УДК: 681.3

О.Ю. Ніколаєвський, здобувач

АВТОМАТИЗАЦІЯ УКЛАДАННЯ КОМПОНЕНТІВ ЛІНГВІСТИЧНОГО ЗАБЕЗПЕЧЕННЯ МОДУЛЯ АВТОМАТИЧНОГО МОРФОЛОГІЧНОГО АНАЛІЗУ РІЗНОМОВНИХ ТЕКСТІВ

В статті розглянуто питання побудови автоматизованого робочого місця лінгвіста-дослідника. Запропоноване автоматизоване робоче місце дозволяє автоматизувати процес побудови компонентів лінгвістичного забезпечення для автоматичного морфологічного аналізу в системі машинного перекладу. Особливу увагу приділено опису можливостям інтерфейсу.

Ключові слова: автоматизоване робоче місце, автоматичний морфологічний аналіз, система машинного перекладу, аналітичний граматичний словник, лінгвістичне забезпечення.

The questions automated workstation linguist researcher are considered in the article. The workstation allows to ***automatize the process of building the components of linguistic software for automatic morphological analysis in machine translation system. Particular attention is paid to the description of interface features.***

Keywords: workstation, automatic morphological analysis, machine translation system, an analytical grammar vocabulary, linguistic support.

Відповідні словники, що забезпечують опрацювання вхідного тексту на морфологічному рівні мовної системи, входять до складу лінгвістичної бази даних і потребують спеціалізованого інструментарію для їх розробки відповідно до сучасного рівня лінгвістичних знань та інформаційних технологій. Аналітичні граматичні словники (АГС) є невід'ємним компонентом лінгвістичного забезпечення системи машинного перекладу (СМП), які призначені для автоматичного морфологічного аналізу вхідного тексту. Словникова стаття АГС містить в собі інформацію щодо частини мови та відповідних цій частині мові граматичних категорій (відмінок, рід, число, час, особа тощо). Структура словникової статті залежить від підходу до автоматизації морфологічного аналізу зокрема в СМП.

Аналіз сучасних систем машинного перекладу показав три принципово різних підходи щодо організації аналітичних граматичних словників. Лінгвістичне забезпечення за одним із підходів може формуватися на основі словника словоформ, але такий словник повинен мати не менше 0,5 млн. слів для флективних мов (російська, українська) і близько 150-200 тис. слів для аналітичних мов (англійська). Формування такого словника є дуже трудомістким процесом, який потребує багато часу та людських ресурсів, хоча для англійської мови він вважається прийнятним. Крім того, при цьому підході СМП не зможе аналізувати нові слова [1-4].

Другий підхід – формування АГС на основі словника квазіоснов та таблиці закінчень. Обсяг такого словника складає близько 100 тис. словникових одиниць, що теж є достатньо великою кількістю та потребує значного часу на його складання. За структурою словникова стаття такого АГС містить квазіоснову+відповідний словозмінний клас. Недоліком цього підходу також є неможливість автоматичного морфологічного аналізу нових слів, тобто слів які не належать до однієї парадигми. Третій підхід визначає на основі словника квазізакінчень всі закономірності словозмінення для певної вхідної даної мови [2]. Даний підхід дозволяє при обсязі АГС близько 5 тис. словникових одиниць аналізувати всі слова певної мови, включаючи і нові слова, що не можливо при інших підходах. Отже, доцільно будувати АГС на основі словника квазізакінчень в якості компоненти лінгвістичного забезпечення для автоматизації морфологічного аналізу в СМП [3].

Мета даної роботи – висвітлити досвід автоматизації формування компонентів лінгвістичного забезпечення для автоматичного морфологічного аналізу.

Викладення основних результатів дослідження. Автоматизація формування АГС значно може скоротити час підготовки компонентів лінгвістичного забезпечення, але для цього потрібно розробити спеціальні інструментальні засоби. Таким спеціальним засобом ми пропонуємо автоматизоване робоче місце лінгвіста (АРМ "ПАРАДИГМА"), яке призначено для формування аналітичних словників розпізнавання текстових одиниць на морфологічному рівні. Така віртуальна лінгвістична лабораторія розрахована на фахівця-мовознавця й надає необхідні засоби для швидкого й високоякісного створення аналітичних словників.

У розроблюваній СМП АРМ "ПАРАДИГМА" виступає як окрема система, яка дає змогу фахівцю-лінгвісту оптимізувати розроблення компонентів лінгвістичного забезпечення, забезпечити його функціональну повноту.

Результатом роботи АРМ "ПАРАДИГМА" є:

1) словник службових слів для відповідної вхідної мови;

2) словник квазізакінчень для відповідної вхідної мови.

АРМ лінгвіста підтримує формування словника словоформ для 3-х мов: української, російської, англійської. З цією метою розроблено єдину систему кодування для зазначених мов. Інтерфейс системи АРМ "ПАРАДИГМА" представлений на рисунку 1.

Класифікація, покладена в основу автоматичного морфологічного аналізу (АМА), зорієнтована на те, що результати слугують вихідними даними для автоматичного синтаксичного, лексичного та семантичного аналізів декількох мов. Список граматичних класів прийнятої в роботі класифікації складається із 27 виділених груп (лексико-граматичних класів) [5,6]. Звісно, виокремлюємо традиційні частини мови: іменник, дієслово, прикметник, числівник, прийменник, сполучник та ін. Визначаємо: артикль, герундій.

Уведені зміни до АРМ "ПАРАДИГМА" стосуються класів:

- дієслова – в окрему групу виділяємо дієслова минулого часу, неминулого часу, інфінітив, вказуючи на особливості керування, дієслова наказового способу;
- в окремі групи виділяються дієслівні форми – дієприкметник та дієприслівник;
- числівника (виокремлено числівник-іменник, числівник-прикметник);
- для займенників ураховується як характер їх значень, так і специфіка словозміни та функціонування в мові. Відповідно, розглядаються займенники, що мають прикметниковий тип відмінювання (деякий, кожний, всякий), займенники-іменники (вона, він), особові займенники;
- для прийменник додається лексико-граматична категорія керування.

© Ніколаєвський О.Ю., 2012

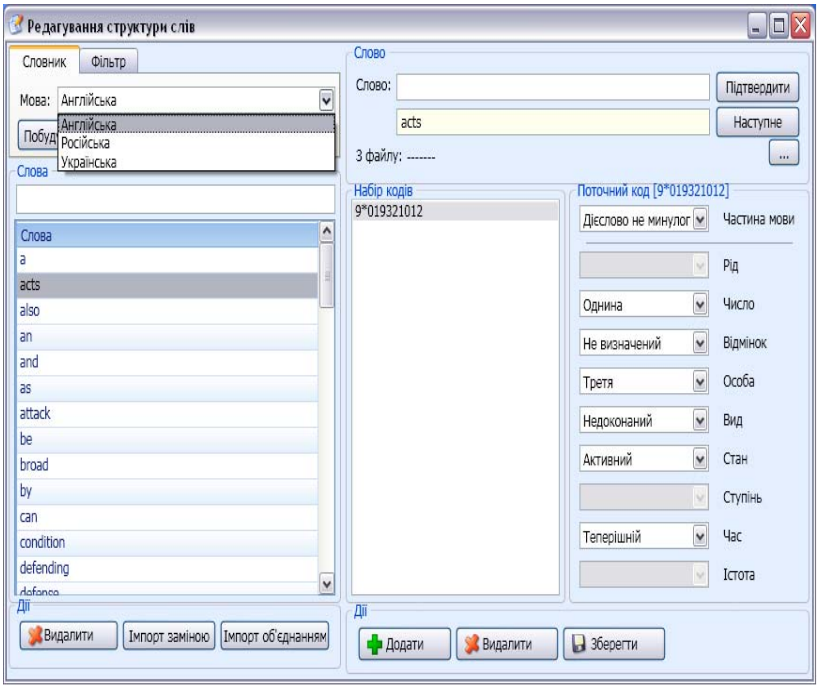


Рис. 1

Щоб сформувати словник квазізакінчень, який би забезпечував апріорну повноту, необхідно набрати достатню вибірку словоформ з текстів певної вхідної мови. З цією метою до АРМ ведено функції:

- формування словника словоформ за текстом, при цьому якщо дана словоформа вже є словнику АРМ, то вона не завантажується до вікна користувача;
- формування словника словоформ безпосередньо набором із вікна користувача. Це зручно, коли потрібно ввести до базового словника словоформи, що виключенням з певного правила.

Отриманий словник словоформ формується за текстами різної жанрово-тематичної спрямованості. У разі необхідності залучались дані з граматик відповідних мов. Основою пропонованого принципу є розроблений метод, який спирається на позиційно-цифрове кодування граматичної інформації в словниковій статті. У такий спосіб кожна аналізована словоформа отримує свій код, що містить інформацію про її частиномовну приналежність та конкретне граматичне значення.

При введенні нового слова лінгвіст визначає йому лексико-граматичний клас (див. рис. 2).

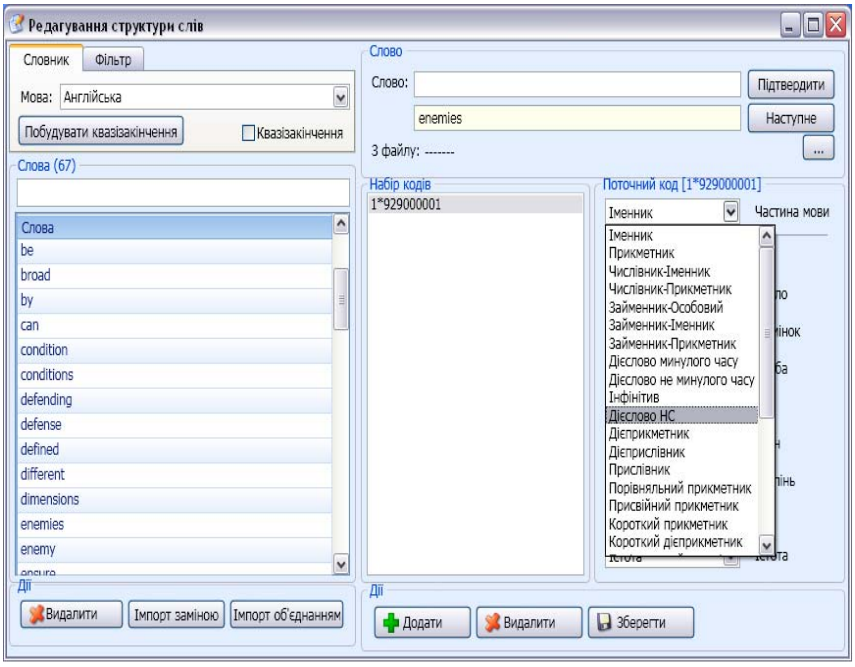


Рис. 2

Для дослідницьких цілей в АРМ передбачено автоматичну побудову словників (на основі базового словника):

- словника службових слів;
- оберненого словника, який є основою для побудови синтетичних словників квазізакінчень: лематизаційного і парадигматичного;
- окремих словників для кожного лексико-граматичного класу, це дає змогу зручного корегування і виправлення помилок при визначенні кодів оператором;
- словника квазізакінчень.

Повнота словника квазізакінчень визначається експериментальним шляхом. Вважається, що словник є

повним, якщо при додаванні нових словоформ до базового словника, словник квазізакінчень не змінюється.

Після визначення лексико-граматичного класу автоматично з'являються лише ті граматичні категорії, які визначаються для даного лексико-граматичного класу, інші граматичні категорії заблоковані (див. рис. 3). Це дозволяє уникнути зайвих помилок при кодуванні граматичної інформації з боку лінгвіста.

Крім того, АРМ має можливість:

- фільтрації даних за різноманітними граматичними ознаками;
- експорту даних в текстовий формат за прямим та інвертованим впорядкуванням;
- імпорту даних з різних текстових форматів (що допомагає лінгвісту вводити та перевіряти дані).

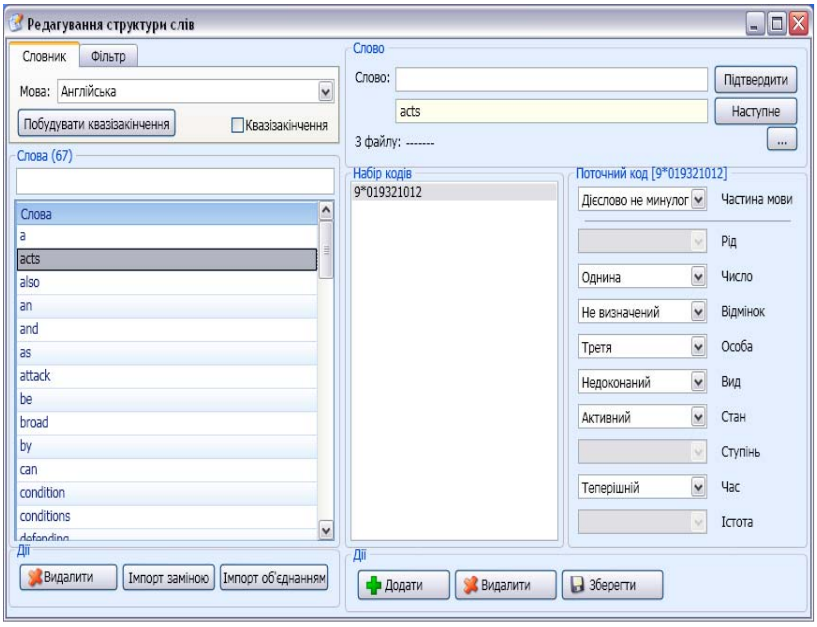


Рис. 3

Для дослідницьких цілей в АРМ передбачено автоматичну побудову підсловників на основі базового словника (режим словник на рис. 4):

- словник службових слів (спеціальний словник на рис. 4), до даного словника входять всі службові частини мови;
- низка словників за окремими частинами мови. Це є зручною функцією при редагуванні базового словника, оскільки для словозмінних частин мови можна відслідкувати всю парадигму слова: перевірити коректність коду (праве віконце на рис. 4), додати відсутні словоформи тощо;
- словник словоформ, що мають декілька граматичних класів, тобто словник слів, яким притаманна міжчастинно-мовна омонімія.

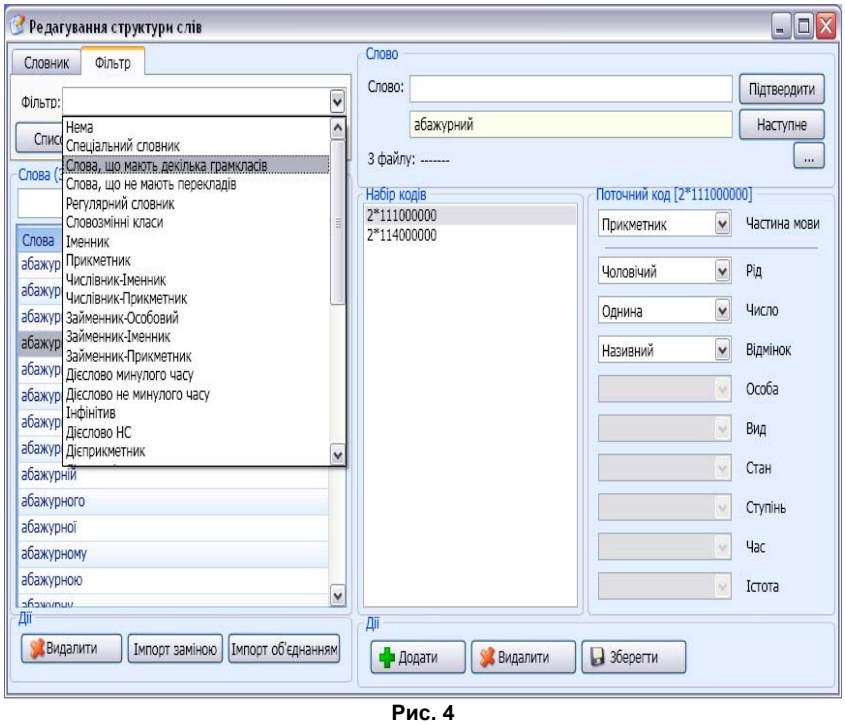


Рис. 4

Крім того (див. рис. 5), АРМ має такі функції, як:

• автоматичне формування базового словника в прямому алфавітному порядку (режим СПИСОК), з подальшим його виведенням в текстовий формат;

• автоматичне формування базового словника в зворотному алфавітному порядку, тобто з кінця слова (режим ІНВЕРТОВАНИЙ СПИСОК);

• автоматичне формування базового словника, в якому слова представлені в перевернутому форматі (режим ПЕРЕВЕРНУТИЙ СПИСОК).

Перераховані словники є основою для побудови синтетичних словників квазізакінчень: лематизаційного і парадигматичного.

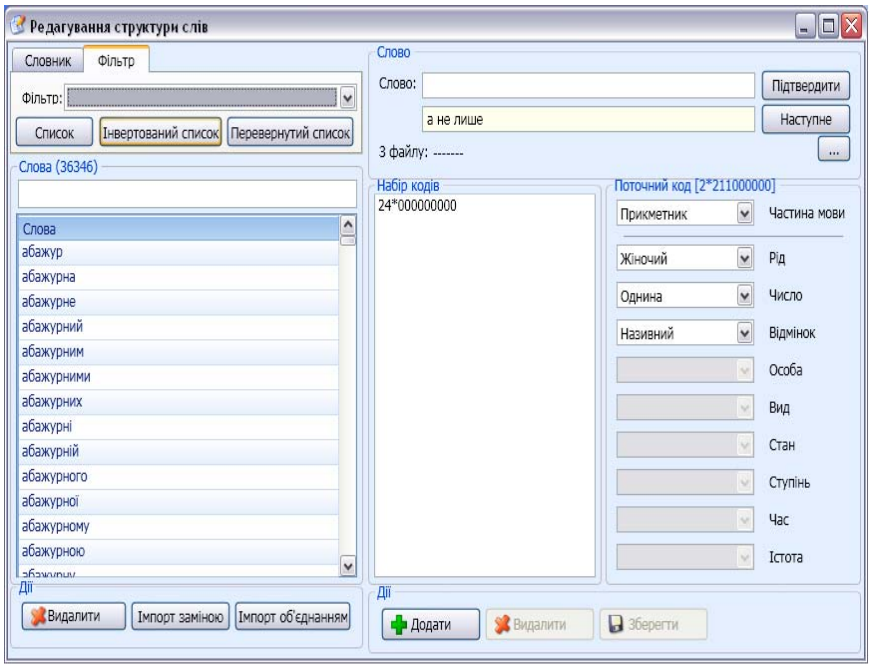


Рис. 5

Як вже зазначалося, кінцевим результатом є автоматичне формування словника службових слів та словника квазізакінчень (режим КВАЗІЗАКІНЧЕННЯ, див. рис. 6), які безпосередньо представлені у вигляді аналітичних граматичних словників (АГС) як складових лінгвістичного забезпечення автоматичного морфологічного аналізу. Словник квазізакінчень будується на основі базового словника для кожної вхідної мови. На мал. 6 у правому віконці представлений фрагмент квазізакінчення для російського базового словника словоформ.

Для автоматичної побудови словника квазізакінчень необхідно вибрати мову та запустити процедуру ПОБУДУВАТИ КВАЗІЗАКІНЧЕННЯ.

Крім того, АРМ дає можливість експорту даних в текстовий формат за прямим та інвертованим впорядкуванням (режим ЕКСПОРТ); можливість імпорту даних з різних текстових форматів (режими ІМПОРТ ЗАМІНОЮ та ІМПОРТ ОБ'ЄДНАННЯМ). Введення цих режимів дозволяє розпаралелити роботу фахівців-лінгвістів по різним робочим місцям з подальшим об'єднанням результатів їх роботи.

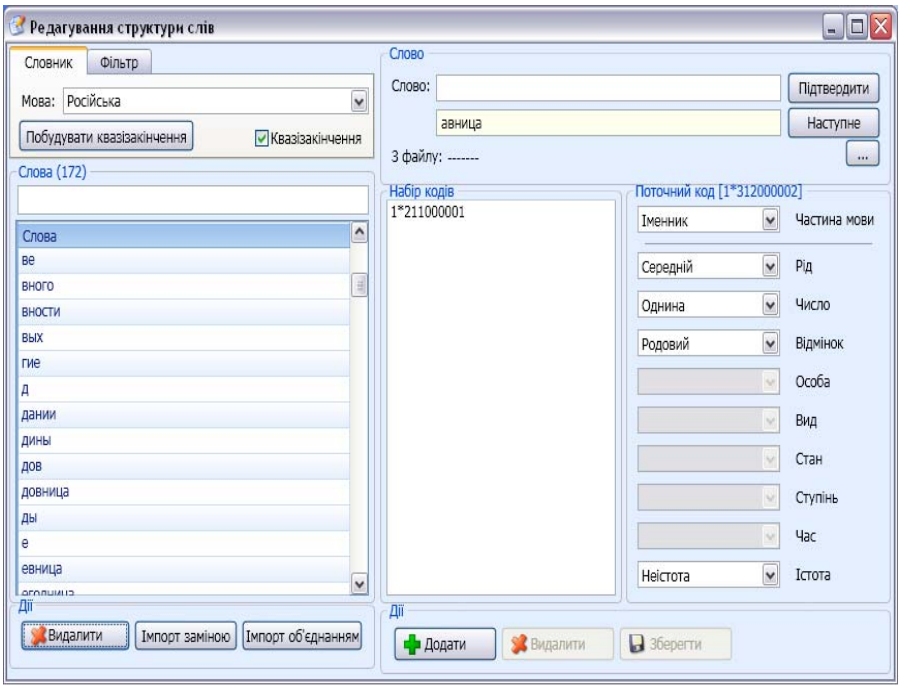


Рис. 6

Висновки. Власне для автоматизації роботи фахівця-лінгвіста пропонується АРМ (автоматичне робоче місце) лінгвіста, особливістю якого є уніфіковане представлення граматичної інформації для трьох мов. З цією метою визначаються частини мови та інші морфологічні характеристики (герундій, артикль, допоміжне дієслово, тощо), що охоплюють перелік можливих характеристик для всіх мов вказаної групи.

1. Оцінка ефективності системи автоматичного морфологічного аналізу (АМА) залежить від обсягів словника, що застосовується, несуперечливості інформації, швидкості опрацювання текстів і можливості аналізу нових слів.

2. Розроблене автоматизоване робоче місце лінгвіста дозволяє максимально оптимізувати розробку аналітичного словника для автоматичного морфологічного аналізу а саме: застосований принцип флективного аналізу на основі позиційно-цифрового кодування допомагає зменшити обсяги словника квазізакінченість, забезпечує компактність збереження лінгвістичних даних, а відтак дозволяє підвищити ефективність системи автоматичного морфологічного аналізу; робить АМА відкритим, оскільки уможливлює аналіз "невідомих для системи" нових слів; способи поповнення бази словоформ – "уведення слова вручну" та "опрацювання повнотекстових масивів" – надає морфологічній моделі змісту, який може відповідати лінгвістичній реальності.

3. Поповнення словника словоформ – як дослідного масиву (базового словника) для формування списку квазіфлексій – завдяки створеному АРМ лінгвіста можна здійснювати двома шляхами: безпосередньо вводячи словоформу вручну або ж за допомогою текстового файлу. Така можливість дозволяє спиратися як на знання мови, скажімо, вводячи унікальні класи словозміни вручну, так і на реальні тексти, які дозволяють реалізувати знання-орієнтовний підхід і до формування морфологічної моделі мови.

4. Єдина система параметризації граматичної інформації дозволяє за допомогою одного АРМ (програмного модуля) опрацьовувати англомовні, російськомовні та україномовні тексти, що є однією з головних вимог до розробки багатомовної системи машинного перекладу.

1. Гельбух А. Ф., Сидоров Г. О. К вопросу об автоматическом морфологическом анализе флективных языков // www.dialog-21.ru/Archive/2005. 2. Лазарева О.Я. Метод формирования словаря квазиокончаний // Інформатизація та нові технології. – № 1. – 1997. – С.9-12. 3. Грязнухіна Т. О., Нікула М. В. Система автоматичного морфологічного аналізу українського наукового тексту // Пробл. україніізації комп'терів. Матеріали 2-ї міжнар. конф. – Київ, 1993. – С. 42-46. 4. Морфологический анализ научного текста на ЭВМ. – К.: Наук. думка, 1989. – 264 с. 5. Замаруєва І.В., Шипнівська О.О. Морфемна обробка текстів в системах машинного перекладу. // Вісн. КНУ ім. Тараса Шевченка. Військово-спеціальні науки. – К., 2008. – №20. – С.61 – 63. 6. Балабін В.В., Замаруєва І.В., Лєнков С.В., Пампуха І.В.Технологічні аспекти реалізації автоматизованих систем машинного перекладу.// Зб. наук. пр. ВІКНУ. – К., 2010. – № 26. – С. 55 – 64.

Надійшла до редколегії 06.06.12

УДК 81'322.2

О.О. Шипнівська, канд. філол. наук

СТВОРЕННЯ БАЗИ ДАНИХ ДЛЯ РОЗРОБКИ АВТОМАТИЧНОГО СИНТАКСИЧНОГО АНАЛІЗАТОРА УКРАЇНОМОВНИХ ТЕКСТІВ (НА МАТЕРІАЛІ ПРОСТОГО РЕЧЕННЯ)

У статті представлено основні засади та специфіку розробки лінгвістичної бази даних для побудови автоматичного синтаксичного аналізатора україномовних текстів на матеріалі простого речення. Головна увага зосереджена на можливостях розроблюваного дослідницького середовища.

Ключові слова: автоматичний синтаксичний аналіз, лінгвістична база даних, просте речення, синтаксична структура речення.

In the paper the instrumental environment for the automatic syntactical analysis of Ukrainian is described. The main characteristics of databases for simple sentence their advantages for the automatic syntactical analysis and general principles of automatic syntactical analysis are presented.

Key words: formal syntactical structure, general principles of automatic syntactical analysis, linguistic databases for automatic syntactical analysis, simple sentence.

1. Автоматичний синтаксичний аналіз: проблеми та перспективи їх вирішення. Проблема розробки алгоритмів автоматичного синтаксичного аналізу є однією з ключових при автоматичному опрацюванні текстів (АОТ) природної мови. Незважаючи на значний доробок у цій царині як теоретичного, так і практичного плану, все ще залишається затребуваним напрацювання відповідних процедур. Актуальною є ця проблема для україномовних текстів у контексті побудови знання-орієнтованої системи машинного перекладу (ЗСМП).

Завдання автоматичного синтаксичного аналізу полягає в побудові формально-синтаксичної структури тексту, представлення речення в синтаксичних категоріях [Апресин 1978; Синтаксический анализ 1999; Дарчук 2005]. У нашій роботі автоматичний синтаксичний аналіз розуміємо як незалежний модуль АОТ текстів природної мови, що покликаний аналізувати та синтезувати всі можливі варіанти синтаксичної структури речення, зокрема україномовного [Замаруева 1999].

Розробка модуля автоматичного синтаксичного аналізу акумулює низку складних практичних та теоретичних проблем [Апресин 1978; Анно 1975; Гладкий 1985; Синтаксический анализ 1985; Баталина 2001; Кобзарева 2007]. Представлені в публікаціях сучасні дослідження проблеми стосовно слов'янських мов зосереджують свою увагу головно на підготовці та побудові синтаксично розмічених корпусів текстів, розробці необхідних для цього засобів аналізу [Кобзарева 2004, Шимкова 2005].

Головним теоретичним підґрунтям автоматичного синтаксичного аналізу, зазвичай, виступає граматика залежностей та граматика безпосередніх складників [Синтаксический анализ 1999]. Представлений у праці підхід, крім цих визначних методів, зорієнтований ще й на врахування комунікативного аспекту синтаксису української мови. Однією з вихідних точок аналізу є поняття правильної синтаксичної структури [Иорданская 1964, с. 215]. Правильною є структура, а у нашій праці, лінійна модель речення, яку може зрозуміти носій мови, не знаючи конкретної семантики слів.

Українській мові як одній із флективних мов притаманна велика кількість типів лінійної структури простих та складних речень. У мовленні можна спостерігати багато структурних компонентів та їх комбінацій. Такий високий рівень синтаксичної розгалуженості ускладнює процеси аналізу, синтезу та моделювання механізмів автоматичного породження тексту. Застосування процедури автоматичного синтаксичного аналізу продемонструвало, що багато синтаксичних одиниць, таких як

© Шипнівська О.О., 2012