УДК 004.65
Шклярук М., Карнаухов О.
Тернопільський національний технічний університет імені Івана Пулюя

ПРОЕКТУВАННЯ ЛІНГВІСТИЧНИХ БАЗ ДАНИХ

Науковий керівник: к.т.н., доцент Козак Р. О.

Shkliaruk M., Karnaukhov O.
Ternopil Ivan Pul'uj National Technical University

LINGUISTIC DATABASES DESIGN

Supervisor: Kozak R. O.

Ключові слова: лінгвістика, база даних, автоматичний словник
Keywords: linguistics, database, automatic dictionary

Процес автоматичного опрацювання тексту є складним і ресурсозатратним, тому для його оптимізації варто проводити одноразове повне опрацювання тексту, результати якого можна використовувати для таких задач як анотування текстів, машинний переклад, автоматичне реферування та інших. Найбільш доцільно для цього створювання лінгвістичні бази даних та бази знань.

Знаннями, які використовують в інтелектуальних системах, є спеціальним чином організовані дані. Для їх використання у лінгвістичних системах ці знання формалізують з використанням математичного апарату. Залежно від виду та характеру залученого математичного апарату представити знання можна у різний спосіб.

Дослідники приводять два етапи проектування бази даних: інфологічний – відбір інформації та її структурування, моделювання змісту інформації, та датологічний – оформлення інформації мовою представлення, яка придатна для комп'ютерного опрацювання (перетворення інформації на дані). При проектуванні лінгвістичної бази даних завданням першого етапу є створення концептуальної інформаційної моделі предметної галузі, а другого етапу – зовнішня формалізація мовних об'єктів.

Залежно від обраних джерел формування розрізняють словниково- та текстозорієнтовані лінгвістичні бази даних. При побудові реально існуючих баз даних їх склад, структура та принципи зазвичай визначаються конкретними цілями проектування.

Основним компонентом лінгвістичної бази даних є автоматичний словник, в якому зберігається основна інформація для реалізації алгоритмів. Основним елементом автоматичного словника є стаття, яка містить усю інформацію про характеристики даної лінгвістичної одиниці. Вона характеризується трьома факторами, які є взаємозумовленими та визначають ефективність лінгвістичної бази даних. Цими факторами є обсяг лінгвістичної інформації, що закладається у словникову статтю, спосіб її компонування у словниковій статті та організація самої словникової статті.

Отже, процес проектування лінгвістичних баз даних з одного боку використовує загальні принципи побудови баз даних, а з іншого боку має певні особливості, які пов'язані з завданнями автоматичного опрацювання тексту. В даному процесі необхідно застосовувати як комп'ютерні, так і лінгвістичні знання, для того щоб побудувати максимально узгоджену, повну та ефективну для практичного застосування базу даних, яка буде оптимально розв'язувати покладені на неї завдання.