *свого професора, здалека впізнають його постать у солом'янім брилі, в одежі селянській, з очима, повними розуму, і жарту, і мрії…*

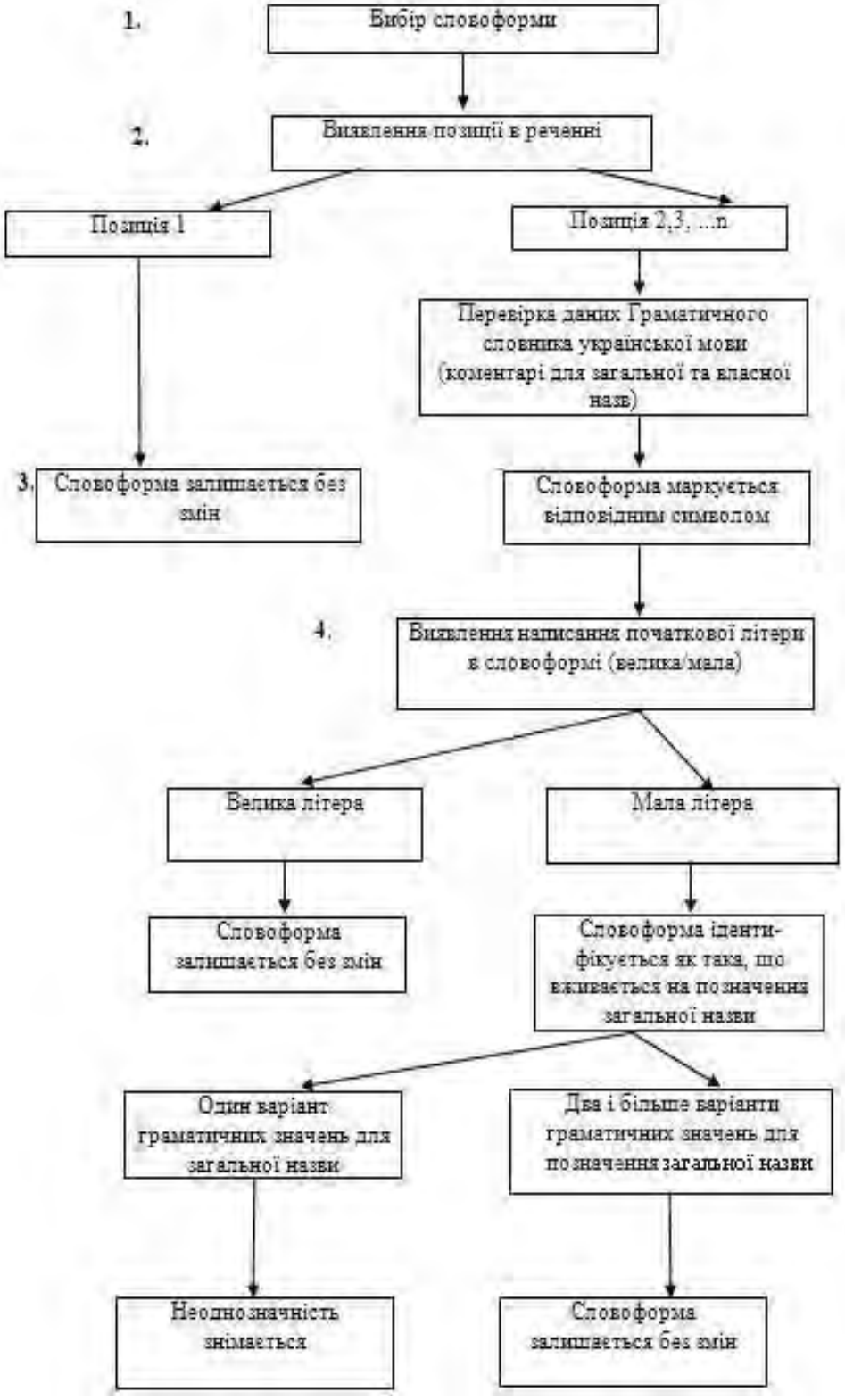
Словоформа, *w*	Лема	Частина мови, v	Коментар	Граматичне значення, g	Промарковано	Має бути
села	Село село	іменник іменник	населений пункт в Україні загальна назва	родовий відмінок, середній рід, однина родовий відмінок, середній рід, однина; називний відмінок, середній рід, множина; знахідний відмінок, середній рід, множина; кличний відмінок, середній рід, множина	іменник (загальна назва), родовий відмінок, середній рід, однина	іменник (загальна назва), називний відмінок, середній рід, множина

2. Неправильно встановлено частиномовну приналежність.

Приклад: *Щоліта мандрівний наш професор гостює на тутешніх пасіках, ночує в куренях на баштанах, а вечорами розповідає різні історії про ці плавні, де росло все на світі, де людина без суєти* ***жила*** *поруч з чистими озерами, з нелякливими лелеками, з журавлями і добрими черепахами.*

Словоформа	Лема	Частина мови	Коментар	Граматичне значення	Промарковано	Має бути
жила	жилий жила жила Жила жити	прикметник іменник (істота) іменник іменник дієслово	 скупа людина судина річка в Україні	називний відмінок, жіночий рід, однина називний відмінок, чоловічий рід, однина називний відмінок, жіночий рід, однина називний відмінок, жіночий рід, однина недоконаний вид, минулий час,жіночий рід, однина	іменник (судина), називний відмінок, жіночий рід, однина	дієслово, недоконаний вид, минулий час,жіночий рід, однина

Мал. 2


1.
Вибір словоформи
2.
Виявлення позиції в реченні
Позиція 1
Позиція 2,3, ...n
Перевірка даних Граматичного словника української мови (коментарі для загальної та власної назв)
3.
Словоформа залишається без змін
Словоформа маркується відповідним символом
4.
Виявлення написання початкової літери в словоформі (велика/мала)
Велика літера
Мала літера
Словоформа залишається без змін
Словоформа ідентифікується як така, що вживається на позначення загальної назви
Один варіант граматичних значень для загальної назви
Два і більше варіанти граматичних значень для позначення загальної назви
Неоднозначність знімається
Словоформа залишається без змін

Таблиця 4

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					
Речення	10	5 50%	5 50%	1 10%	1 10%	0	1 10%	5 50%	9 90%	5 50%	1 10%
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%
Твір	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%

Отримані результати подано у вигляді таблиці (див. табл. 4).

Виконане дослідження попри кількісну та стильову обмеженість дає змогу визначити деякі особливості поведінки словоформ у тексті, написаному українською мовою (у цьому випадку – прозовий твір художнього стилю), а відповідно – і зазначеної програми загалом. Під час проведення експерименту було встановлено, що неоднозначні словоформи, які мають варіанти граматичних характеристик і для власних, і для загальних назв, налічують у досліджуваних текстах художнього стилю близько 8 % від загальної кількості. Після застосування запропонованого алгоритму зняття граматичної неоднозначності досліджуваних одиниць виявлено, що в невеликих за обсягом текстах (до 200 словоформ) 54,54 % одиниць (від загальної кількості досліджуваної групи) і 3,44 % (від загальної кількості словоформ у тексті) програма ідентифікувала правильно. Нерозпізнаними залишились відповідно 45,45 % і 2,87 % одиниць. Неправильно ідентифікованих словоформ не виявлено.

Результати апробації алгоритму в тексті, що налічує 20 208 словоформ, щодо відношення розпізнаних і нерозпізнаних лінгвістичним процесором одиниць не зазнали суттєвих змін: 52,00 % (від загальної кількості досліджуваної групи) і 39,03 % (від загальної кількості словоформ у тексті). Однак при аналізі результатів було встановлено, що 8,96 % і 0,74 % відповідно програма ідентифікувала неправильно. Суттєво, що помилку допущено після того, як алгоритм «зняв» неоднозначність, пов‘язану з поняттям власна/загальна назва (за умови наявності декількох варіантів граматичних значень для словоформ на позначення загальної назви). Отже, можна зробити висновок, що застосування алгоритму на практиці надало змогу виявити неточності в роботі програми загалом.

Аналіз отриманих результатів показав, що використання запропонованого алгоритму дає змогу розраховувати на підвищення відсотка точності автоматичної дизамбігуації словоформ у тексті. У перспективі доцільною є апробація алгоритму на матеріалі текстів інших стилів, що сприятиме накопиченню адекватної навчальної вибірки, яка, в свою чергу, допоможе створити більш точний статистичний портрет будь-якого тексту для виконання відповідних досліджень.

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. Виноградов В. В. Русский язык. Грамматическое учение о слове / Виноградов В.В. – М. : Высшая школа, 1972. – 614 с.
3. Корпусна лінгвістика / [В. А. Широков, О. В. Бугаков, Т. О. Грязнухіна та ін.] ; під ред. В. А. Широкова. – К. : Довіра, 2005. – 407 с.
4. Кочерган М. П. Слово і контекст / Кочерган М. П. – Львів : Вища школа, 1980. – 182 с.
5. Крыгин М. Снятие грамматической омонимии в тексте с помощью стастических методов / Крыгин М., Шкурко В. Афанасьева О. // Прикладна лінгвістика та лінгвістичні технології : MegaLing-2009. – К. : Довіра, 2009. – 527 с.
6. Крыгин М. Текст на естественном языке как объект статистического анализа / М. Ю. Крыгин // Біоніка інтелекту : наук.-техн. журнал. – 2000. – ₣ 1 (72). – С. 75 – 82.
7. Пиотровский Р. Г. Инженерная лингвистика и теория языка / Пиотровский Р. Г. – Ленинград : Наука, 1979. – 112с.
8. «Українська мова». Енциклопедія / [Редкол.: Русанівський В. М., Тараненко О. О. (співголови), Зяблюк М. П. та ін.] – К.: «Укр. енцикл.», 2000. – 752 с.
9. Український правопис / Ін-т мовознавства ім. О. О. Потебні НАН України, Ін-т укр. мови НАН України. – К. : Наук. думка, 2007. – 288 с.
10. Шевченко И. В. Электронный грамматический словарь украинского языка. / Шевченко И. В., Рабулец А. Г., Широков В. А. // Труды Международной конференции «MegaLing-2005. Прикладная лингвистика в поиске новых путей», 27 июня – 2 июля 2005 г. – Меганом., 2005. – С. 124 – 129.
11. Шипнівська О. О. Структурно-семантичні та функціональні характеристики міжчастиномовної морфологічної омонімії сучасної української мови: дис. ... канд. філол. наук: 10.02.01 / Шипнівська Ольга Олександрівна. – К., 2007. – 238 с.
12. Шипнівська О. О. Функціонування міжчастиномвних морфологічних омонімів в українських текстах / О. О. Шипнівська // Мовознавство. – 2005. – ₣ 6 (233). – С. 70 – 80.

В статье рассматривается вопрос снятия грамматической неоднозначности словоформ в системах автоматической обработки текстов. Предложен алгоритм идентификации словоформ, обозначающих имена собственные и нарицательные. Выявлена специфика работы алгоритма в текстах художественного стиля.

Ключевые слова: грамматическая неоднозначность словоформ, омонимы, дизамбигуация (снятие неоднозначности словоформ), алгоритм, имя собственное, имя нарицательное, грамматическое значение, морфологический анализ, лингвистический процессор.

This Article touches upon the issue of elimination of grammatical ambiguity of separate word forms in systems of automatic texts processing. It is offered some algorithm of word forms identification which may designate both proper and common names. It is found specificity of algorithm in belles-lettres texts.

Key words: grammatical ambiguity, homonyms, disambiguation (elimination of grammatical ambiguity of separate word forms), algorithm, proper name, common name, grammatical meaning, morphological analysis, linguistic processor.