

Василь Войнов, Сергій Тимофєєв
Харківський національний університет ім.В.Н.Каразіна

АРМ ЛЕКСИКОГРАФА: КРИТИЧНИЙ АНАЛІЗ СУЧАСНОЇ КОНЦЕПЦІЇ ЛЕКСИЧНОЇ КАРТОТЕКИ

© Василь Войнов, Сергій Тимофєєв, 2002

У роботі розглядаються два підходи до використання комп'ютера у складанні лексиконів. Основна ідея : комп'ютер – помічник укладальника словника, який накопичує текстові факти у постійній довготривалій пам'яті, оперативно показує людині усі проміжні результати і формує остаточний доробок – словник у звичному людині вигляді.

In this report the using of computers in lexicographical studies is considered. The main idea: computer assists to human student but it doesn't replace him utterly.

Автоматизована лексична картотека (АЛК) – це інформаційна система, що забезпечує створення лексичних карток і роботу з ними для заданого масиву текстів [1, с.242]. Лексична картотека є найважливішим інструментом вивчення лексики цікавлячого нас тексту. “Лексична картотека як достатньо повний набір прикладів вживання лексичних одиниць має настільки велике значення для філологічної науки, що вже сама проблема її формування набуває самодостатнього характеру ...” [1, с.236].

В.А.Широков [1, с.235–248] наводить опис сучасної концепції АЛК. Тексти, що аналізуються, вводяться в комп'ютер. За допомогою особливих алгоритмів сегментації тексти членуються на сегменти – інтервали, що не перекриваються, що містять цілком контексти всіх ужитих у сегменті лексем. Створюється конкорданс – список уживань усіх словоформ (крім стоп-слів) з їхніми координатами в текстах. За допомогою алгоритмів лематизації конкорданс перетворюється у словник – основний компонент бази даних АЛК. За допомогою алгоритмів парадигматизації формуються морфологічні парадигми лексем. На основі інформації, закладеної в БД, формується її користувальницьке представлення у формі масиву лексичних карток. Ілюстративна зона кожної картки, що містить спочатку деякий автоматично обраний сегмент, піддається ручному постредагуванню для скорочення до точного контексту заголовної лексеми. Тепер АЛК готова до використання.

Описана концепція базується на 2-х основних принципах: 1) автоматичної сегментації і 2) автоматичного формування записів лексем у словнику. Однак автор роботи [1] Широков В.А. визнає, що сучасні алгоритми сегментації, незважаючи на складність, далекі від досконалості і тому не можна обійтися без ручного постредагування сегментів. Але це взагалі ставить під сумнів корисність автоматичної сегментації: адже лексикографу не має значення, відкіля вичленовувати текст, чи з сегмента, чи прямо з тексту, потрібно тільки, щоб комп'ютер надав усі вживання лексеми.

Далі, автоматичне формування записів лексем, засноване лише на алгоритмах лематизації і парадигматизації, призводить до функціональної обмеженості АЛК: не реалізуються лексеми різних рівнів, гіперлексеми і складені лексичні одиниці (словосполучення), тобто лексичні одиниці (ЛО) різних типів [2, с.180]. Потрібні додаткові алгоритми, а це ще більш ускладнить систему.

Основними недоліками концепції [1] ми вважаємо функціональну обмеженість і невиправдану складність реалізації. Головну причину зазначених недоліків ми бачимо в неправильному розподілі функцій між людиною і комп'ютером. Щодо сегментації це виявляється в спробі «перекласти на плечі» машини задачу семантичного аналізу тексту. Щодо автоматичного формування парадигм: вони використовуються тільки для пошуку входжень лексем, граматичні ознаки словоформ по суті не потрібні, тобто суттєвим є лише перелік словоформ парадигми. Щодо автоматизації формування основної форми лексеми: її уведення вручну вимагає незрівняно менше часу, аніж аналіз входжень. Таким чином, без алгоритмів сегментації, парадигматизації і лематизації в АЛК цілком можна обійтися.

Необхідно виходити з призначення АЛК як інструмента аналізу лексики. Відповідно до цього ми пропонуємо більш загальне визначення: «АЛК – це інформаційна система, що забезпечує доступ до всіх контекстів уживання необхідних лексичних одиниць і їхніх характеристик у заданому масиві текстів».

Перегляд поняття електронної лексичної картки.

Для здійснення доступу до контекстів уживань ЛО зовсім немає необхідності у формуванні лексичних карток (далі – карток) старого зразка – їх ми будемо називати «картками-виписками», досить швидкого переходу в ті місця тексту, де розташовані вживання ЛО. Але на екрані комп'ютера зручно представити інформацію про входження ЛО, як і раніше, у формі картки. Таку картку ми назвемо картокою-закладкою, тому що вона імітує роботу з закладками в книзі. Зона ілюстрації цієї картки є віконцем у те місце тексту, де ужита ЛО. Для зручності ЛО виділяється особливим кольором і розташовується в центрі. Лексикограф може «прокрутити» текст у віконці в обидва боки на будь-яку відстань, таким чином контекст ЛО не зобов'язаний міститися цілком на картці. Картки-закладки формуються автоматично для усіх уживань ЛО – треба лише задати її парадигму, як це описано нижче. Але ми не відмовляємося і від карток-виписок, що дійсно необхідні на завершальному етапі аналізу лексики для фіксації результатів. Формування карток-виписок – трудомістка ручна робота, як у нашій концепції, так і в концепції [1]. Лексикограф повинен переглянути усі вживання ЛО (тобто картки-закладки) і залишити лише ті, котрі в найбільш зручному вигляді відбивають лексико-семантичні варіанти й особливості уживання аналізованих ЛО. Кожна картка-закладка перетвориться в картку-виписку в такий спосіб: лексикограф виділяє блоком контекст ЛО у віконці картки-закладки і натискає відповідну кнопку на панелі керування програми АЛК. Створюється картка-виписка, у якій можна дореддагувати контекст, а також занести коментар щодо даного вживання в зону паспорта картки. Ми вважаємо корисним існування обох типів карток, причому для картки-виписки завжди доступна картка-закладка: якщо кінцевого користувача АЛК не задовольнить контекст, підготовлений для нього експертом-лексикографом, він завжди може звернутися до тексту джерела.

Формування парадигм.

Лексична одиниця будь-якого типу в нашій концепції АЛК визначається своєю основною формою, номером омоніма і парадигмою. Парадигма простої ЛО складається зі словоформ, парадигма складеної ЛО – з послідовностей словоформ. Принциповою тут виявляється наступна обставина. Лексикографу не треба задавати усі форми ЛО і, тим більше, вводити їх у комп'ютер. Потрібно лише відзначити на екрані в списку всіх словоформ тексту фактично присутні для парадигми, яка формується. У випадку

словосполучень така операція повторюється для кожної складової. Зазначена процедура не обтяжлива для філолога: усі словоформи перед очима, нічого не треба відтворювати по пам'яті і тим більше вводити в комп'ютер. Зрештою в словнику буде створений запис ЛО в складі її основної форми і списку посилань на усі вживання. Кожне посилання відповідає картці-закладці. Якщо в подальшому для якогось посилання буде створена картка-виписка, то її контекст і коментар будуть також занесені в запис ЛО. Запис має додатково службові поля - частоту входжень ЛО, дату формування, ім'я лексикографа і т.п. Через граматичну омонімію, щойно сформований запис ЛО може містити посилання на вживання інших ЛО, тому необхідне «очищення» списку посилань, що виконується в процесі аналізу усіх уживань ЛО. Для словникових омонімів передбачена спеціальна команда, що дублює запис ЛО і встановлює в ній номер омоніма.

Пропонована загальна структура АЛК.

АЛК складається з програми і бази даних. Програма може працювати з декількома БД, реалізуючи тим самим кілька картотек. БД складається з масиву текстів, конкордансу і словника. Кожному масиву відповідає його екранна форма. Масив текстів складається з пронумерованих образів джерел лексики «текстів» і їх БО. Текст є послідовністю квазислів [1], так що вказівкою на входження словоформи є пара (№ тексту, № квазислова). Посиланням на вживання ЛО є посилання на її перше слово. Конкорданс має форму інвертованого індексу з записами виду (словоформа, список посилань). У його вікні здійснюється перегляд і оцінка словоформ парадигми; одна словоформа є поточною – для неї можна обійтися списком посилань і побачити ділянки тексту з уживаннями словоформи. Словник містить всю інформацію про сформовані лексеми і складається з описаних вище записів ЛО. Екран словника в лівій частині містить список ЛО, права частина екрана має вигляд лексичної картки.

Процес створення і функціонування АЛК. Етапи створення АЛК.

Створення картотеки здійснюється в 3 етапи: підготовчий етап: сканування, розпізнавання і редагування текстів джерела. Результат – набір файлів WORD' а. Тексти повинні бути ретельно відредаговані, тому що наступне їхнє виправлення не допускається через можливість зміни нумерації квазислів. Етап виконується без використання програми АЛК; технологічний етап: створення порожньої БД, імпорт до неї текстів, їхнє додаткове структурування; створення конкордансу, загалом – формування структури доступу до БД по словоформам.. Етап може бути виконаний допоміжним персоналом; основний етап : створення і редагування записів лексем, заповнення лексикографічних параметрів і формування контекстів. Етап вимагає участі кваліфікованого експерта - лексикографа.

Режими роботи програми АЛК.

Програма АЛК проектується таким чином, щоб не встановлювати жорстку послідовність дій при створенні картотеки. Експерт планує роботу, як йому зручно: він може спочатку створити певну групу заготовок лексем, потім їх усі відредагувати і, нарешті, окремим кроком сформувати лексичні картки. Альтернативою є полексемна обробка. Гнучкість планування операцій дозволяє розподілити роботу між кількома експертами. У той же час є можливість задання бажаного сценарію роботи і тоді програма буде підказувати експерту послідовність дій. У зв'язку з вищесказаним, програму АЛК зручно описувати не алгоритмом, а діаграмою режимів. Режим – це стан програми, що характеризується змістом екрана і набором доступних операцій, виконуваних у будь-якій

послідовності. Переходи між режимами задаються або експертом, або підпрограмою «майстром».

Висновки.

На основі чіткого усвідомлення мети створення АЛК і природної вимоги мінімізації витрат на її виготовлення, виявилось можливим значно спростити структуру системи й обійтися без програмування складних лінгвістичних алгоритмів.

АЛК є підсистемою автоматизованого робочого місця (АРМ) лексикографа, проблеми створення і використання якого досліджуються в лабораторії лексикографічного моніторингу української мови на філологічному факультеті ХНУ ім.В.Н.Каразіна.

1. Широков В.А. *Інформаційна теорія лексикографічних систем*. К.: Довіра, 1998. – 331с. 2. Грязнухина Т.А., Дарчук Н.П. і ін. *Використання ЕОМ у лінгвістичних дослідженнях*. – К.: Наук. думка, 1990. – 226.

Софія Денисенко

Український державний лісотехнічний університет, м.Львів

ЛЕКСИКОГРАФІЯ ТА МІЖМОВНІ ЗВ'ЯЗКИ В УКРАЇНІ НА ПОЧАТКУ ХХІ СТ.: НОВІ ЛЕКСИКОГРАФІЧНІ ТВОРИ

© Софія Денисенко, 2002

Укладання словників - вид діяльності, яка вимагає високого професійного інтелекту та самовідданості у важкій праці, що неможливо без натхнення і творчого пошуку. Вершиною лексикографічного мистецтва вважається тлумачний словник, який найбільш систематизовано і послідовно відображає ступінь розвитку лексичного складу мови. Тлумачний словник ніби відкриває межі між рівнями мовної системи, він конкретно представляє всю лінгвістику в цілому.

The information about “German – Russian – Ukrainian with Greco-Roman etymology veterinary thesaurus” of prof. Khomitsenko has been given in this article.

Початок ХХІ ст. знаменується продовженням інтенсивного вивчення лексикографії в Україні, зокрема навчальної перекладної лексикографії. Навчальні перекладні словники допомагають тим, для кого мова не є рідною, вивчати іноземну мову на різних етапах навчання. Цьому сприяє також розвиток навчальної лексикографії як науки.

Укладаються нові перекладні словники різноманітних типів, такі як "Німецько-російсько-український з греко-латинською етимологією тлумачний зооветеринарний словник" проф. Хоміцького О.В.

Перейдемо до характеристики цієї праці. Як відомо, тлумачними словниками називаються лінгвістичні словники, у яких пояснюються (тлумачаться) значення слів і фразеологізмів якої-небудь мови засобами самої цієї мови. Тлумачні словники є, як правило, комплексними: їхні словникові статті крім певних значень слова вміщують дані про його написання, особливості вимови, місце наголосу, граматичні словотворчі властивості, ті стилістичні функціональні сфери до яких належить слово, типові сполучення і фразеологічні вирази, в яких воно може вживатися, його синтагматичні і парадигматичні