УДК 681.3.06

В.В. ВОЙТКО, М.П. БОЦУЛА, Г.Л. ЛУЦИШИН

РОЗРОБКА ТА РЕАЛІЗАЦІЯ ЛІНГВІСТИЧНИХ МЕТОДІВ АНАЛІЗУ ВЛАСНИХ НАЗВ

Вінницький національний технічний університет
95, вул. Хмельницьке шосе, 21001, м. Вінниця, Україна
E-mail: Hennadiy@bk.ru

Анотація. У статті запропоновано лінгвістичний метод аналізу власних назв, орієнтований на автоматизацію процесів словотворення у текстових і пошукових комп'ютерних системах. Розроблені семантичні моделі відмінкових форм власних назв реалізують правила чинного українського правопису в програмних засобах автоматизації.

Аннотация. В статье предложен лингвистический метод анализа собственных имён, ориентированный на автоматизацию словообразовательных процессов в текстовых и поисковых компьютерных системах. Разработанные семантические модели падежных форм собственных имён реализуют правила действующего украинского правописания в программных средствах автоматизации.

Abstract. The paper presents the linguistic analyze method of proper nouns which aligned on word-formative process automation for textual and searching computer-based systems. There had been developed semantic models for proper nouns case forms which realized orthography rules of ukrainian language for automation software.

Ключові слова. Специфікація COM, український правопис, власні назви, семантичні моделі, пошукові системи.

ВСТУП

Сьогодні, в епоху розвитку електронного документообігу, широкого розповсюдження набувають програмні засоби автоматизації процесів роботи з текстовою та графічною інформацією. Сучасні офісні програми, бази даних, бухгалтерські та банківські системи, пошукові системи працюють зі складними семантичними конструкціями. У зв'язку з цим виникає проблема перевірки правильності синтаксису та семантики речень, словосполучень та слів. Крім того, постає проблема розробки та реалізації україномовних пошукових систем, у яких пошук має відбуватися не лише за вказаним словом, а й за його відмінковими формами, що забезпечить більший варіантний масив шуканих об'єктів та розширить простір вибору. Одна з важливих задач програмної реалізації таких засобів автоматизації зводиться до аналізу варіантів з урахуванням семантичних змін при відмінюванні слів. Широко відомі засоби перевірки орфографії MS Office, Open Office Org, словники Mozilla не забезпечують користувачів такою можливістю. Тому актуальним є питання розробки засобів автоматизованого вибору слів з урахуванням їх відмінкових форм, що, у свою чергу, потребує створення електронних бібліотек. Крім того, важливою задачею є реалізація та використання програмних засобів, спрямованих на перевірку правильності подання власних назв і спеціалізованих термінів та на забезпечення можливості утворення їх нових форм відмінкового характеру.

Метою роботи є автоматизація процесів словотворення тексту на основі реалізації правил української орфографії у моделях форм відмінювання власних назв. Поставлена мета передбачає розробку алгоритму та його реалізацію як бібліотеки динамічної компоновки (DLL) за стандартом Component Objects Model (COM) для програмних додатків MS Windows та веб-орієнтованих систем. Об'єктом дослідження постають методи автоматизації процесів словотворення, лінгвістичні правила їх реалізації та методології створення DLL за специфікацією COM [1]. Під предметом дослідження розуміємо україномовні орфографічні правила формування відмінкових форм власних назв та можливість їх реалізації у моделі DLL.

У зв'язку з цим першочергово постають задачі дослідження правил відмінювання власних назв в українській мові, аналізу існуючих лінгвістичних методів та розробки моделей реалізації лінгвістичних

правил відмінкових словотворень.

АНАЛІЗ ІСНУЮЧИХ МЕТОДІВ РЕАЛІЗАЦІЇ МОДЕЛЕЙ ВІДМІНКОВИХ ФОРМ ВЛАСНИХ НАЗВ

Досліджуючи правила відмінювання власних назв у російській мові, розробники лінгвістичних методів зупинилися на виборі методів, орієнтованих на аналіз кінцівки початкової форми вхідного слова [2]. Загалом, відмінювання слів не передбачає обов'язкового визначення частин слова: кореня, префікса, суфікса і т. п. За відмінками змінюється лише закінчення. Тому правила відмінювання власних назв у російськомовних моделях відмінкових форм базуються саме на аналізі закінчення та букви, що безпосередньо йому передує. Закінченнями в словах виступають кінцеві букви, які програмно ідентифікувати досить легко, що обумовлює перспективність використання методу. Таким чином, алгоритм словотворення зводиться до аналізу кінцевих букв слова, так званої кінцівки початкової форми (КПФ), та заміни закінчення початкової форми закінченням потрібного відмінку. Вагомою проблемою для методів лінгвістичного аналізу постає наголос, який програмно визначити неможливо. Крім того, існує багато слів-винятків, які не підпадають під жодне правило. Тому автоматизовані системи передбачають обов'язкове створення словника виключень відкритого типу з забезпеченням можливості додавання до нього додаткових членів.

РОЗРОБКА МЕТОДУ РЕАЛІЗАЦІЇ УКРАЇНОМОВНИХ МОДЕЛЕЙ ВІДМІНЮВАННЯ ВЛАСНИХ НАЗВ

Правила відмінювання власних назв в українській мові є дещо складнішими, порівняно з російськими, за рахунок різноманітних неалгоритмізовано утворених закінчень запозичених слів та винятків [3], що, однак, не заперечує використання методу аналізу КПФ у процесі творення відмінкових форм. Проте в українській лінгвістиці цей метод потребує деякої модернізації.

Процес аналізу формозмінної власної назви починається з пошуку найбільш конкретизованої КПФ у бібліотеці кінцівок форм (КФ) та ідентифікації КФ шуканого відмінка. Якщо така знайдена, то реалізується перетворення початкової КФ відповідно до потрібної відмінкової форми. Якщо ж потрібна КПФ на першому етапі не знайдена, то продовжується пошук більш загального варіану у бібліотеці DLL. При цьому алгоритмічно передбачається можливість пошуку незмінних форм власних назв.

Розглянемо роботу лінгвістичного методу на прикладах відмінювання прізвищ, які початково задаються в називному відмінку. В бібліотеці відмінкових КФ наявні такі можливі варіанти КПФ власних назв називного відмінку однини [3]:

-га:	-ка:	-ха:	-шиплячий+а:	-приголосний+а:
Н.-> -га	Н.-> -ка	Н.-> -ха	Н.-> -шипл.+а	Н.-> -приг.+а
Р. -> -ги	Р. -> -ки	Р. -> -хи	Р. -> -шипл.+і	Р. -> -приг.+и
Д. -> -зі	Д. -> -ці	Д. -> -сі	Д. -> -шипл.+і	Д. -> -приг.+і
З.-> -гу	З.-> -ку	З.-> -ху	З.-> -шипл.+у	З.-> -приг.+у
О.-> -гою	О.-> -кою	О.-> -хою	О.-> -шипл.+ею	О.-> -приг.+ою
М.->-зі	М.->-ці	М.->-сі	М.->-шипл.+і	М.->-приг.+і

Нехай за вхідними даними маємо прізвище "Стоха" називного відмінку однини; вихідними даними передбачається отримання прізвища "Стоха" в давальному відмінку однини. За алгоритмом програмного засобу автоматизації перевіряється можливість співпадання останніх букв шуканого слова, які репрезентують собою КПФ, зі стандартним масивом варіантів з бібліотеки: -га, -ка, -ха і т. д. У наведеному прикладі шуканою КПФ буде КФ "–ха". Далі відбувається заміна ідентифікованої КФ обраного варіанту на її аналог шуканого відмінку "-сі". Таким чином, у вихідних даних отримуємо слово "Стосі".

Тепер проаналізуємо роботу лінгвістичного методу на прикладі ідентифікації прізвища "Сажа"в орудному відмінку однини. Зазначимо, що "ж" є шиплячим приголосним [4]. За алгоритмом методу проводиться послідовний пошук об'єкта у бібліотеці відмінкових КФ, результатом якого буде вибір

варіанту ”-шиплячий+а”. Реалізація лінгвістичної заміни КФ забезпечує отримання слова “Сажею” у вихідних даних програми. Зауважимо, що варіант “-приголосний+а” є найбільш загальним, що обумовлює його розташування в кінці лінгвістичного масиву бібліотеки КФ. Зрозуміло, що розташування КФ у бібліотеці DLL має бути жорстко структурованим, оскільки місце знаходження кожного варіанту пропорційно впливає на правильність ідентифікації об’єкта. Так, наприклад, розташування варіанта “-приголосний+а” у бібліотеці КФ перед варіантом “-шиплячий+а” обумовило б вибір і утворення помилкової форми слова - “Сажою”. А в першому прикладі розташування варіанта “-приголосний+а” перед потрібним “-ха” призвело б до отримання помилкового результату ”Стохі”. Алгоритми пошукових процесів форм власних імен та по-батькові реалізовані за аналогічними принципами. Основною вимогою ефективності алгоритмів є початкове впорядкування відмінкових КФ у бібліотеці DLL.

РОЗРОБКА МОДЕЛЕЙ РЕАЛІЗАЦІЇ УКРАЇНОМОВНОГО ЛІНГВІСТИЧНОГО МЕТОДУ ТВОРЕННЯ ВІДМІНКОВИХ ФОРМ ВЛАСНИХ НАЗВ В АВТОМАТИЗОВАНИХ СИСТЕМАХ

Модель україномовного лінгвістичного методу творення відмінкових форм (УЛМТВФ) власних назв повинна включати словник виключень (СВ), модуль порівняльного аналізу кінцівок відмінкових форм, алгоритми реалізації пошукової ідентифікації варіантів. На рис.1 наведено узагальнену модель УЛМТВФ, де МПІ – модуль порівняльної ідентифікації, який реалізує алгоритми розробленого методу пошуку варіантів шляхом порівняльного аналізу КФ власних назв з обов’язковим структурним упорядкуванням бібліотеки КФ; СВ – словник виключень; МВД – модуль введення даних; МВР – модуль відображення результату відмінювання.

Рис.1. Узагальнена модель УЛМТВФ власних назв

Модель реалізації алгоритмів пошукової ідентифікації запропонованого методу аналізу відмінкових форм власних назв розглянуто на прикладі аналізу лінгвістичної форми по-батькові (ФПБ) (рис.2,3).

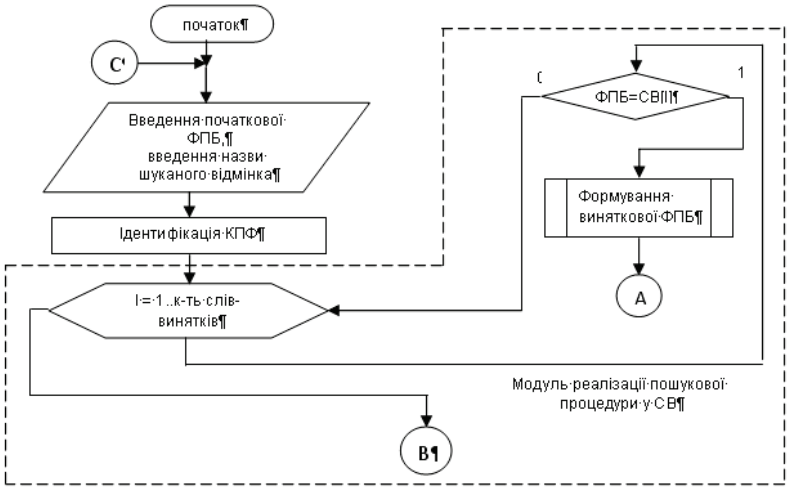

Рис. 2. Модель реалізації алгоритму МПІ на прикладі аналізу лінгвістичної ФПБ

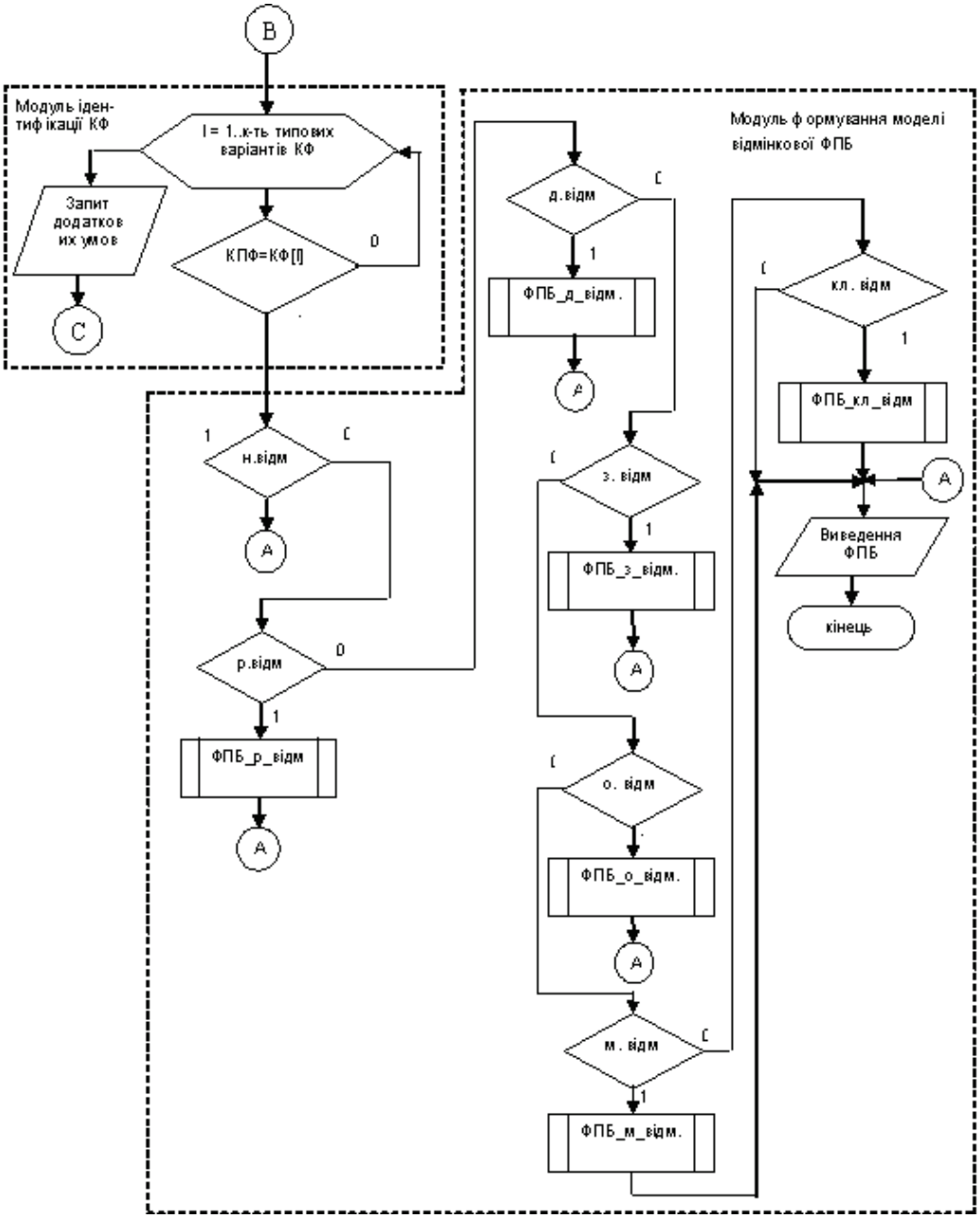

Рис. 3 Модель реалізації алгоритму МПІ на прикладі аналізу лінгвістичної ФПБ (закінчення)

Вхідними даними є початкова ФПБ в називному відмінку та відмінок потрібного словотворення. Процедури "ФПБ_р_відм" – "ФПБ_кл_відм" надають можливість реалізації перетворення ідентифікованої КФ початкової форми до вказаного у вхідних даних відмінку словотворення. Всі процедури мають ідентичну структуру та відрізняються лише масивом відмінкових КФ. Типова модель такої процедури розглянута на прикладі формування ФПБ у давальному відмінку (рис.4).

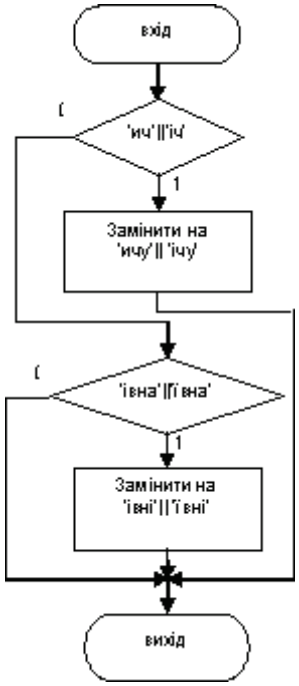

Рис. 4. Модель процедури перетворення кінцівки ФПБ давального відмінку

ВИСНОВОК

Розроблений лінгвістичний метод аналізу власних назв може забезпечити офісні, банківські, пошукові та інші інформаційні системи можливістю автоматизованої орфографічної перевірки власних назв та форм їх відмінювання. Запропоновані моделі здатні забезпечити україномовні пошукові системи ширшим простором пошуку інформації шляхом реалізації розроблених семантичних схем, що базуються на правилах української орфографії. Варто зазначити, що перспективним є створення лінгвістичних методів аналізу не лише власних назв, а й професійних та наукових термінів, географічних назв, та розширення можливості їх застосування для загального аналізу словотворних форм відмінкового характеру іменників, прикметників, займенників та числівників.

Надійшла до редакції 10.04.2008 р.

ЛІТЕРАТУРА

1. Бокс Д. Б78 Сущность технологии COM. Библиотека программиста. — СПб.: Питер, 2001. — 400 с.
2. Русский язык: Учебник для иностранных студентов подготовительных факультетов / Э.В.Витковская, Э.К.Горлова, Г. Е. Маевская и др. – Х.: Мир детства, 2002. – 320 с.
3. Академія наук України. Інститут мовознавства ім. О. О. Потебні. Інститут української мови. Український правопис, 6-е видання, виправлене й доповнене.- Київ: Наукова думка, 1997. – 240 с.
4. Українська мова. Підручник. Ч. 1/За ред. П. С.Дудика. – К.: Вища школа, 1993. – 415 с.

ВОЙТКО В.В. – к.т.н., доцент кафедри програмного забезпечення, Вінницький національний технічний університет, Вінниця, Україна.

БОЦУЛА М.П. – к.т.н., доцент кафедри моделювання і моніторингу складних систем, начальник центру дистанційної освіти, Вінницький національний технічний університет, Вінниця, Україна.

ЛУЦИШИН Г. Л. – студент 3–го курсу факультету комп'ютерного інтелекту інституту інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, Україна.