

АВТОМАТИЗАЦІЯ ПОБУДОВИ ТЕЗАРУСУ

Яхимович О. В.

Науковий керівник – д.т.н., проф. Бісікало О. В.

На сьогодні одним з найважливіших завдань лексикографії є проектування таких словників, які б на рівні світових стандартів задовольняли велику потребу сучасної інформатизованої спільноти в систематизованій лінгвістичній інформації.

Кінцевим продуктом розробки є саме комп'ютерний тезаурус як найоптимальніша за впорядкуванням лінгвістичного матеріалу та швидкодією система, а тому потребує удосконалення формалізована, алгоритмізована методика його побудови.

Тезаурус – це термін, який широко використовується в галузі інформатики як складова частина інформаційно-пошукових систем. Можна виділити два визначення тезауруса:

1. Тезаурус – це словник, що відображає семантичні відношення між лексичними одиницями дескрипторної інформаційно-пошукової мови (дескрипторами) і призначений для пошуку слів за їх смисловим змістом.

2. Тезаурус – це контрольований словник термінів предметної області, який створюється для поліпшення якості інформаційного пошуку в певній предметній сфері.

Для визначення відношень у тезаурусі було розроблено програму. Вона призначена для знаходження скільки разів те чи інше слово зустрічається у визначенні. Результатом роботи програми є таблиця, де рядки – це слова-визначення, а стовпці – слова, які містяться у визначенні. На перетині рядків і стовпців стоїть цифра, що вказує скільки разів слово, у всіх його словоформах, зустрічається у даному визначенні. Також є можливість залишити рядки і стовпці утвореної таблиці, в яких є хоча б один елемент більший або рівний за число вказане користувачем.

Кожне слово-визначення зберігається в окремому файлі, де назва файлу є слово-визначення, а вміст – всі його значення. Слова, які не відносяться до повнозначних частин мови, або є словами виключеннями: займенниками, числівниками, також зберігаються в окремому файлі і при аналізі не додаються до таблиці. Словоформи слів теж зберігаються в окремому файлі.

Створення тезаурусів, словників термінів дотепер залишається вкрай складною й трудомісткою роботою, ступінь автоматизації якої дуже низький. По суті, всі тезауруси створюють вручну. Автоматично може перевірятися лише узгодженість накопичених визначень. Альтернативою міг би бути підхід, коли визначення понять створюють за наявними текстами, а потім, у разі необхідності, корегують в процесі діалогу з експертом. Для реалізації такого підходу необхідно вміти здійснювати докладний аналіз семантики текстів.