# РОЗРОБКА ЕЛЕКТРОННОГО ТЛУМАЧНОГО СЛОВНИКА ТЕРМІНІВ З КОМП'ЮТЕРНОЇ ЛІНГІСТИКИ НА ОСНОВІ КОРПУСУ ТЕКСТІВ

**Скуменко С.І.**
*Національний технічний університет*
*«Харківський політехнічний інститут»*
*г. Харків, ул. Пушкінська, 79/2, тел. 707–63–60*
e-mail: *s.relaxed@yandex.ua*

Дане дослідження присвячене вивченню однієї з наук, що динамічно розвивається – комп'ютерна лінгвістика, яка є новою гілкою прикладного мовознавства.

Спеціалізований тлумачний словник термінології комп'ютерної лінгвістики належить до словників комп'ютерного типу, який має базу даних і в якому передбачено різні режими його використання: пошук і систематизація інформації, яка використана у дефініціях, а також можливість перетворення словника на дослідну, навчальну або фактографічну базу даних. Комп'ютерний спеціалізований тлумачний словник термінів з комп'ютерної лінгвістики репрезентує терміносистему цієї галузі у її сучасному стані, на рівні передових лексикографічних й інформаційних технологій і є першим досвідом її систематизації та стандартизації.

Під терміном "комп'ютерна лінгвістика" (computational linguistics) зазвичай розуміється широка область використання комп'ютерних інструментів - програм, комп'ютерних технологій організації та обробки даних - для моделювання функціонування мови в тих чи інших умовах, ситуаціях, проблемних областях, а також сфера застосування комп'ютерних моделей мови не тільки в лінгвістиці, а й у суміжних з нею дисциплінах [1].

"Термін" комп'ютерна лінгвістика "задає загальну орієнтацію на використання комп'ютерів для вирішення різноманітних наукових і практичних завдань, пов'язаних з мовою, ніяк не обмежуючи способи вирішення цих завдань" [2].

Словотвір - постійний шлях поповнення лексичними засобами будь-якої мови і підмови, що дає найбільш значущий результат у кількісному відношенні. Способи словотворення, які традиційно виділяють в системі мови це: афіксація, конверсія, словоскладання, абревіація.

Афіксація в українській та англійській термінології активно використовує приставки латинського походження, характерні для загальнонаукової мови (табл. 1).

Корпус текстів - це зроблена за певними правилами вибірка з проблемної області, тобто під корпусом текстів розуміється великий, структурований і оброблений спеціальним чином масив мовних даних кінцевого розміру, призначений для вирішення різних лінгвістичних завдань. Всі тексти, що входять в масив об'єднані деяким логічним задумом, логічної ідеєю [3].

Таблиця 1 - Приклади приставок латинського походження, як одного із способів словотворення

| Афікс | Приклад |
|---|---|
| inter- | interactive (интерактивный) |
| super- | superuser (привилегированный пользователь) |
| mini- | minidriver (минидрайвер) |
| macro- | macrocommand (макрокоманда) |
| micro- | microfile (микрофайл) |
| auto- | autodump (авторазгрузка) |
| multi- | multisystem (мультисистема) |
| mega- | megaword (мегаслово) |
| re- | recopy (повторно копіювати);<br>recreate (відновлювати дані);<br>reformat (пере форматувати); |
| e- | e-mail (електронна пошта);<br>e-book (електронна книга);<br>e-cash (електронні гроші);<br>e-form (електронна форма, електронний бланк). |

Для вирішення різних лінгвістичних завдань мало лише наявності масиву текстів. Потрібно також, щоб тексти містили в собі різного роду додаткову лінгвістичну та екстралінгвістичну інформацію.

Розмітка полягає в приписуванні текстам і їхнім компонентам спеціальних міток: зовнішніх, екстралінгвістичних (відомості про автора і відомості про текст: автор, назва, рік і місце видання, жанр, тематика; відомості про автора можуть включати не тільки його ім'я, але також вік, стать, роки життя та багато іншого; це кодування інформації має назву метарозмітка), структурних (глава, абзац, речення, словоформа) і власне лінгвістичних, що описують лексичні, граматичні та інші характеристики елементів тексту.

Серед лінгвістичних типів розмітки виділяються [4]:

а) морфологічна розмітка. В іноземній термінології вживається термін part-of-speech tagging (POS-tagging), дослівно - частиномовна розмітка. Насправді морфологічні мітки включають не тільки ознаку частини мови, але й ознаки граматичних категорій, властивих даній частині мови.

б) синтаксична розмітка, є результатом синтаксичного аналізу, або парсинга (англ. parsing), виконуваного на основі даних морфологічного аналізу. Цей вид розмітки описує синтаксичні зв'язки між лексичними одиницями і різні синтаксичні конструкції (наприклад, підрядне речення, дієслівне словосполучення тощо).

в) семантична розмітка. Хоча для семантики немає єдиної семантичної теорії, найчастіше семантичні теги позначають семантичні категорії, до яких належить дане слово чи словосполучення, і вужчі підкатегорії, які специфікують його значення.

Розмітка корпусів - це трудомістка операція, особливо враховуючи розміри сучасних корпусів. Якщо деякі види розмітки, здійснюються вручну, то морфологічний та синтаксичний аналіз можливо здійснити автоматично [5].

Підсумковим завданням даної роботи є розробка алгоритму автоматичної розмітки тексту та вибору з тексту можливих термінів англійської комп'ютерної лінгвістики на основі корпусу текстів.

На першому етапі вручну виділяємо найпоширеніші префікси, суфікси та основи, за допомогою яких утворюються терміни комп'ютерної лінгвістики, з якими у подальшому будуть порівнюватися словоформи.

Нерозмічений лінгвістичний текст автоматично розбивається на окремі словоформи, які порівнюються з морфемами та основами, які найчастіше використовуються для словотворення термінів комп'ютерної лінгвістики.

Якщо збіг не знайдено, програма виводить на екран «Нажаль у цьому тексті немає термінів комп'ютерної лінгвістики.»

Якщо програма знайшла збіг, вона автоматично помічає його тегом <ap></ap>, який вказую, що це слово є терміном комп'ютерної лінгвістики.

На виході ми отримаємо розмічений текст, який можна занести до корпусу текстів, та надати можливість користувачеві переглянути імовірні терміни англійської термінології з комп'ютерної лінгвістики та дізнатися їх дефініцію.

Створення корпусів текстів значно полегшило збір і зберігання інформації. Це дуже цінується при створенні словників, глосаріїв, лексикографічних робіт. Також спосіб зберігання корпусів текстів дозволяє надійніше і довше зберігати будь-який мовний матеріал, що є важливою знахідкою та інструментом в лексикографії.

**Список літератури**

1. Баранов А.Н. Введение в прикладную лингвистику. - Эдиториал УРСС, 2001.
2. Довідково-інформаційний портал: http://www.gramota.ru/slovari/types/17_2
3. Карпіловська Є. А. Вступ до комп'ютерної лінгвістики. - Донецьк, 2003
4. Апресян, Ю.Д., Иомдин, Л.Л., Санников А.В., Сизов, В.Г. Семантическая разметка в глубоко аннотированном корпусе русского языка. - Издательство Санкт-Петербургского университета, 2004.
5. Баранов, А. Н. Автоматизация лингвистических исследований: корпус текстов как лингвистическая проблема. - Просвещение, 1998.
6. Труды Международного семинара по компьютерной лингвистике и ее приложениям «Диалог-2000», «Диалог-2001», «Диалог-2002», «Диалог-2003», «Диалог-2004», «Диалог-2005».