АВТОМАТИЗОВАНЕ НАПОВНЕННЯ ЛІНГВІСТИЧНОЇ БАЗИ ДАНИХ

Алєшко Є.О., Борисова Н.В.
Національний технічний університет «Харківський політехнічний інститут», м. Харків

Сучасне лінгвістичне дослідження навряд чи можливе без залучення лінгвістичних інформаційних та мовних ресурсів. Важливе місце серед них займають лінгвістичні бази даних, що містять структуровану інформацію про мовні одиниці різних рівнів (від морфеми до тексту) і різноманітну інформацію про ці одиниці. Можна виділити дві основні сфери їх застосування: 1) забезпечення функціонування різних автоматизованих систем, пов'язаних з обробкою тексту і мови (інформаційні, експертні, навчальні системи, системи аналізу мови, машинного перекладу та ін.); 2) автоматизація лексикографічної діяльності різного призначення і підготовки словників різного типу (навчальних, перекладацьких, нормативних, тлумачних та ін.).

Лінгвістичні бази даних мають певні особливості: розгалуженість структури, відхід у підлеглі таблиці, багато незаповнених лакун і наявність розгалуженої системи зв'язків між таблицями. Визначені особливості дозволили обрати цей вид інформаційного ресурсу для збереження інформації про слова-запозичення української мови. Дослідження цього мовного явища є необхідним та актуальним, оскільки згідно з сучасними лінгвістичними дослідженнями, запозичених слів в українській мові зараз 10-15% її лексичного складу. І хоча виникнення таких слів зумовлюється або тим, що у мові немає необхідних слів, що відображають те чи інше явище чи предмет, або тим, що існуючі слова з тієї чи іншої причини не влаштовують носіїв мови, їх використання має бути виправданим та збагачувати мову, а не засмічувати її. У зв’язку з цим завданнями даної роботи є дослідження особливостей слів-запозичень, що функціонують в українській мові, аналіз переваг та недоліків існуючих ресурсів, що зберігають слова-запозичення, та визначення механізму наповнення лінгвістичної бази даних таких слів. Аналіз недоліків існуючих ресурсів дозволив розробити наступний алгоритм автоматизованого наповнення лінгвістичної бази даних слів-запозичень:

1. Вибір слів-запозичень за визначеними правилами з існуючих електронних орфографічних словників української мови.
2. Вибір визначень слів-запозичень з існуючих тлумачних словників української мови.
3. Занесення обраної зі словників інформації у відповідні таблиці лінгвістичної бази даних. При цьому мова, з якої запозичене слово, вноситься у відповідне поле в залежності від того, яке правило вибору слова спрацювало.
4. Слово-оригінал (тобто слово вихідної мови, з якої відбулося запозичення) додається у відповідну таблицю вручну.
5. Поле «Ступінь запозичення» (повністю або частково) також заповнюється вручну.

Для реалізації лінгвістичної бази даних обрано СУБД Microsoft Access.