Я. А. Єрмолова

Частотні словники та їх використання

Сучасне мовознавство все ширше використовує математичний інструментарій та комп'ютерні технології в контексті переходу від описових до аналітичних методів досліджень. Відповідно до цього сучасна філологічна освіта немислима без опертя на обчислювальну лінгвістику як для осмислення отриманих наукових результатів, так і для проведення самостійного лінгвістичного експерименту з використанням чисельних методів (Б. Л. Ван дер Варден).

Найбільш поширеним і доступним кількісним методом аналізу тексту є статистичний аналіз, який полягає у підрахунку кількості вживань у лексичному складі заданого тексту окремих слів. Статистичний аналіз широко використовується для:

• математично точного розрізнення літературних стилів і жанрів (статистична стилістика);

• встановлення авторства анонімних або підроблених текстів (атрибуція тексту);

• опису поведінки різних мовних одиниць (букв, морфем, слів) у тексті (їх розподіл, сполучуваність, частота вживання);

• вимірювання інформативності текстів (кількість інформації, що міститься в тексті та в його складових частинах);

• відновлення текстів та мов за їхніми фрагментами;

• визначення рівня спорідненості, швидкості мовних змін і часу поділу різних мов.

Основою проведення будь-якого статистичного аналізу тексту є частотні словники (М. П. Концевой).

Відомості про найбільш частотні й комунікативно важливі слова тієї чи тієї мови значно розширюють можливості як успішного викладання іноземної мови, так і більш глибокого оволодіння рідною мовою. Частотні словники є також основою для створення електронних словників, комп'ютерних перекладачів, систем семантичного пошуку, автореферування й автоанотування текстів і т.п. Сьогодні для прискорення і полегшення статистичних досліджень у мовознавстві та літературознавстві широко застосовуються електронні частотні словники. Однак, як правило, вони є дорогим програмним продуктом, а тому не завжди доступні. Тому в навчальному процесі доцільно використовувати частотні словники, створені на основі окремих текстів з використанням загальнодоступних програмних засобів.

Частотні словники мають велику сферу застосувань. На основі ЧС можна визначити автора непідписаного рукопису, оскільки ЧС різних текстів настільки індивідуалізовані, що можуть діагностувати автора, якщо зіставити ЧС непідписаного рукопису з ЧС текстів тієї людини, яку ми вважаємо автором непідписаного рукопису. Звідси випливає використання ЧС у криміналістиці для визначення авторства анонімних текстів.

Аналіз ЧС документів певної організації може розкрити напрямок спрямування її діяльності. Такий аналіз широко застосовується в політиці, економіці, соціології.

Статистичні (імовірнісні) методи можна застосовувати для фільтрації спаму. Для цього використовуються частотні словники, створені в процесі навчання фільтру. Береться архів старих вручну відсортованих повідомлень і обробляється програмою навчання. Вона складає частотні словники для кожного типу повідомлень і вимірює відстані між заданим текстом і класом спам-текстів (Л. Н. Беляєва).