Інститут української мови
Національної академії наук України

Лексикографічний бюлетень

Випуск 13

Київ 2006

УДК 61.2.81’373’374
Л 43 Лексикографічний бюлетень: Збірник наукових праць
Заснований у 1951 р.
Відновлений у 2004 р. випуском 10

РЕДАКЦІЙНА КОЛЕГІЯ
Відповідальний редактор
В. Німчук, *д. філол. н., професор, чл.-кор. НАН України*
Заступник відповідального редактора
О. Демська-Кульчицька, *к. філол. н.*
Відповідальні редактори випуску
В. Перебийніс, *д. філол. н., професор*
Н. Дарчук, *к. філол. н.*
Відповідальний секретар
О. Тищенко, *к. філол. н.*

А. Бурячок, *д. філол. н., професор*
К. Городенська, *д. філол. н., професор*
С. Єрмоленко, *д. філол. н., професор*
Л. Ставицька, *д. філол. н.*

Рекомендовано до друку ухвалою вченої ради
Інституту української мови НАН України
(протокол № 5 від 30 травня 2006 р.)

ISBN 966-02-3450-3 (серія)
ISBN 966-02-4002-3 (вип. 13)

Збірник висвітлює питання, пов’язані з історією лексикографії, методикою й технологією укладання словників різних типів, аналізом українського та іноземного словникарства, комп’ютерною лінгвістикою, а також із проблемами системного й комунікативного аналізу мовного матеріалу з погляду лексикографії.

Випуск 13, зокрема, присвячений корпусній лінгвістиці, комп’ютерному розпізнаванню й синтезу усної мови, лінгвістичним процесорам, комп’ютерній лексикографії – основним темам Міжнародної конференції з прикладної (комп’ютерної) лінгвістики (27–28 квітня 2006 р., Київ; організатори: Інститут української мови НАНУ, Інститут філології КНУ ім. Тараса Шевченка, Київський національний лінгвістичний університет).

Для науковців, студентів-філологів.

© Інститут української мови НАН України

ЗМІСТ

ПРИКЛАДНА (КОМП'ЮТЕРНА) ЛІНГВІСТИКА 5

Корпусна лінгвістика 5
Демська-Кульчицька О. Національний корпус української мови: концептуальний аспект 5
Кукушкина О., Поликарпов А., Токтонов Г. Корпусная неография и неология: системный анализ характеристик лексических неологизмов (на материале газетного раздела «Полистилевого корпуса современного русского языка»)..10
Коцюк Л. Використання корпусів текстів у процесі укладання словників.....15
Сидорко Т. Рекламний текст як особлива одиниця корпусу даних 19
Лозова Н. Дослідження дієслів повідомлення на базі робочої версії Національного корпусу української мови 22
Монахова Т. Корпус творів Валерія Шевчука: формат розмітки базових концептів26
Grabowski Ł. Comparative Characteristics of Polish and Russian Language Corpora.....29
Дарчук Н., Сорокін В. Корпус текстів як джерело для мовознавчих і літературознавчих досліджень.....34
Зубань О. Параметризована база даних як інструмент дослідження корпусу текстів37
Старко В. Фразеологія, лексикографія і корпус.....43
Тищенко О. Засади створення корпусу української жестової мови глухих.....47
Тименко Л. Наукова термінологія в національному корпусі української мови.....52
Кузнєцова А. Параметризація публіцистичних неолексем у корпусній лінгвістиці .54
Garabík R. Computer(ized) linguistic resources at the Ľ. Štúr Institute of Linguistics.....56
Radić-Bojanić B. Methodology of register analysis: computer-mediated communication60
Ющенко Р., Гудзенко В. Применение параллельных компьютеров для автоматизации морфологической разметки 63
Загнітко А., Данилюк І., Ситар Г., Щукіна І. Принципи і закономірності укладання бази "Українські прийменники та їхні еквіваленти".....66
Карчова А., Шимкова М. Морфологічна анотація текстів словацького національного корпусу 71
Лучкув И. Польская лингвистическая литература XIX – нач. XX вв. в цифровом формате. На пути к созданию библиографии польской лингвистики в фондах интернет-библиотек.....76

Лінгвістичні процесори 81
Карамишева І., Карамишева Р. Аналіз морфемного складу неологізмів засобами сучасних інформаційних технологій (на основі англійської мови).....81
Юзепольська Т. Парадигматичні класи іменника англійської мови (категорії числа: аналіз та синтез).....83
Дембіцький В. Ступені порівняння прикметників і прислівників. Синтез і розпізнавання в тексті парадигматичних класів англійських прикметників та прислівників 85
Бугаков О. Встановлення зон прийменникових зв'язків в українському тексті (модуль автоматичного синтаксичного аналізу).....91
Ерреро Бікус Н. Аналіз і синтез перфектних форм англійського дієслова активного стану.....95
Чаварга Я. Семантичний потенціал прикметників англійської мови.....97
Курстак С. Алгоритм зняття омонімії типу «іменник-дієслово» в англійській мові та імплементація його в програмному продукті.....102
Прыстаўка А. Тыпалогія семантычных пераўтварэнняў тэкстаў.....105
Коваль Н., Лангенбах М. Парадигматичні і синтагматичні характеристики прийменникових та безприйменникових дієслівних сполучень.....109

Комп'ютерне розпізнавання й синтез усної мови.....112
Бардіна Н., Каташинська І. Системність лінгвістичного аналізу при проведенні ідентифікаційних експертиз.....112
Глізнуца Н. Вимова довгих та коротких приголосних як ідентифікаційна ознака .115
Близнюк В. Моделювання фонемної структури лексем сучасної італійської мови 118

Дудник З. Сегментація на акустичні склади для компілятивного синтезу англійського мовлення.....123
Наумова Н. Ідентифікація користувачів англійської мови у діловому дискурсі 125
Володкевич О. Створення й використання бази даних складів англійського мовлення для блоку лінгвістичної обробки текстів у системі конкатенативного синтезу.....128
Біскуб І. Лінгвістична параметризація текстів для тренування користувачів у системі автоматичного розпізнавання англійського мовлення Dragon Naturally Speaking.....130
Філіппова Н. Когнітивні засади аналізу концепту „прикладна лінгвістика" як спеціальності 134

Лінгвістичні портали, інформаційний пошук.....137
Дарчук Н., Сірук О., Сорокін В. Лінгвістичний портал MOVA. Info: три роки в мережі Інтернет 137
Могилевский В., Королева Т. Критериальная модель соотнесенности интенсивности семантического признака с количественным проявлением его акустических коррелятов.....143

Комп'ютерна лексикографія.....146
Калинчук Д. Проблеми перекладу при укладанні навчального італійсько-українського словника 146
Лазарєва О. Про можливість автоматичного укладання термінологічних словників за фаховими текстами.....149
Шульгіна В. Проблеми комп'ютерної лексикографії у галузі інформаційної діяльності.....151
Перебийніс В., Назаров С., Бобкова Т. Комплекс електронних навчальних словників „ГЛОСА".....153
Шуневич Б. , Паплик О. Проект англійсько-українського комп'ютерного словника з робототехніки.....156
Сірук О. Комп'ютерний тезаурус дієслів української мови: проблеми та методика укладання 158
Васильєва І. Використання комп'ютерного тезауруса в дослідженні мови поетів.161
Бук С. Статистичні характеристики лексики основних функціональних стилів української мови: спроба порівняння.....166
Андрейчук Н., Волошиновська І. Дієслова, що визначають глибину опрацювання предмета дослідження, та їх статистичні характеристики у науково-технічних текстах.....170
Балог В. Передумови створення корпусу словників.....174

ЛЕКСИКОГРАФІЯ, ЛЕКСИКОЛОГІЯ: ТЕОРІЯ ТА ПРАКТИКА.....177
Цимбалюк-Скопненко Т. Контамінація як структурно-семантичний спосіб трансформації фразеологічних одиниць (на матеріалі перекладів Миколи Лукаша) 177
Осінчук Ю. Історія назв добових богослужінь 181
Щипська М. Особливості використання фразеологізмів у мові сучасної поезії 188
Тищенко О. Концепція і принципи укладання "Словника української жестової мови глухих" 192
Прядко І. Фразеологічні синоніми з нумеративним компонентом в українській мові.....195

РЕЦЕНЗІЇ.....200
Поликарпов А. Об одной рецензии 200

РІЗНЕ 206
Лексикографічний бюлетень 206

ПРИКЛАДНА (КОМП'ЮТЕРНА) ЛІНГВІСТИКА

КОРПУСНА ЛІНГВІСТИКА

О. Демська-Кульчицька, к. філол. н.*
Інститут української мови НАНУ, (Київ)
УДК 161.2.81'374.72'22.322

НАЦІОНАЛЬНИЙ КОРПУС УКРАЇНСЬКОЇ МОВИ: КОНЦЕПТУАЛЬНИЙ АСПЕКТ

У статті висвітлено засади формування та формалізації матеріалів Національного корпусу української: обґрунтовано принципи добору текстів, визначено кількість та обсяг текстових фрагментів, подано адаптовану систему кодування первинних даних.

Сучасні мовознавчі дослідження сьогодні активно послуговуються корпусним методом і для усіх європейських мов, серед яких не забуваємо про слов'янські, існує по кілька корпусів текстів різного типу, обсягу, структури, наповнення та призначення. Й узагальнивши понад сорокалітню історію становлення корпусної лінгвістики та побудови корпусів, мовознавці визнали, що створення корпусу є „обов'язком щодо рідної мови" [4: 5].

Створення корпусів на сьогодні виформувало в межах загального напряму корпусної лінгвістки певний субнапрям, а саме *теорію і практику побудови текстових корпусів. В*иокремлення такого типу досліджень передовсім умотивовано тим, що процедура побудови текстового корпусу, в термінах корпусної лінгвістики, до моменту його експлуатації як лінгвістичного об'єкта, а також використання як програмістського ресурсу, вимагає розв'язання низки наукових проблем, зокрема типології корпусу, його призначення, обсягу, параметризації предметної галузі, репрезентативності, структурування та принципів відбору базових одиниць зберігання etc, тобто йдеться про процедуру концептуалізації корпусу, в нашому випадку – *Національного корпусу української мови (НКУМ).*

Національний корпус української мови: теоретико-практичний аспект є науково-дослідною темою Інституту української мови НАН України, затвердженою Вченою радою ІУМ НАНУ від 26.01.06 протокол № 1, Бюро відділення ЛММ НАНУ від 02.03.06 протокол № 2. Це ставить ряд вимог до концепції *НКУМ. Н*асамперед ідеться про ідеологію проекту, яка чітко окреслює некомерційний, науковий статус корпусу, призначеного для здійснення наукових досліджень різних рівнів сучасної української мови в Інституті української мови НАН України. Здійснення цих досліджень сучасної української мови накладає особливі вимоги щодо параметризації предметної галузі корпусу, яка повинна базуватися на традиційній параметризації українських текстів у лінгвоукраїністиці, та щодо необхідності його морфологічної розмітки, яка також за основу має брати кваліфікації слів, вироблені в межах української граматичної традиції. Тобто науковий проект *НКУМ* повністю ґрунтується на досягненнях української мовознавчої науки.

Національний корпус – це зібрання текстів, що репрезентують національну мову на певному етапі(ах) її існування в усьому різноманітті жанрів, стилів, історичних, територіальних і соціальних варіантів. Електронний текстовий корпус додатково передбачає машиночитану форму існування текстового ресурсу, стандартне подання й можливість застосування у дослідженнях мови та в програмуванні. Виходячи зі сказаного, визначаємо **Національний корпус української мови (НКУМ)** як *перетворену на електронну форму, організовану згідно з корпусними стандартами, програмно опрацьовану, репрезентативну вибірку текстів сучасної української мови, призначена для лінгвістичного застосування. Л*інгвістична орієнтованість передбачає здійснення академічних досліджень а) української мови, зокрема лексики, фразеології,

* © *О. Демська-Кульчицька, 2006*

термінології, морфології та синтаксису, словотвору, орфографії тощо; б) методики викладання / навчання / вивчення української мови як рідної та іноземної.

З огляду на призначення, типологію й коло завдань *НКУМ*, описуваний корпус повинен бути корпусом:

- **загальнонародної мови:** з урахуванням територіальної специфіки як у межах України, так і поза ними;
- **одномовним:** тексти, що ввійшли до корпусу, є результатом мовної діяльності носіїв української мови;
- **дослідницьким:** зорієнтованим на широкий спектр лінгвістичних завдань;
- **динамічним:** тобто передбачати постійне поповнення множини корпусних текстів;
- **синхронним:** має охоплювати рівень сучасної української мови;
- **фрагментним:** збудованим із текстових фрагментів, тобто уривків текстів, відібраних за попередньо визначеними засадами відбору текстових даних до корпусу;
- **мішаним:** передбачати введення до корпусу писемних і усних текстових фрагментів;
- **морфологічно розміченим:** усі текстові дані розмічені до рівня слова, і кожне слово передбачає маркування частиномовної належності та відповідних морфологічних форм.

Концепція довільного корпусного проекту вимагає параметризації предметної галузі, репрезентованої корпусом. Оскільки національний корпус є зібранням електронних текстів, які репрезентують національну мову на певному етапі(ах) її існування в усьому різноманітті жанрів, стилів, територіальних і соціальних варіантів, то предметною галуззю *НКУМ* буде українська загальнонародна мова в таких формах її існування, як літературна мова, територіальний діалект і соціальний діалект.

Найповніше в *НКУМ* повинна бути представлена літературна мова – основна наддіалектна форма існування природної мови, ознаками якої є опрацьованість, унормованість, поліфункціональність, стилістична здиференційованість, фіксованість. Найважливішим аспектом літературної мови є її співвіднесеність (унаслідок розвиненої здиференційованості) з усіма сферами людської діяльності, що, відповідно, забезпечує всі основні типи суспільної інформації.

Щодо літературної мови, то не підлягає сумніву потреба її відображення в корпусі національного типу. Натомість вимагає додаткової аргументації введення діалектних та соціолектних текстів.

Діалектна лексика, репрезентована діалектними текстами, як зазначає О. Гердт, є частиною лексики сучасного усного мовлення, однією з синхронних форм реалізації мовної системи, і, якщо йдеться про репрезентативність корпусу національного типу, то введення діалектного мовлення до його структури є беззаперечним. Важливою характеристикою територіального діалекту – різновиду національної мови, якому властива відносна структурна близькість і який є засобом спілкування людей, об'єднаних спільною територією, елементами матеріальної й духовної культури, традиціями й самосвідомістю, – це функціонування в синхронії та одночасно належність до діахронії. Належність до синхронії та діахронії власне є основним аргументом на користь уведення діалектного матеріалу до корпусу. Ще одним аргументом на користь цього є те, що будь-який діалект значно старший за будь-яку літературну мову. Крім того, територіальні діалекти охоплюють як найдавнішу питому мовну специфіку, так і результати інноваційних процесів, потенційних для літературної мови. Наявність діалектного матеріалу в *НКУМ* дає змогу досліджувати фактичний матеріал літературної мови в зіставленні з діалектними даними в межах однієї корпусної побудови.

Уведення до *НКУМ* текстів, які б репрезентували соціальні діалекти, пов'язано, з одного боку, з загальною лінгвістичною тенденцією до „термінологічної універсалізації соціально маркованої лексики" [1: 47], на що вказує Л. Ставицька, а з іншого – із завданням оптимальної репрезентації мови в корпусі загальномовного типу. Тобто йдеться про понятійне визначення соціально маркованої лексики і принципи відбору соціолектного текстового матеріалу. До соціальних діалектів, за Л. Ставицькою, зараховуємо арго (особлива мова певної відокремленої професійної чи соціальної групи, яка складається з видозмінених елементів однієї або двох природних мов), жаргон (напіввідкрита лексико-фразеологічна підсистема, яку застосовують та чи та соціальна група з метою відособлення від решти мовної спільноти) і сленг (різновид розмовної

мови, яку суспільство оцінює як підкреслено неофіційну („побутова”, „фамільярна”, „довірлива”).

Очевидно, що в межах *НКУМ* соціолектні тексти не повинні кількісно переважати діалектні тексти а становити лише незначний відсоток від загального обсягу корпусу. Тут також варто зазначити, що соціолектні тексти матимуть переважно звукову форму, тобто входитимуть до усної частини корпусу.

Уведення до *НКУМ* діалектних і соціолектних текстів додатково зумовлює й літературна традиція вплетення в полотно художнього твору, відповідно, діалектних і соціолектних елементів для естетичної та культурно-змістової виразності твору. Крім того, наявність у корпусі цих форм загальнонародної мови є одним із шляхів досягнення репрезентативності *НКУМ.*

За аналогією до *Британського національного корпусу*, який розглядаємо як еталонний у сучасній корпусній лінгвістиці, вважаємо за доцільне, особливо на етапі впровадження напряму корпусної лінгвістики в українську мовознавчу традицію, визначити хронологічні межі *НКУМ* межами сучасної української літературної мови. Тобто від останніх років XVIII ст. і до сьогодні, залишаючи поза увагою всі попередні періоди існування української мови, а саме: давньоукраїнський період від середини XI ст. до кінця XIV ст., ранньосередньоукраїнський – від початку XV ст. до середини XVI ст., середньоукраїнський – від середини XVI ст. до перших років XVIII ст., пізньосередньоукраїнський – від середини і до кінця XVIII ст.

Змістова параметризація предметної галузі *НКУМ* передбачає визначення джерельної текстової бази корпусу з подальшою стилістично-жанровою стратифікацією відібраних текстів. Враховуючи традицію відбору та організації фактичного матеріалу для лінгвістичних досліджень у лінгвоукраїністиці, до *НКУМ* повинні передусім увійти літературно опрацьовані масові друковані тексти, які визначає багатство лексики й нормативність слововживання. Важливою ознакою таких текстів є відображення в них не лише писемного варіанта мови, а й усного. До основних вимог текстових джерел належать: а) адекватне представлення сучасної української мови в усіх найважливіших сферах функціонування; б) охоплення всіх жанрів, у яких реалізуються функціональні стилі мови; в) орієнтація на масового читача; г) достатність щодо кількісних параметрів. Остання вимога є надзвичайно важливою, оскільки вона, з одного боку, впливає на встановлення реального лексичного складу мови, а з іншого – на забезпечення кожного мовного факту достатньою кількістю контекстів. Таким чином, опрацюванню підлягатимуть: художня проза, драматургія, поезія, публіцистика, наукова, науково-популярна, суспільно-політична література, офіційні документи, підручники, посібники й листи.

Стилістично-жанровий аспект текстового ресурсу вимагає додаткового опису. Так, стильова диференціація мови можлива, по-перше, за характером мовної експресії, а по-друге, за характером суспільної функції мови. За характером мовної експресії традиційно розрізняють високий, середній і низький стилі. В основі такої тристильової диференціації лежить концепція залежності між предметом викладу, тематикою і добором мовних засобів та жанрів. Цей поділ, успадкований європейською традицією доби Відродження й бароко з александрійської філософської школи античного періоду, застосовано до параметризації предметної галузі в багатьох національних корпусах європейських мов.

В українській традиції тристильова диференціація мови, яку, зокрема, розробляли Ф. Прокопович, М. Довгалевський, Г. Кониський, втратила актуальність у процесі історичного розвитку стилістичної системи української мови, коли три традиційні стилі (‘слоги’) занепали, а основними одиницями стильової диференціації української мови стали функціональні стилі. Таким чином, в українську лінгвістичну традицію увійшов функціональний принцип стилістичної диференціації мови, і така диференціація покладена в основу змістової параметризації предметної галузі *НКУМ*.

Сучасна українська мова на рівні стильової диференціації вкладається в семиелементну систему: 1) художній, 2) науковий, 3) офіційно-діловий, 4) публіцистичний, 5) релігійний, 6) розмовний та 7) епістолярний стилі. Крім того, виділяємо усний і писемний варіанти мови. Художній, публіцистичний, науковий та офіційно-діловий функціональні стилі в межах *НКУМ* репрезентовані головно писемними текстами, а розмовний – усними.

Для багатьох галузей мовної практики писемна форма національної мови є пріоритетною, вона зазвичай реалізує себе в художньому, публіцистичному, науковому й офіційно-діловому стилях. Тому для *НКУМ* приймаємо, що фактичним матеріалом

писемного варіанта української мови будуть тексти таких функціональних стилів сучасної української мови:

(1) художнього;
(2) публіцистичного;
(3) наукового;
(4) офіційно-ділового;
(5) релігійного;
(6) епістолярного.

Художній стиль у *НКУМ* репрезентуватимуть прозові, поетичні й драматургічні тексти відповідних хронологічних меж (від останніх років XVIII ст. і до сьогодні) з такою жанровою диференціацією:

– проза: роман, повість, оповідання, новела;
– поезія: вірш, сонет, поема, балада, сатира;
– драматургія: драма, комедія, трагедія.

Публіцистичний стиль у *НКУМ* головно репрезентуватиме періодика і частково рекламні тексти. Тематично-функціональний спектр періодичних видань надзвичайно широкий і покриває всю стильову диференціацію мови, тому періодичні видання, незалежно від їхнього типу – газета, журнал, бюлетень, тижневик, місячник, квартальник тощо, – вкладемо в прийняту нами шестиелементну стильову систему і відповідно структуруватимемо фактичний матеріал. Крім текстів періодики, до *НКУМ* буде уведено фрагменти публіцистичних творів Івана Франка, Лесі Українки, Бориса Грінченка, Ліни Костенко, Юрія Мушкетика та ін.

Науковий стиль у *НКУМ* поділено на підстилі: власне науковий, науково-популярний і науково-методичний та здиференційовано за галузями наук – гуманітарні, природничі й технічні.

Офіційно-діловий стиль у *НКУМ* доцільно представляти як усним текстовим матеріалом, так і писемним, що засвідчуватиме його існування в усному й писемному варіантах української мови. Офіційно-діловий стиль репрезентуватимуть:

– законодавчо-правові державні документи, наприклад, Акт проголошення незалежності України, Конституція України, Постанови Верховної Ради, Укази Президента;
– організаційно-службові документи: протокол, наказ, розпорядження, ухвала, заява;
– суспільна документація, наприклад, колективний договір, статут державних або недержавних організацій тощо.

Уведення до *НКУМ* текстів релігійного стилю, сформованого на книжній і народнорозмовній основі, умотивоване, по-перше, розширенням сфери його побутування в сучасному суспільстві, по-друге, його статусом в історії розвитку української мови, і, по-третє, потребою доповнення лексичного складу мови цим пластом лексики. Джерелами фактичного матеріалу релігійного стилю в *НКУМ* будуть Біблія, богослужбові книги, богословська й духовна література.

Епістолярний стиль, який сьогодні існує в паперовій і електронній формі, відповідно, зумовлює виділення в *НКУМ* класичних епістолярних текстів, чи листів, і електронної кореспонденції. Крім того, класичні епістолярні тексти додатково поділено на: а) офіційно-ділове внутрішньодержавне листування, б) офіційне міждержавне листування і в) приватне листування.

Усну частину в *НКУМ* репрезентуватимуть лише усні тексти розмовного стилю, а саме: а) публічні виступи, тексти парламентських дебатів; б) інтерв'ю й коментарі; в) ділове мовлення; г) повідомлення в транспорті; ґ) аудіореклама; д) побутове діалогічне й монологічне спілкування; е) діалектне мовлення; д) соціолектне мовлення.

Отже, до *НКУМ* увійдуть тексти:

I. Писемного варіанта української мови.
   1. Художнього стилю:
   2. Наукового стилю (власне наукового, науково-популярного та науково-методичного підстилів):
   3. Офіційно-ділового стилю.
   4. Публіцистичного стилю.
   5. Релігійного стилю.
   6. Епістолярного стилю.

II. Усного варіанта української мови.

1. Розмовного стилю.

Щодо стратегій обсягу, то на початковому етапі *НКУМ* доцільно будувати як середній, визначивши нижню межу в **1 млн слів,** передбачивши можливості перманентного поповнення корпусу текстовими даними до **100 млн слів** і не ставлячи обмежень на верхню межу кількості слів у корпусі.

Концептуалізація процедури створення корпусу обов'язково ставить вимогу на окреслення кола дослідницьких завдань, здійснюваних на копусі. І на початковому етапі експлуатації *НКУМ* дослідницькі завдання доцільно згрупувати за: а) мовними рівнями й б) прикладними аспектами.

Щодо першої групи *НКУМ* передусім доцільно використати для морфологічних, синтаксичних та лексичних досліджень. Так, морфологічні дослідження можуть стосуватися, зокрема, вивчення специфіки реалізації морфологічних категорій і значень, типології флективних змін у парадигматичних формах іменника та прикметника, реальної іменної та дієслівної системності, прийменникової та безприйменникової сполучуваності слів у мовленні тощо. Синтаксичні – вивчення типології простих і складних речень сучасної української мови, довжини речень (мінімальної, максимальної, типової), функціональної специфіки активних і пасивних конструкцій, засад упорядкування слів у реченнях сучасної української мови тощо. Лексикологічні – вивчення динаміки лексичного складу української мови, аспектів неологізації vs архаїзації українського лексикону, механізмів поповнення лексичного ядра та периферії на сучасному етапі функціонування мови, шляхів збагачення лексичного складу мови. Крім наведених вище, не варто забувати також про можливість корпусного дослідження фонетики й словотвору.

Прикладний аспект є чи не найважливішим застосуванням корпусів, оскільки опосередковано викликав і появу самих корпусів, і формування корпусної лінгвістики. Тут головно йдеться про лексикографію, а ще й, крім того, корпусну лексикографію та методику мови. У лексикографічних дослідженнях *НКУМ* можемо використовувати для укладання, наприклад, нового тлумачного словника української мови, шкільного тлумачного словника сучасної української мови, словника динаміки української лексики в XX ст., словника неологізмів, словника архаїзмів, словника дієслівних керувань, не кажучи вже про різні словники так званого реєстрового типу й частотні словники.

Щодо методики української мови як рідної та іноземної, то *НКУМ*, з одного боку, може слугувати базою для написання підручників і посібників, бути частиною інтерактивних методичних ресурсів, а з іншого боку – корпусні дані та можливість швидкого різнорівневого пошуку дадуть змогу підібрати дидактичний матеріал для практичного вивчення мови; а „здатність викликати комбінації слів, а не індивідуальні слова́" [3] уможливлює працю з реальними контекстами слів, за допомогою яких простіше зрозуміти правила вживання різних лексичних одиниць і, відповідно, опановувати мову.

Створення будь-якого текстового корпусу, як ще у 60-х роках минулого століття зазначав У. Френсис, вимагає детального плану чи, висловлюючись лексиконом початку XXI ст. – концепції, а якщо „планування зроблено правильно, то методика відбору та формування корпусу стає абсолютно зрозумілою, а практичне його створення виявляється цілком формальною процедурою" [2: 348].

Література

1. Ставицька Л. Арго, жаргон, сленг. – К.: Критика, 2005. – 462 с.
2. Френсис У. Н. Проблемы формирования и машинного представления большого корпуса текстов // Новое в лингвистике. – 1983. – Вып. XIV. – С. 334–352.
3. McEnery T., Wilson A. Corpus Linguistics. – 1996. – http://www.comp.lancs.ac.uk.
4. Przepiórkowski A. Korpus IPI PAN: wersja wstępna. – Warszawa: Instytut podstaw informatyki PAN, 2004. – 89 st.

О. Кукушкина, А. Поликарпов, д. филол. н., ***Г. Токтонов****
МГУ им. М.В.Ломоносова (Россия, Москва)
УДК 81'33

КОРПУСНАЯ НЕОГРАФИЯ И НЕОЛОГИЯ: СИСТЕМНЫЙ АНАЛИЗ ХАРАКТЕРИСТИК ЛЕКСИЧЕСКИХ НЕОЛОГИЗМОВ (НА МАТЕРИАЛЕ ГАЗЕТНОГО РАЗДЕЛА «ПОЛИСТИЛЕВОГО КОРПУСА СОВРЕМЕННОГО РУССКОГО ЯЗЫКА»)

В статье обоснованы цель, объект, предмет, теоретическая база анализа системных характеристик новой русской лексики газетного языка конца XX века, а также указаны основные результаты такого исследования.

1. Целью настоящей работы является изучение системных характеристик новой лексики русского газетного языка 90-х годов XX века. Мы полагаем, что исследование газетного языка позволяет вскрыть важные особенности данного языка в целом. Более того, в этих особенностях русского языка должны проявляться и универсальные неологические характеристики новых слов.

Под «системными характеристиками» подразумеваются наиболее важные, принципиальные особенности новых лексических единиц, находящиеся во взаимообусловливающей связи. Эти особенности вскрываются на основе теоретических представлений о закономерностях процессов семантического развития слов (изменения у них значений и образования новых значений) и закономерностей словообразовательного процесса (образования новых слов) [9; 7; 4; 8; 10]. Верификация этих представлений производится на основе анализа экспериментальных данных, получаемых в ходе эмпирического сопоставления характеристик новой лексики с соответствующими характеристиками их производящих единиц и единиц общей лексической системы, представленной в репрезентативных общих словарях и словаря текстовых корпусов русского языка.

Несмотря на то, что современная русистика располагает рядом фундаментальных работ, описывающих неологические аспекты лексической и словообразовательной систем (достаточно назвать имена Е.А.Земской, И.С.Улуханова, Н.З.Котеловой, Г.Н. Скляревской, В.В.Лопатина, Н.А.Янко-Триницкой, А.А.Брагиной и др.), представляется, что системный аспект неологического процесса в лексике исследован еще недостаточно. Системность нашего подхода заключается не только в том, что мы опираемся на системную теоретическую концепцию развития лексики (см. п. 4 ниже и ссылки по данной проблематике), но и в том, что проводится сплошной, без изъятия, анализ представительного современного текстового материала, которым является ядерная часть «Компьютерного корпуса газетных текстов русского языка конца XX века» (являющегося, в свою очередь, частью «Полистилевого корпуса современного русского языка». «Ядерный корпус» (общим объемом ок. 1 млн. 350 тыс. словоупотреблений, более 3 тыс. текстов) был создан в лаборатории общей и компьютерной лексикологии и лексикографии филологического факультета МГУ при финансовой поддержке Российского фонда фундаментальных исследований в 1999–2002-ом гг. (гранты № 01-07-90386 и 02-06-80435) [1; 3; 2;] и потом вошел в качестве составной части в более обширный «Полистилевой корпус современного русского языка», созданный там же по гранту РФФИ № 03-07-90265 в 2003-2005 гг. [2]. (Описание «Компьютерного корпуса газетных текстов русского языка конца XX века» и демонстрационный вариант корпуса для Интернета см. на сайте Лаборатории общей и компьютерной лексикологии и лексикографии [http://www.philol.msu.ru/~lex/corpus/]. Описание «Полистилевого корпуса современного русского языка» см. на том же сайте [http://www.philol.msu.ru/~humlang/articles/polystylcorp.html]).

Анализ такого представительного и контролируемого материала, каковым является материал размеченного корпуса, усиливает возможности объективного моделирования общеязыковых соотношений единиц и категорий языковой системы, в т.ч. и количественными методами. Этим материал корпуса выгодно отличается от исследовательских картотек и существующих словарей новых слов, для которых

* *© О. Кукушкина, А. Поликарпов, Г. Токтонов, 2006*

источники, объем и полнота материала, как правило, не регламентированы достаточно определенно.

2. Объектом настоящего исследования является список 1887 новых слов русского газетного языка 90-х годов XX века, отобранный в ходе автоматической фильтрации и многостороннего экспертного анализа лексического материала представительного корпуса газетных текстов. А именно, материалом для исследования послужили новые слова, появившиеся в русском языке в период 1990-1997 гг., отобранные из указанного корпуса текстов.

3. Предметом исследования являются некоторые важные словообразовательные, морфологические, семантические, жанровые и частотные свойства новых слов в их взаимосвязи, а также по сравнению с аналогичными характеристиками прочих слов русского языка и, в частности, тех слов, которые послужили производящими для обнаруженных в корпусе газетных текстов неологизмов.

4. Теоретические основания анализа. Работа по выявлению закономерностей появления новых слов опирается на теоретические положения, сформулированные в Модели жизненного цикла знака [4; 7; 9; 8; 10]. Эта теоретическая концепция позволяет выдвинуть и экспериментально проверить ряд прогнозов о процессах исторического движения в знаковом материале языка, в т.ч. и об особенностях неологического процесса в лексике различных функциональных сфер (стилей и жанров) современного русского языка. Согласно этой концепции, в частности, прогнозируется, что слова с возрастом не только постепенно теряют семантическую специфичность, предметность своих значений, но и в качестве производящих порождают слова, в среднем, все более абстрактные, менее предметные (лексически и категориально-грамматически), чем они могли порождать в начале своей жизни. Кроме того, вне зависимости от возраста производящих, новые (производные) слова должны быть по своей категориальной семантике более абстрактными, чем их производящие.

Более специфичное лексическое значение производящего имеет меньше шансов сохраниться в неизменном виде (без серьезных изменений своих компонентов) в производном. Таким образом, для слов относительно более молодых, слов, находящихся на относительно более низких ступенях деривации, слов, принадлежащих к категориально более конкретной части речи (т.е. существительных) должны быть в большей степени свойственны нерегулярности в соотношении их лексических значений с лексическими значениями их производных, что характерно для мутационной разновидности лексической деривации. И, наоборот, для слов, находящихся на относительно более высоких ступенях деривации, и слов, принадлежащих к признаковым частям речи, должны быть в большей степени свойственны модификация и, особенно, синтаксическая деривация.

Новые слова, называющие нечто бывшее известным носителям языка и ранее обозначавшееся другим словом или целым словосочетанием, должны противопоставляться по ряду признаков словам, обозначающим совершенно новые смыслы. Можно предполагать, первые будут чаще мотивироваться уже существующими словами (например, из прежней фразеологической номинации) и поэтому будут находиться на более отдаленных ступенях деривации, чем «совсем» новые, будут менее предметными (и по лексической, и по категориальной семантике) и т.п.

5. Основные полученные результаты.

а) Классификация актуальных способов словообразования, словообразовательных формантов и средств. Она разработана на основе анализа неологического лексического материала Газетного корпуса, а также расширения существующих классификаций И.С. Улуханова и Е.А. Земской (см. Таблицу 1 в Приложении). На ее основе удается выяснить, что среди новых слов корпуса газетных текстов, полученных различными словообразовательными способами, наиболее многочисленными являются суффиксальные слова и чистые сложения. Количество слов, образованных по этим структурным моделям, будет примерно одинаковым, если объединить чистое сложение с близкими ему случаями, когда в качестве одного из двух компонентов-основ выступает связанная или усеченная основа (т.н. сложносокращенный способ, или сложение со связанным компонентом).

Наиболее употребительными среди всех словообразовательных классов слов являются каждая из новых аббревиатур (в среднем, почти 12 употреблений каждого такого слова, что почти в 6 раз выше, чем средняя употребительность любого другого типа лексических неологизмов). Высокая употребительность этого рода образований

объясняется, видимо, высокой актуальностью смысла, для обозначения которого адаптируются (сокращаются) уже связанные с ним знаковые ресурсы (полные слова и развернутые словосочетательные номинации). Кроме них, высокой, в среднем, частотностью на общем фоне выделяются заимствования (в среднем, 4,2 употреблений). При относительно небольшом количестве новых чистых заимствований и заимствований, осложняемых аффиксацией, каждое из таких слов, в среднем, употребляется существенно чаще, чем новые слова многих других классов. Можно полагать, что заимствования осуществляются, в основном, в тех случаях, когда обозначаемая область уже очень актуальна, и средство обозначения берется готовым напрямую из материала языка-источника.

б) ***Наиболее активными с точки зрения образования новых слов являются такие производящие слова, которые находятся на относительно низких ступенях деривации*** (более близкие к исходному слову словообразовательного гнезда). Производящие слова новых слов по данным Корпуса чаще находятся между нулевой и первой ступенями деривации (среднее арифметическое значение ступени деривации для производящих составляет примерно 0,84). В то же время, средняя ступень деривации, например, всех слов «Словообразовательного словаря» А.Н.Тихонова равняется 1,77 (т.е. среднее слово данного словаря находится ближе ко второй ступени). Т.е. новые слова, зафиксированные в Газетном корпусе, образованы от слов, находящихся в деривационной цепи, в среднем, почти на одну ступень ниже, чем это свойственно среднему слову русского языка. Кроме этого, распределение слов по ступеням деривации наглядно демонстрирует доминирующую активность участия в словообразовании непроизводных слов (нулевая ступень деривации). Увеличение величины ступени деривации слов пропорционально ведет к снижению их словообразовательной активности. В среднем, каждое слово нулевой ступени деривации участвует в производстве более полутора новых слов, первой ступени – 1,4, второй – 1,38, третьей 1,17, четвертой – 1,05, пятой и шестой – 1,00. Начиная с седьмой ступени словообразовательной активности слов по материалам данного корпуса не наблюдается.

в) Активнее участвуют в словопроизводстве относительно более употребительные в речи слова. **Средняя частота употребления в газетном корпусе слов, мотивирующих обнаруженные новые слова (85,33 употребления), примерно в четыре раза выше средней употребительности всех слов корпуса (21,77).**

г) Соотношение категорий новых сложных и новых несложных слов. Кроме аффиксальных производных, к числу несложных новых слов мы относим слова, образованные обратными (десуффиксация, депрефиксация), а также некоторыми другими окказиональными способами словообразования (неморфемное усечение, искажение). Полученные нами данные демонстрируют, что расслоение новой лексики на сложные и несложные производные выявляет принципиальные категориальные различия этих групп новых слов как между собой, так и по сравнению со всеми словами Корпуса. Среди новых сложных слов значительно больше существительных (приблизительно на 12%), несколько больше прилагательных (приблизительно на 2%) и намного меньше глаголов и наречий (в составе новых сложных слов эти части речи имеют всего лишь по одному представителю), чем в составе новых несложных слов.

Из наших данных также следует, что если степень предметности новых несложных слов на 10% ниже, чем у всех слов корпуса, то у новых сложных – наоборот, более чем на 2% выше. Т.е. подтверждается наше общее представление о разном механизме образования новых аффиксальных (они составляют большинство среди несложных) и новых сложных производных. Образование сложений близко образованию словосочетаний по результату – получению более специфической в семантическом плане номинации, чем ее составляющие. При образовании новых аффиксальных слов получаются, как правило, семантически и категориально менее специфические слова.

д) При образовании новых аффиксальных и иных несложных слов качественно меняется соотношение слов предметной (существительные) и признаковой (прилагательные, глаголы и наречия) категориальной семантики среди слов нового поколения по сравнению с аналогичным соотношением частеречных категорий производящих слов. Происходит количественный категориальный сдвиг в сторону слов более абстрактной, признаковой семантики, увеличивающих свое присутствие среди «молодых» слов. Иначе говоря, среди новых не-сложных слов по сравнению с их производящими отмечается сокращение доли существительных и увеличение доли слов признаковых категорий.

е) Зависимость между ступенью деривации, на которой находится производящее, и структурно-семантическим типом деривации при образовании нового слова. Наши данные показывают, что чем ближе находится производящее к исходному слову в своем словообразовательном гнезде, т.е. чем ниже его ступень деривации, тем более вероятна при образовании нового слова мутационная разновидность лексической деривации. И наоборот, чем выше ступень деривации производящего (чем оно дальше от исходного слова), тем более

частотны случаи синтаксической деривации (транспозиции) и модификации.

ж) Зависимость между частеречной принадлежностью производящего и типом деривации при образовании нового несложного слова. Наши данные показывают, что имена существительные, производящие новые слова газетного корпуса, в большей степени склонны к мутации, прилагательные – к синтаксической деривации, а глаголы – к модификации.

з) Лексическая семантика новых слов: новизна обозначаемых понятий. Новые слова корпуса различаются не только способом их образования или количеством мотивирующих их слов, но еще и новизной обозначаемых ими внеязыковых реалий. По этому критерию можно различать новые слова с новыми и новые слова с неновыми означаемыми (так, *банкомат, видеопират, фитнес-центр* – это новые слова с новыми означаемыми, а новые слова с неновыми означаемыми – это, например, *креативный* «творческий», *суперхитовый* «очень популярный», *наркозелье* «наркотик»). Новые слова с новыми означаемыми рассматриваются нами как «эталонные» представители нового поколения лексических единиц русского языка, поскольку они новы и по форме, и по содержанию. В процентном соотношении эта разновидность новых слов составляет чуть больше четверти от их общего количества в корпусе. Они обладают следующими свойствами, противопоставляющими их словам, называющим нечто и ранее бывшее известным носителям языка и обозначавшееся другим словом или целым словосочетанием:

– категориальная семантика новых слов с новыми означаемыми более предметна, иначе говоря, в их составе более велика доля существительных, чем в составе слов с неновыми означаемыми (примерно на 8%);

– новые слова с новыми означаемыми, в среднем, находятся на более низких ступенях деривации, чем новые слова с неновыми означаемыми (средняя ступень деривации первых – 1,3, вторых – 1,8);

– среди новых существительных с новыми означаемыми доля слов мужского рода больше, чем среди новых существительных с неновыми означаемыми (примерно на 16%), а доли слов женского и среднего рода, соответственно, меньше;

– среди новых прилагательных с новыми означаемыми доля относительных по отношению к качественным больше, чем среди новых прилагательных с неновыми означаемыми (почти на 25%);

– новые слова с новыми означаемыми характеризуются большим количеством заимствований в своем составе, тогда как среди новых слов с неновыми означаемыми выделяются префиксальные, префиксально-суффиксальные и сложно-аффиксальные производные;

- новые слова с новыми означаемыми относительно более употребительны в текстах газетного корпуса, чем новые слова с неновыми означаемыми (средняя употребительность в газетном корпусе первых – 2,83 употребления каждого слова, вторых – 1,79).

Литература

1. Виноградова В.Б., Поликарпов А.А., Пирятинская Е.Ф, Савчук С.О. Компьютерный корпус текстов современной русской газеты // Язык СМИ как объект междисциплинарных исследований. – М.: Изд-во МГУ, 2003.
2. Кукушкина О.В., Пирятинская Е.Ф., Поликарпов А.А. «Полистилевой корпус текстов современного русского языка»: задачи, принципы, первичный анализ (2006) http://www.philol.msu.ru/~humlang/articles/polystylcorp.html.
3. Кукушкина О.В., Поликарпов А.А. Жанровое распределение русских предлогов в корпусе современной газеты // Язык и социум. Материалы V международной научной конференции. Ч.1. – Минск, 2003.
4. Поликарпов А.А. Закономерности образования новых слов: моделирование процесса и его экспериментальное исследование // Язык. Глагол. Предложение. К 70-летию Георгия Георгиевича Сильницкого. – Смоленск, 2000, С. 213–214.

5. Поликарпов А.А. Когнитивное моделирование циклических процессов в становлении лексической системы языка // Труды Казанской школы по компьютерной и когнитивной лингвистике. – Казань, 2001. С. 57-99.
6. Поликарпов А.А., Филимонова Т.А. О системной зависимости негативных денотативных и субъективно-оценочных признаков фразеологических единиц от их возраста // Cognitive Modeling in Linguistics. CLM-2005. – Москва, 2005.
7. Поликарпов А.А. Циклические процессы в становлении лексической системы языка: моделирование и эксперимент. Автореферат докторской диссертации. – М., 1998.
8. Polikarpov A. Cognitive Mechanisms for Emergence of Some Regularities of Human Language Evolution // VII Международная конференция «Когнитивное моделирование в лингвистике» (September 4-11, 2004 Varna). Труды (Обработка текста и когнитивные технологии №9). – М., МИСИС, 2004. – С. 10-20.
9. Polikarpov A. On the Model of Word Life Cycle // Koehler, R., Rieger, B. (eds.) Contributions to Quantitative Linguistics. – Dordrecht: Kluwer, 1993. – P. 53-66.
10. Polikarpov A.A. Towards the Foundations of Menzerath's Law (On the Functional Dependence of Affix Length on their Positional Number within Words) // Peter Grzybek (ed.). Contributions to the Science of Text and Language. Word Length Studies and Related Issues. – Dordrecht: Springer, 2006. – Pp. 215-240.

Приложение

Таблица 1. Способы словообразования, их продуктивность и употребительность по данным Ядерного корпуса газетных текстов русского языка

Способ образования нового слова	Кол-во слов, образованных данным способом	Кол-во их употреблений	Средняя их употребительность	Доля этих слов только среди новых произв. Исключая заимствования и слова с новыми корнями	Доля употреблений этих слов, среди всех употреблений новых произв-х слов. Исключая заимствования и слова с новыми корнями.	Доля этих слов среди всех новых слов	Доля употребл-й этих слов среди всех употребл-й новых слов
Собственно способы словообразования							
суффикс.	535	918	1,72	31,45%	28,62%	28,35%	23,99%
чистое слож-е	376	483	1,28	22,10%	15,06%	19,93%	12,62%
префик.	260	414	1,59	15,29%	12,91%	13,78%	10,82%
слож. с перв. связ. комп.	252	624	2,48	14,81%	19,45%	13,35%	16,31%
преф.+суф.	101	131	1,3	5,94%	4,08%	5,35%	3,42%
слож.+суф.	50	74	1,48	2,94%	2,31%	2,65%	1,93%
аббревиац.	37	445	11,74	2,23%	13,90%	2,01%	11,65%
сращение	20	23	1,15	1,18%	0,72%	1,06%	0,60%
контамин.	19	25	1,32	1,12%	0,78%	1,01%	0,65%
сл.со втор.связ.комп.	15	21	1,4	0,88%	0,65%	0,79%	0,55%
искажение	9	11	1,22	0,53%	0,34%	0,48%	0,29%
неморф. усечение	5	10	2	0,29%	0,31%	0,26%	0,26%
десуффик.	4	4	1	0,24%	0,12%	0,21%	0,10%
междуслов наложение	4	5	1,25	0,24%	0,16%	0,21%	0,13%
постфикс.	3	4	1,33	0,18%	0,12%	0,16%	0,10%
суффик.+ постфик.	2	4	2	0,12%	0,12%	0,11%	0,10%
депрефик.	2	2	1	0,12%	0,06%	0,11%	0,05%

префик.+ постфик.	2	3	1,5	0,12%	0,09%	0,11%	0,08%
декомп.+ суффик.	1	3	3	0,06%	0,09%	0,05%	0,08%
префик.+ суффик.+ постфик.	1	1	1	0,06%	0,03%	0,05%	0,03%
усеч.+ редупл.	1	1	1	0,06%	0,03%	0,05%	0,03%
усеч.+ сложение	1	1	1	0,06%	0,03%	0,05%	0,03%
Итог	1700	3207	1,89	100,00%	100,00%	90,14%	83,83%
Заимствования и немотивированные новые слова							
заимств.	173	590	3,42			9,11%	15,36%
заимств.+ суффик.	7	23	3,29			0,37%	0,60%
заимст.+ слож.со втор. связ. компонен.	3	4	1,33			0,16%	0,10%
заимств.+ сложение	2	2	1			0,11%	0,05%
созд. нов. корня+ префик.+ суффик.	1	1	1			0,05%	0,03%
созд.нов. корня+ суффик.	1	1	1			0,05%	0,03%
Итог	187	621	3,33			9,86%	16,17%
Общ. итог	1887	3828	2,03			100,00%	100,00%

Л. Коцюк*

УДК 81’373.374.33

ВИКОРИСТАННЯ КОРПУСІВ ТЕКСТІВ У ПРОЦЕСІ УКЛАДАННЯ СЛОВНИКІВ

The article represents the difference that a corpus, and corresponding advances in lexicology and lexicography can make to form new generation of corpus-based dictionaries. The author raises the question of what the word is and how corpus study can influence the representation of all its features in the dictionary.

У будь-якій діяльності можна виділити формальну чи рутинну і творчу, неформальну компоненту. Це в повній мірі стосується і роботи лексикографа. Ще В.В. Виноградов цитував слова, які стали класикою, – нема нічого тяжчого, ніж укладати словник і писати граматику. Не завжди можна чітко провести грань між рутинною та творчою роботою при укладанні словників. Однак можна стверджувати, що методологія корпусної лінгвістики не тільки змінила це співвідношення, але й внесла багато нового в саму лексикографію. В значній мірі потреби лексикографії стимулювали появу та розвиток корпусної лінгвістики.

Метою даної статті, однак, не є розібрати всі аспекти співвідношення між цими двома лінгвістичними дисциплінами, а звернути увагу на ті переваги, які отримує лексикограф при укладанні словника, використовуючи дані з корпусів текстів. Основними **завданнями** дослідження є встановлення тих характеристик лексичної одиниці, які можуть бути представлені з допомогою корпусу текстів та аналіз готових лексикографічних робіт на предмет використання даних з корпусів текстів.

* *© Л. Коцюк, 2006*

Укладання сучасного словника, як правило, базується на деякому корпусі текстів. Це, до певної міри, одне з найбільш вдалих та впливових застосувань корпусів текстів, оскільки навіть ті люди, які ніколи не чули про корпус користуються результатами його аналізу. Дослідження щодо використання корпусів текстів у процесі укладання словників не численні, проте набувають все більшої актуальності. Цієї теми у своїх наукових доробках торкаються Сінклер (1987), Аткінс (1992), Саммерс (1996), Бо та співавтори (1996), Клер та співавтори (1996). Проте всі вони підкреслюють важливість використання результатів аналізу корпусу текстів у лексикографії.

Будь-який лінгвіст (чи студент) потребує відповіді принаймні на п'ять запитань, коли йдеться про невідоме слово: 1. Що це слово означає? 2. З чим воно асоціюється? 3. З якими значеннями воно асоціюється? 4. З якими граматичними функціями воно асоціюється? 5. Як це слово розміщується у тексті?

Перші два запитання здаються найбільш типовими. Всі відомі словники мають справу зі значеннями слів. З огляду на друге запитання, то ще з початку заснування науки про укладання словників і до нашого часу вважається, що правильний опис будь-якого слова буде включати інформацію про лексичне оточення цього слова, його колокацію [5]. Наприклад, слово *consequence* зустрічається з *serious*, а *consequences* – з *economic*. Напевне не існує жодного слова в англійській мові (за винятком деяких граматичних слів, і навіть цей виняток до певної міри суперечливий), яке б не мало певних правил, що стосується вживання інших слів навколо нього. Словосполучення об'єднуються у ідіоми, і іноді важко визначити межу між словами, які регулярно використовуються с якимсь іншим словом, та словами, які утворюють ідіоматичний вираз (хоча у контексті ці дві групи досить легко ідентифікуються через те значення, яке вони мають як єдине ціле чи кожного складового слова по-окремості).

Не менш важливим за сполучуваність, однак менш описаним, є термін "семантична асоціація", феномен, що стосується третього питання. Його можна визначити як тенденцію слова до вживання з певним семантичним класом чи групою слів. Так, наприклад, розглянемо, як працює семантична асоціація у випадку вживання слова *consequence*. *Consequence* має семантичні асоціації з концептами: логіки (зовсім не дивно, оскільки ядерне значення цього слова стосується логічних зв'язків): *unavoidable, inevitable, inexorable, inescapable, ineluctable, direct, ultimate, long-term, immediate*; сподівання/несподіванки: *likely, possible, probable, natural, unintended, odd, strange, planned-for;* чи, з негативною оцінкою: *awful, dire, appalling, sad* з зазначенням ступеня (не)важливості: *serious, important, dramatic, enduring, prominent.*

Як і у випадку з лексичною сполучуваністю, на поведінку слова у значній мірі впливає його вживання у однині/множині. Наприклад, слово *consequences*, зазвичай семантично асоціюється і відповідно сполучається з прикметниками на позначення логіки та негативної оцінки (хоча воно частіше асоціюється з останнім), та ще й додатково з таким концептом, як володіння: *biological, constitutional, economic, political, cultural*.

Як бачимо, лексична сполучуваність та семантична асоціація безпосередньо впливають на значення. Інші характеристики слова, однак також, можуть, хоча й менше, бути важливими для його вживання. Четверте запитання у вищезгаданому списку перевіряє, які граматичні асоціації, чи зв'язки, може мати слово. Це запитання видається важливим з огляду на те, що іноді зв'язки слова характеризуються певними граматичними рисами. Звернімося до вже аналізованого нами слова *consequence*. Навряд чи воно може вживатися як додаток речення на відміну від інших абстрактних іменників, таких як *preference* чи *use*. Важко зустріти приклад (можливо й дивно) такого речення: *Unfortunately it also had this tragic consequence that the baby became grossly bloated.* У той час як такі речення, як *The homeless are asked if they have a preference.* та *The minister called on schools to make more use of the colleges' vocational experience ….* досить часто вживаються. З іншого боку, *consequence* вживається як (частина) комплементу набагато частіше за інші абстрактні іменники: *It is the natural consequence of a deep recession ….*

Не важко помітити, що слово *consequence* дуже часто вживається у прийменниковій групі: *as a consequence* чи *in consequence* (чи їх варіантах). У свою чергу, ці фрази зазвичай з'являються на початку речення, на відміну від звичного розташування прийменникових груп. Ця вибірковість щодо функціонування на початку речення слугує прикладом відповіді на п'яте запитання – які типові розташування слова у тексті?

Все вищесказане свідчить про те, що "знати слово" означає набагато більше, ніж "знати значення слова". Тому лексикограф, щоб відповісти принаймні на цих п'ять запитань, змушений звернутися до аналізу корпусу текстів. З розвитком комп'ютерних

технологій цей процес набув масштабного характеру. Більшість сучасних лексикографічних праць своєю метою поставили розширити опис слова з врахуванням того, з якою частотою і в якому оточенні слово у всіх його значеннях функціонує у текстах різних стилів. Традиційні словники поряд з інформацією про значення слова показують статус іменника (злічувальний / незлічувальний), його вимову, а в деяких випадках зазначають, що це слово у певному контексті буде вживатися в офіційному стилі. Також вони можуть інформувати читача про використання даного слова у сталих виразах, таких як *in consequence* чи *as a consequence.*

Використання корпусів текстів при укладанні словника значно розширили ці можливості. Ось перелік основних акцентів, з допомогою яких корпуси текстів модифікують формування словників:

- акцент на частоту;
- акцент на сполучуваність та фразеологію;
- акцент на варіативність
- акцент на граматичні форми лексем
- акцент на автентичність [6: 96-109].

Частота вживання слів у різних значеннях та їх розподіл у тексті є, напевне єдиною сферою, на яку не поширюється людська інтуїція щодо вживання мовних форм. Саме тому важливим вкладом корпусів текстів у створення словника є включення інформації про відносні частоти певних одиниць. Таку інформацію може використати як безпосередній читач словника, так і його укладач, вирішуючи, яке зі значень певної лексичної одиниці повинно подаватися у словниковій статті першим. Наприклад, порівняємо кількість значень одного й того ж частотного слова *know*, представлену у другому виданні Longman Dictionary of Contemporary English (1987), укладеного без врахування аналізу корпусу текстів, та 3-му виданні цього ж словника 1995 року, сформованого з допомогою даних, взятих з корпусу. У виданні 1987 року це слово має 20 значень, у пізнішому виданні цих значень уже нараховують 40. Збільшення кількості значень пояснюється появою нової інформації щодо частоти вживання, включаючи поділ значення на два і більше. Також можна зафіксувати появу нових значень чи вживань, які здавалися не важливими до того, як корпус показав, наскільки вони є частотними. Одним з прикладів може слугувати фраза *I know*, включена в першому з аналізованих словників тільки як фраза на позначення раптової ідеї: *What can we get her for birthday? Oh, I know, we'll get her some flowers.* У новому виданні словника додаються ще два випадки використання цієї фрази: погодження (як в *'I'm so worn out! Yeah, I know'*) та попереднього непогодження (як в *It sounds silly I know, but try it anyway*).

Іншою інновацією в словниках, яка стала можливою тільки завдяки використанню корпусу текстів, є включення безпосередньої інформації про частоту. Словник Longman Dictionary of Contemporary English 1995 року видання, наприклад, відмічає слова, які особливо частотні в усному чи писемному мовленні і в деяких випадках порівнює ці дві сфери вживання.

Використовуючи корпус текстів з зазначенням частин мови, стає можливим порівняти частоту вживання одного й того ж слова, наприклад, будучи іменником чи дієсловом. Дефініції словника тоді можна розташувати у правильному порядку, зважаючи, що першими будуть показані частотніші частини мови [9: 262]. Наприклад, *gorge* як іменник (у значенні 's valley with steep sides') приблизно у чотири рази частіше вживається, ніж дієслово (у значенні 'eat greedily'). Таким чином, у словнику більш резонно буде представити значення іменника перед значеннями дієслова.

Ще більше допомогти лексикографам зможе автоматичне розпізнавання значень слова [9; 4]. Це досягнення дасть можливість ідентифікувати окремі значення слова з урахуванням частоти вживання даного слова у цьому значенні. Зважаючи на те, що слова в окремих значеннях вживаються у різних оточеннях, автоматичне розпізнавання значень теоретично можливе, але все ще не досяжне.

В "нових" словниках особливу увагу приділяють сполучуваності та фразеології, що відбиває загальну тенденцію корпусів текстів акцентувати увагу саме на цих аспектах дослідження лексичних одиниць [7: 116-136; 8: 104-115, 150-159; 9]. З цієї точки зору, словники, засновані на аналізі корпусу текстів характеризуються наступними особливими рисами. По-перше, спостерігається тенденція до подачі визначення цілій фразі, а не окремому слову (наприклад, перевага буде надаватися вислову *I know,* ніж просто *know*). По-друге, деякі словники використовують фразу вже у дефініції для ілюстрації сполучуваності. Наприклад, визначення слова *brink* у словнику Collins Cobuild English

Dictionary 1995 року видання – 'if you are *on the brink of* something, usually something important, terrible, or exciting, you are just about to experience it'. Особливістю подачі визначення таким чином є ще й доповнення інформації про лексичну сполучуваність чи семантичну асоціацію цього слова (див., наприклад, 'usually something important, terrible, or exciting'). Ще однією важливою рисою цих словників є обов'язкове представлення прикладу для підтвердження визначення та демонстрації сполучуваності.

Фразеологія здається особливо важливою у випадках частотних слів, більшість з вживань яких можуть бути представлені у сталих виразах [7; 9]. Саммерс, наприклад, відмічає, що слово *day* найчастіше зустрічається у таких фразах, як *one day, the other day, some day* [9], а Сінклер зауважує, що у словниках недостатньо приділяється увага саме такій фразеології, особливо це стосується граматичних слів, на противагу лексичним (наприклад, слово *a* зустрічається у таких фразах, як *come to a head* як альтернатива артиклю *the*, проте словники часто ігнорують таке його вживання) [7].

Інший важливий акцент при укладанні словника, на який корпус текстів зробив можливість наголосити – це варіативність мовлення в залежності від регіону, статі, соціальної групи, часового періоду чи стилю. Особливістю словників, які беруть інформацію з корпусу текстів, є подача випадків варіативності щодо стилю мовлення. Згідно спостережень Байбера, наприклад, частотні дієслова англійської мови *say, get, go, know* та *think* набагато частіше вживаються у розмовному стилі, ніж у інших [3]. Разом вони складають 40% всіх лексичних одиниць в усному мовленні, у той час, як на письмі відсоток їх вживання становить тільки 20 % від усіх лексем.

Традиційно у лексикографії словосполучення розглядали тільки з лексичної точки зору. Різниця між лексемами та грамемами базується на принаймні трьох твердженнях: по-перше, граматичні слова частіше вживаються за прості лексеми; по-друге, граматичні слова утворюють парадигми, а лексичні значення слів – синтагми; по-третє, граматичні слова не мають власного значення на відміну від лексичних слів. Останнім часом кожне з цих тверджень можна піддати сумніву за допомогою доказів з корпусу текстів. Наприклад, у згаданому вже нами вислові *come to a head* статус артикля *a* як граматичного слова дуже сумнівний, оскільки його парадигматичні характеристики згладжуються тим фактом, що це слово не може бути замінене артиклем *the*, і відповідно *a* відіграє таку ж роль як і слово *head*, тобто його можна вважати таким, що має лексичне значення.

У процесі укладання словника за допомогою корпусу текстів, можна запропонувати приклади для ілюстрації найбільш типового вживання слова чи фрази, і, якщо приклади взяті безпосередньо з корпусу, автентичність гарантується. Важливо розрізняти поняття автентичності та типовості. Будь-який корпус текстів включає безліч автентичних прикладів вживання певного слова, вони є частиною реальних текстів, проте в той же час – продуктом інновацій, гри слів, або просто певних обставин, і таким чином вони не можуть бути типовими. На противагу, деякі укладачі словників вигадують речення, які показують типові випадки вживання, але вони не є автентичними.

Зважаючи на той факт, що всі лексикографи погоджуються, що типовість є важливим фактором, не всі з них наполягають на повній автентичності. Бо стверджує, що "більшість цитат не підходять для учнівських словників тому що вони надто складні граматично, включають важкі слова та ідіоми, чи звертаються до специфічних контекстів" [2]. У вступі до Longman Dictionary of Contemporary English 1995 року видання говориться, що всі приклади у цьому словнику взяті з усних та письмових матеріалів корпусу Longman Corpus Network. Деякі приклади беруться безпосередньо з корпусу, а інші дещо видозмінені з метою уникнути складних слів, а деякі були створені спеціально для певного лексичного входу. Укладачі Collins Cobuild English Dictionary 1995 року видання наголошують на виключно автентичності прикладів. Як бачимо думки лексикографів розходяться у цьому питанні і передбачаємо, що це питання буде підніматися у процесі розвитку комп'ютерної лексикографії.

Отже, словники, які використовують результати аналізу корпусу текстів відрізняються від своїх попередників. Дослідження автентичних текстів дає змогу зробити сучасний словник "живим", тобто таким, який відображає реальне мовлення, усне чи то письмове, офіційне чи неформальне. Такі словники будуть належне оцінені як іншими лексикографами, так і користувачами: студентами чи науковцями.

Література

1. Atkins B.T.C. Tools for Computer-Aided Corpus Lexicography: the Hector Project. // Acta Linguistica Hungarica 41 (1-4), 1992-93. – P. 5-71.

2. Baugh D., Harley A. and Jellis S. The Role of Corpora in Compiling the Cambridge International Dictionary of English. // International Journal of Corpus Linguistics 1, 1996. – P. 39 – 59.
3. Biber D., Johansson S., Leech G., Conrad S., Finegan E. Longman Grammar of Spoken and Written English. – London: Longman, 1999.
4. Clear J., Fox G., Francis G., Krishnamurthy R. and Moon R. Cobuild: The State of the Art. // International Journal of Corpus Linguistics 1, 1996. – P. 303 – 314.
5. Hoey M. What's in a word. // MED Magazine. – Issue 10, August 2003. - http://www.macmillandictionary.com/med-magazine.htm
6. Hunston S. Corpora in Applied Linguistics. – Cambridge University Press, 2002. – 241 p.
7. Sinclair J.M. Looking Up: An Account of the Cobuild project. – London: HarperCollins, 1987.
8. Sinclair J.M. A Way with Common Words. – In: Hasselgard H., Oksefjell S. Out of Corpora: Studies in Honour of Stig Johanson. – Amsterdam: Rodori, 1999. – P. 157 – 179.
9. Summers D. Computer Lexicography: The Importance of Representativeness in Relation to Frequency. – In: Thomas J. and Short M. Using Corpora for Language Research. – London: Longman, 1996. – P. 260-265.

Т. Сидорко*
Інститут філології КНУ ім. Т. Шевченка (Київ)
УДК 161.2.81’322

РЕКЛАМНИЙ ТЕКСТ ЯК ОСОБЛИВА ОДИНИЦЯ КОРПУСУ ДАНИХ

У статті йдеться про актуальність та принципи створення корпусу рекламних текстів, збалансованого за жанрами, моделями тощо. Укладання корпусів рекламних текстів з їх наступним лінгвістичним аналізом дозволяє створити надійну базу для проведення лінгвістичних та психолінгвістичних досліджень, використовуючи різні методи для отримання статистично надійних результатів.

Створення корпусу рекламних текстів як особливого виду сугестивних одиниць є дуже важливим. Рекламні тексти мають високий ступінь впливовості на свідомість споживача (у порівнянні зі звичайним текстом), у них використовують специфічні мовні й психолінгвістичні прийоми для досягнення кінцевої мети реклами (купівлі рекламованого продукту). Для опису рекламного тексту можуть бути використані різні схеми, що містять як загальнолінгвістичні, так і специфічні для рекламного тексту характеристики: фонетичні, морфологічні, синтаксичні.

Матеріалом для створення корпусу є рекламні тексти з популярних газет і журналів. *Предметом* вивчення в роботі був корпус рекламних текстів. *Мета нашого дослідження* – прослідкувати специфіку функціонування мовних засобів рекламного тексту та різних методів лінгвістичного та психолінгвістичного маніпулювання свідомістю споживача. *Завдання* роботи: на основі лінгвістичного аналізу бази даних рекламних текстів визначити її вплив на масову свідомість і колективну поведінку.

Інформація про рекламний текст, яка міститься в корпусі, надає можливість наочно побачити весь масив проблемного матеріалу. Проте формуванню корпусу текстів повинен передувати філологічний аналіз фактичного матеріалу.

Насамперед потрібно паспортизувати усі рекламні тексти: слід навести джерело, означити рекламований предмет, визначити довжину тексту (як одного із факторів дієвості). Отже, базу рекламних текстів для зручності користування можна розділити на такі масиви: **джерело рекламного тексту, цільова аудиторія, предмет рекламування, вплив** (лінгвістичний, психолінгвістичний).

Одна з найважливіших умов репрезентативності текстів передбачає, що склад корпусу має бути деяким чином збалансований за жанрами:

- ✓ Реклама-діалог: рекламне звернення представлене у формі діалогу;
- ✓ Порівняльна реклама: у ній зіставляються споріднені предмети з метою пояснення одного за допомогою іншого;
- ✓ Реклама-інструкція: у цьому випадку рекламна аргументація подана у формі інструкції;
- ✓ Реклама-тест: пропонується відповісти на запитання, а далі іде пояснення правильної відповіді;
- ✓ Реклама-загадка: у цій моделі рекламний текст починається з питання;

* © *Т. Сидорко, 2006*

✓ Реклама за моделлю “оберненої піраміди”: найважливіші аргументи йдуть на початку, а далі наводяться всі інші;

✓ Драматизована реклама: основні характеристики цієї реклами – це наявність конфлікту і його подолання.

Проаналізувавши базу з 200 рекламних текстів ми виявили, що найпопулярнішими є моделі “оберненої піраміди” та реклами-діалогу:

Нові можливості професіного прання – імпрегування на дому! SIEMENS. У кожного з нас є верхній одяг, відштовхувальні властивості якого з часом вимагають відновлення, – куртки, плащі, пуховики, вироби з мікрофази. З цією метою використовується імпрегування – просочування тканин спеціальними речовинами. Раніше це було можливе лише в деяких спеціалізованих хімчистках. Сьогодні це можна робити вдома в нових пральних машинах SIEMENS WIQ 1633EU/WIQ 1433EU, у яких передбачена унікальна програма "Імпрегування". Варто лише залити спеціальний засіб у кювету для ополіскування і обрати відповідний режим.

– Що хвилює багатьох?

– Неврози, страхи, депресії, порушення харчової поведінки, конфліктні ситуаці.

– Хто допоможе?

– Допомога професійних психотерапевтів, психологів та психіатрів, індивідуальні програми лікування.

– Як зробити перший крок?

– Зателефонуєте до нас, зустріньтеся зі спеціалістом, почніть вирішування проблем за допомогою професіоналів.

Зважаючи на прогрес сучасної психолінгвістики, рекламісти широко використовують її досягнення. Зокрема, залучають до написання рекламних текстів розробки в галузі NLP (Neuro-Linguistic Programming). Виходячи з цього, ми можемо отримати дуже потужний корпус рекламних текстів, розбитий за такими жанрами, де реклама спрямована на:

- Аудіалів
- Візуалів
- Кінестетиків

Приклади використання таких рекламних текстів:

- *TherapyAir – Вбиває бактерії, плісняву та віруси – **Безшумно** нейтралізує нав'язливі запахи* (для аудіалів);

- *Impression стійка крем -фарба **Найяскравіші** кольори Фантазії! Здорове волосся і неперевершене **сяйво!** Завжди впевненість у результаті* (для візуалів);

- ***Втамуй спрагу** творчості за допомогою цифрового фотоапарату BenQ DC E 510 із сенсорною панеллю та потужною функцією редагування фотографій !* (для кінестетиків).

Ефективна комунікація, на думку спеціалістів NLP, передбачає діалог зі споживачем „його мовою”. Цікаво, що за даними досліджень уплив на співрозмовника зумовлюється на 55% мовою рухів тіла: позами, жестами, візуальним контактом; 38% – тоном голосу; і лише 7 % – змістом вислову [1]. Практичне застосування NLP також пов’язано з подвійною дією слова (емоційно-образною та раціонально-логічною). Це дозволяє за допомогою маніпуляцій зі словами сформувати у людини неусвідомлене нею позитивне або негативне ставлення до рекламного тексту.

Хотілося б також звернути увагу на ще один із жанрів рекламування. Зважаючи на психологічні чинники, досліджуваний корпус рекламних текстів можна розподілити на группи згідно з різними підходами, що використані рекламістами при описуванні рекламованих продуктів:

- фактологічний;
- образний;
- емоційний.

Оголошення, створені за першим принципом, найчастіше наводять факти, і мають раціональний характер. Автори таких текстів мають справу з реальністю. Передусім ідеться про сам рекламований об’єкт: з чого він складається, з чого зроблений, у чому полягає послуга. Концентруючи увагу на фактах, найважливіших для адресата, автор переконує в перевагах рекламованого продукту чи послуги. У таких оголошеннях не спостерігаємо використання емоційно насичених мовних одиниць, лексичних засобів увиразнення викладу. Характерним є вживання стандартизованих словесних зворотів, які дозволяють економити час, не відволікають читача від сприйняття основної інформації:

Iriver
T30 – нова серія MP3-плеєрів

- 256Mb /512Mb/ 1Gb
- MP3-плеєер
- Диктофон
- FM-тюнер
- Функція запису з FM-тюнера
- Аудіоформати: MP3, WMA, Ogg Vorbis, ASF
- 3-х рядковий ЖК-дисплей
- Годинник з таймером, будильник

При застосуванні **образного** підходу факти залишаються фактами, але мистецтво успішного рекламування полягає у вираженні відомих фактів несподіваним способом, створенні образу, здатного буденну інформацію показати під несподіваним кутом зору, рельєфно демонструючи певну перевагу рекламованого об'єкта. Обов'язковою умовою успішного рекламування є не просто перелік пропонованих об'єктів, а й свіжий погляд на описувані предмети. Словесний образ рекламованого об'єкта виступає як певне художнє узагальнення й передається семантичною конструкцією, створеною в результаті використання виразних висловлювань, у який інтерпретуються й перетворюються усталені значення слів, що допомагає показати речі та явища в оригінальному ракурсі. Його також розглядають як найефективніший тип повідомлення, що реалізується в умовах браку інформації та часу, навіть неуважності співрозмовника, яким є адресат рекламного тексту.

Braun Multiquick. Вміє багато. Як і Ви. **Ви – мама і менеджер, вчитель і друг, кухар і артистка, лікар і няня. І щоб з усім впоратись, вам конче потрібен універсальний помічник. Braun розробив нову систему Multiquick – надійного партнера у всьому, що стосується здорового харчування вашої сім'ї щодня. Нова вакуумна** система Braun Multiquick Талановита та інноваційна. Вона вправна і талановита, як ви. **Вакуумна система Braun Multiquick від Braun допомагає легко і швидко змішувати, нарізати, збивати будь-які продукти, і навіть чудово подрібнювати горіхи. З нею так просто куховарити щодня, а за потреби й на кілька днів вперед.**

Виділені мовні одиниці "***вміє багато. Як і Ви***", "***система Braun Multiquick Талановита та інноваційна", "Вона вправна і талановита, як ви***" використані з метою підкреслити, що споживач розумний і талановитий, а техніка цілком відповідає вимогам покупця. Актуалізація у рекламі позитивних якостей потенційного споживача здійснюється для того, щоб споживач підсвідомо налаштувався на те, що техніка така сама гарна, як і він.

Тексти, створені з використанням **емоційного** підходу, відрізняються виразністю й безпосередньо апелюють до емоцій адресата, змушують читача відчути радість, задоволення або навіть почуття розпачу під час сприйняття оголошення.

Faberlic – киснева косметика. Як і Ви, Ваша шкіра дихає. Як і Вам, їй дуже потрібне джерело життєвої енергії – кисень. Завдяки йому і відбувається обмін речовин. Але кисню мало в атомосфері міста. Шкіра ***починає*** *повільно* ***задихатися****, втрачає здатність відновлюватись,* ***старіє. Н****асичуючи шкіру киснем, ви допомагаєте їй добре виглядати. Знову в моді природна краса, а отже, у моді Кисень!*

У цьому оголошенні емоційне тло створюють виділені лексичні одиниці, використані з метою викликати страх споживача перед наслідками, які спричиняє атмосфера міста, і єдиним вирішенням проблеми та порятунком для шкіри є рекламований продукт.

Рекламні звернення, які апелюють до почуття любові, ненависті, страху, виявляються дуже впливовими. Не менш важливим є також емоційний компонент рекламного впливу, який визначає емоційне ставлення до об'єкта рекламної інформації: чи ставиться до нього адресат із симпатією, антипатією, нейтрально або суперечливо.

Зважаючи на представлення корпусу рекламних текстів в електонній базі данних, дуже зручним виявилось те, що на відміну від паперових носіїв, електоронну базу можна редагувати, поповнювати, систематизувати та здійснювати швидкий пошук потрібного матеріалу залежно від поставленої задачі.

На корпусах рекламних текстів можна перевіряти лінгвістичні та психолінгвістичні теорії; їх можна використовувати для будь-яких практичних робіт із мовним матеріалом, зокрема як масиви прикладів для демонстрації певних явищ.

В. В. Риков визначає корпус текстів як деяке зібрання текстів, в основу яких покладений логічний задум, логічна ідея, що об'єднує ці тексти [5]. Часто окремі тексти використовують для різних видів лінгвістичних досліджень. Робота з корпусами, тобто з масивами текстів, представлених у комп'ютерному вигляді, давно вже стала одним із основних методів лінгвістичних досліджень.

Отже, в результаті створення нашої бази рекламних звернень, ми отримаємо корпус текстів, розбитий за певними рубриками (жанрами). Залежно від того, яким із представлених жанрів скористаємось, ми можемо отримати ті чи інші висновки лінгвістичного дослідження. Накопичений корпус може служити базою для подальших філологічних і психолінгвістичних досліджень.

Література

1. Лебедев А. Н., Боковиков А. К. Экспериментальная психология в российской рекламе. – М.: ИЦ «Академия», 1995. – 16 с.
2. Мокшанцев Р. И. Психология рекламы. – М.: ИНФРА-М, Новосибирск: Сибирское соглашение, 2000. – 230 с.
3. Москальская О. И. Текст как лингвистическое понятие // Иностранные языки в школе. – 1978. – №3. – С. 13–15.
4. Пирогова Ю. К., Баранов А. Н. и др. Рекламный текст: семиотика и лингвистика. – М., 2000.
5. Рыков В. В. Корпус текстов как отражение состояния русского языка // Труды Международного конгресса "Русский язык: исторические судьбы и современность" . – Москва: МГУ, 2001.

Н. Лозова*
Інститут української мови НАН України (Київ)
УДК 161.2.81'373.322

ДОСЛІДЖЕННЯ ДІЄСЛІВ ПОВІДОМЛЕННЯ НА БАЗІ РОБОЧОЇ ВЕРСІЇ НАЦІОНАЛЬНОГО КОРПУСУ УКРАЇНСЬКОЇ МОВИ

Дослідження лексики може ґрунтуватися на класичній вибірці з картотеки чи бути корпусно базованим. У статті здійснено спробу виявити активність функціонування окремих слів з лексико-семантичної групи дієслів повідомлення української мови, застосовуючи корпусний підхід. Метою нашого дослідження є аналіз дієслів повідомлення на робочому варіанті підкорпусу художнього стилю Національного корпусу української мови (на основі прозових творів).

Робоча версія підкорпусу художнього стилю характеризується репрезентативністю та збалансованістю [1].

До підкорпусу ввійшли твори, різні за жанрами (роман, повість, оповідання, новела), написані протягом XIX–XX ст. Кількість залучених творів та кількість слів у фрагментах відповідають рекомендаціям до корпусу НКУМ [2] (іноді виникають розбіжності у визначеннях літературознавцями жанру твору; ми пристаємо на найпоширеніші):

роман (100 тис. слів)	20 фрагментів	5 тис. слів у фрагменті
повість (100 тис. слів)	20 фрагментів	5 тис. слів у фрагменті
оповідання (30 тис. слів)	10 фрагментів	3 тис. слів у фрагменті
новела (20 тис. слів)	10 фрагментів	3 тис. слів у фрагменті

У підкопусі представлено тексти:

романи

1. Р. Андріяшик Люди зі страху. II пол. XX ст.
2. Б. Антоненко-Давидович За ширмою. II пол. XX ст.
3. І. Багряний Сад Гетсиманський. Сер. XX ст.
4. І. Білик Меч Арея. II пол. XX ст.
5. В. Винниченко Слово за тобою, Сталіне. Сер. XX.
6. О. Гончар Берег любові. II пол. XX ст.
7. А. Головко Бур'ян. I пол. XX ст.
8. В. Дрозд Катастрофа. II пол. XX ст.
9. П. Загребельний Євпраксія. II пол. XX ст.
10. Панас Мирний Повія. II пол. XIX ст.

* © *Н. Лозова, 2006*

11. Д. Міщенко Синьоока Тивер. II пол. XX ст.
12. Ю. Мушкетик На брата брат. II пол. XX ст.
13. П. Куліш Чорна рада. Сер. XIX ст.
14. С. Скляренко Святослав. Сер. XX ст.
15. М. Стельмах Велика рідня. Сер. XX ст.
16. Григорій Тютюнник Вир. II пол. XX ст.
17. Р. Федоров Єрусалим на горах. II пол. XX ст.
18. І. Франко Борислав сміється. II пол. XX ст.
19. Вал. Шевчук Дім на горі. II пол. XX ст.
20. Ю. Яновський Майстер корабля. I пол. XX ст.

повісті

1. Б. Антоненко-Давидович Смерть. I пол. XX ст.
2. С. Васильченко Олив'яний перстень. I пол. XX ст.
3. Марко Вовчок Інститутка. Сер. XIX ст.
4. О. Гончар Микита Братусь. Сер. XX ст.
5. К. Гриневичева Шоломи в сонці. I пол. XX ст.
6. Б. Грінченко Серед темної ночі. Поч. XX ст.
7. Є. Гуцало Княжа гора. II пол.XX ст.
8. О. Донченко Лісничиха. Сер. XX ст.
9. А. Кащенко Зруйноване гніздо. Поч. XX ст.
10. Г. Квітка-Основ'яненко Маруся. Поч. XIX ст.
11. О. Кобилянська В неділю рано зілля копала. Поч. XX ст.
12. М. Коцюбинський Фата моргана. Поч. XX ст.
13. И. Микитенко Вуркагани. I пол. XX ст.
14. Ю. Мушкетик Обвал. II пол. XX ст.
15. І. Нечуй-Левицький Дві московки. II пол. XIX ст.
16. І. Пільгук Дуби шумлять. II пол. XX ст.
17. А. Тесленко Страчене життя. I пол. XX ст.
18. Ю. Федькович Три як рідні бріти. II пол. XIX ст.
19. І. Франко Boa constriktor. II пол. XIX ст.
20. Г. Хоткевич Аврон. I пол. XX ст.

оповідання

1. Бордуляк Дід Макар. I пол. XX ст.
2. М. Вінграновський Гусенятко. II пол. XX ст.
3. А. Головко Пилипко. I пол. XX ст.
4. А. Кащенко Оповідання про славне військо запорозьке низове. I пол. XX ст.
5. М. Коцюбинський В дорозі. Поч. XX ст.
6. П. Куліш Гордовита пара. Сер. XIX ст.
7. Панас Мирний День на пасівнику. II пол. XIX ст.
8. І. Нечуй-Левицький Афонський пройдисвіт. II пол. XIX ст.
9. А. Тесленко Школяр. Поч. XX ст.
10. А. Чайковський За сестрою. I пол. XX ст.

новели

1. В. Винниченко Студент. Поч. XX ст.
2. С. Васильченко Мужицька арихметика. Поч. XX ст.
3. О. Гончар Кресафт. II пол. XX ст.
4. Є. Гуцало Запах кропу. II пол. XX ст.
5. Г. Косинка В житах. I пол. XX ст.
6. М. Коцюбинський Сон. I пол. XX ст.
7. В. Стефаник Кленові листки. I пол. XX ст.
8. Григір Тютюнник Лісова сторожка. II пол. XX ст.
9. І. Франко Schon Schreiben. Кін. XIX – поч. XX ст.
10. М. Хвильовий Вальдшнепи. I пол. XX ст.

Відібрано слова, що найбільш нейтрально й повно характеризують процес (або відсутність процесу) передання повідомлення, його одержання, сприйняття та володіння інформацією й становлять опозицію сем

„передавати інформацію” – „одержувати інформацію”

„передавати інформацію” – „володіти інформацією”.

Список становлять лексеми **повідомляти, сповіщати, інформувати, довідуватися, сприймати, таїти, знати, пам'ятати.**

Дієслова **повідомляти, сповіщати** та **інформувати** мають значення *"передання інформації в найзагальнішому аспекті"* (без додаткових сем), що мало характерно для художніх творів. Цим можна пояснити досить обмежену поширеність уживання цих дієслів. Так, на **повідомити** натрапляємо 8 разів, на **повідомляти** – 2 рази. Зафіксовано одне вживання віддієслівного іменника **повідомлення. П**ричому більшість із цих фіксацій припадає на 60–80-ті роки XX ст. і дві – у творах А. Кащенка та А. Чайковського – на початок XX ст.

У II половині XX ст. майже в усіх випадках дієслова **повідомляти, повідомити** й віддієслівний іменник **повідомлення** мають значення саме *„передавати інформацію"* (виняток становить роман П. Загребельного „Євпраксія", де **повідомити** означає *„оповістити"*, що можна пояснити стилістично-жанровою специфікою твору). На відміну від творів середини XX століття, в оповіданнях початку XX ст. зафіксовано дієслова **повідомити** зі значенням *„доповісти"*.

Подібна картина й з дієсловами **сповіщати – сповістити. Ц**і дієслова теж уживані у творах 10–20-их рр. та II пол. XX ст. Однак із 5 вживань (3 **сповістити** і 2 **сповіщати**) 4 припадає на початок століття й лише один – в І. Пільгука – на II пол. XX ст. – у знач. *„доповідати"*. Обидва дієслова – і **повідомляти,** і **сповіщати –** зустрічаємо лише в І. Пільгука й А. Чайковського.

Інформувати й похідні з семою *„повідомлення"* вживано 6 разів:

інформувати – 1;
поінформувати –2;
поінформований – 1;
непоінформований – 1;
інформація – 1.

Часові межі вживання лексеми **інформувати** та похідних – початок та друга половина XX ст. (як і в попередніх дієсловах). **Інформувати** поряд з **повідомляти** і **сповіщати** виявлено лише у В. Дрозда в романі „Катастрофа".

Таким чином, спостерігаємо досить малу частотніть дієслів **повідомляти, сповіщати, інформувати** та їх похідних. При цьому жодного вживання цих дієслів до XX ст. не зареєстровано, це дає підстави припустити, що вони стали активно функціонувати відносно недавно, а в художній прозі залишилися маловживаними. Зрозуміло, що зовсім інші дані одержимо з підкорпусу публіцистики. У художній літературі ж більш уживаними є дієслова з додатковими семами на позначення яким саме способом передається повідомлення.

Значення *„передавати нульову інформацію"*, або *„приховувати інформацію"* має дієслово **таїти. Ц**е дієслово та однокореневі й близькі за семантикою слова трапляються 33 рази:

таїти – 1;
затаїти – 1;
притаїти – 1;
таїтися – 1;
таємниця – 2;
тайна – 3;
таємний – 7;
таємничий – 8;
потайний – 3;
нетаєний – 2;
таємничо – 1;
потайки – 2;
потай – 1.

Та лише три з них мають відношення до розглядуваної групи. Це **таємниця, тайна** (обидва зі значенням *„секрет"*) і **затаїти** (*„приховати"*) – останній з „Чорної ради" П. Куліша.

Маловживаним для досліджуваних текстів є дієслово **сприймати** з похідними, що позначають процес одержання інформації:

сприймати – 3;
сприйматися – 3;
сприйняти – 2;
сприймання – 1.

Серед цих слів жодного з семою *„осягати розумом"*, а лише *„трактувати"*, тобто вони не є об'єктом дослідження.

Набагато частіше трапляються дієслова з семою *„мати"* або *„зберігати інформацію"* та їхні похідні. 48 випадків уживання слів з семами *„пам'ятати"* і *„згадати"* (слів з іншими семами, а також вставних слів до уваги не беремо):

пам'ятати – 20;
запам'ятати – 5;
запам'ятатися – 4;
запам'ятовувати – 2;
пам'ятувати – 4;
пам'ятаючи – 1;
пам'ять – 12.

Дієслово **запам'ятати** 4 рази зустрічаємо з семою *„пам'ятати"* і 1 – *„забути"* (*„азбуки запам'ятав уже"* С. Васильченко).

Досить поширеними є словосполучення з іменником **пам'ять** зі знач. *„пам'ятати"* (закарбуватися в пам'яті, зафіксувати в пам'яті, зберегти в пам'яті, мені в пам'яті); *„згадати"* (у пам'яті виринуло, відтворити в пам'яті, випливти (спливти) в пам'яті; *„забути"* (не зітреш з пам'яті; щезнути з пам'яті; не буде нам ні слави, ні пам'яті).

Пам'ятати та похідні є майже в усіх досліджуваних фрагментах, тобто вони залишаються активними протягом XIX–XX ст.

Але найбільш активно вживане дієслово **знати** – *„володіти інформацією"*.

За „Словником української мови" [3], **знати** має 9 значень, із яких зупинимося на першому – *„мати якісь дані, відомості про кого-, що-небудь"*. Ті похідні, які означають *„мати відомості"* або *„одержувати відомості"*, теж беремо до уваги. *„Володіти знанням"* та *„розуміти"* не розглядаємо. Не враховуємо також ті слова, що входять до словосполучень з іншими значеннями, вставні слова тощо. У фрагментах дієслово разом з похідними зустрічаємо 629 разів, із яких 198 відповідають заданим параметрам. Таким чином, виходить така картина:

знати – 177 (з 567);
знаючи – 2 (з 12);
дізнатися – 6 (з 10);
дізнаватися – 3 (з 4);
дізнавшись – 3 (з 5);
дознавшись – 1 (з 1);
пізнати 2 (з 25);
узнати – 4 (з 5).

Дієслово **знати** є в усіх фрагментах підкорпусу з найдавніших часів до сьогодення.

Похідні від **знати** теж з'являються досить рано, після чого функціонують стабільно в текстах усіх розглядуваних етапів.

Цікавою особливістю є те, що **знати** та похідні досить часто вживаються з іншими з досліджуваних слів у тому самому реченні. Наприклад: *„— Хіба я знаю? Десь раніше. Так давно, що й не пам'ятаю?"* або *„Того я не знаю. Та й ніколи мені вже те довідуватися"* (Вал. Шевчук); *„Княжа гора сидить на підземних печерах-чернерах, мов степова орлиця на гнізді, гай-гай, уже скільки століть відшуміло, як вона оберігав своє гніздо, а що в тому гнізді — сама вона лиш і знає, мабуть, таємниці"* (Є. Гуцало).

Часто в одному реченні фіксуємо **знати** поряд з синонімами: *„— Це та вже, що допоможе? — Не знаю, соколику. Котрась має допомогти, а котра — не відаю"* (Д. Міщенко „Синьоока Тивер").

Якщо порівняти одержані дані з даними „Частотного словаря русского языка" [6], виявимо, що вони подібні. Так, при досить широкій уживаності дієслова **сообщать**, він майже не вживається в художній літературі, натомість поширений у газетно-журнальних текстах. Мала частотність у дієслів **информировать** (6 вживання, з них жодного з художньої прози), **воспринимать** (з 18 вживань 1 – художня проза) **– воспринять** (6 – 0), **узнавать** (40 – 17), **таить** (23 – 6) – **утаить** (7 – 4). Набагато частіше зустрічається в художній прозі **помнить** (333 – 70). Правда, **узнать** більш поширений, ніж **помнить** (343 – 131). Але найуживанішим є дієслово **знать** (2192 вживання, з них 568 припадає на художню прозу). Навіть якщо зняти з розрахунків омонім цього дієслова (іменник), це несуттєво вплине на загальну цифру.

Слід зазначити, що в „Частотном словаре русского языка» йдеться про 13 творів 11 авторів винятково XX ст., і становлять вони 25, 4 % від усіх текстів. Таким чином, порівнювати одержані дані можна лише з певним застереженням.

Набагато повнішим є „Частотний словник сучасної української художньої прози" [4; 5]. Тут розглядається 25 творів 22 авторів середини XX ст. Вибірки налічують по 1000 слововживань. Картина вимальовується майже така сама: **повідомляти** (4 вживання з 4 творів) **повідомити (**19 з 11**), сповіщати** (3 з 3) **– сповістити** (12 з 8)**, інформувати** (подається т. іменник **інформація** – 6 з 4), **таїти** (3 з 3), **сприймати** (13 з 10) – **сприйняти** (15 з 10) досить маловживані. Трохи вживанішим є дієслово **довідатися** (36 з 15). Набагато частіше трапляється дієслово **пам'ятати** (97 з 24 – майже з усіх залучених творів). І в 13, 6 разів більше, ніж **пам'ятати,** вживається дієслово **знати** (1323 з 25, тобто з усіх залучених творів) (табл. 1):

Таблиця 1

Дієслова	Кількість досліджених творів (у дужках – відсоток від загальної кількості розглянутих творів)			Частота вживання		
	за „Частотним словарем русского языка"	за „Частотним словником	за одержаними нами даними	за „Частотним слова-рем	за „Частотним словником сучасної	за одержани-ми нами даними

		сучасної україн-ської художньої прози”		русско-го языка”	україн-ської художньої прози”	
сповіщати	5 (38%) сообщать	3 (12%)	1 (2%)	6	3	1
сповістити	9 (69%) сообщить	8 (32%)	3 (5%)	27	12	3
повідомляти		4 (16%)	1 (2%)		4	1
повідомити		11(44%)	7 (12%)		19	8
інформувати	– информировать	–	1 (2%)	–	–	1
сприймати	1 (8%) воспринимать	10 (40%)	3 (5%)	1	13	3
сприйняти	– воспринять	10 (40%)	2 (3%)	–	15	2
довідуватися	7 (54%) узнавать	–	1 (2%)	17	–	1
довідатися	13 (100%) узнать	15 (60%)	11 (18%)	131	36	13
таїти	4 (31%) таить	3 (12%)	1 (2 %)	6	3	1
пам’ятати	11 (85%) помнить	24 (96%)	27 (45%)	70	97	44
знати	13 (100%) знать	25 (100%)	60 (100%)	568	1323	567

Таким чином, наша вибірка майже повністю підтверджує дані розглянутих частотних словників, пропонуючи нові можливості дослідження частотності вживання слів.

Література

1. Демська-Кульчицька О. Репрезентативність як ознака текстового корпусу // УМ. – 2005. – № 3.
2. Демська-Кульчицька О., Балог В. Національний корпус української мови: підкорпус художньої літератури // Лексикографічний бюлетень. – Київ, 2005. – Вип. 12.
3. Словник української мови. – Т. 3. – К.: Наук. думка, 1972. – С. 642–644.
4. Частотний словник сучасної української художньої прози / У 2-х томах. – Т.1. – К.: Наук. думка, 1981. – 864 с.
5. Частотний словник сучасної української художньої прози / У 2-х томах. – Т.2. – К.: Наук. думка, 1981. – 856 с.
6. Частотный словарь русского языка / Ред. Л. Н. Засорина. – М.: Русский язык, 1977. – 936 с.

Т. Монахова*
Інститут української мови НАН України (Київ)
УДК 161.2.81’322

КОРПУС ТВОРІВ ВАЛЕРІЯ ШЕВЧУКА: ФОРМАТ РОЗМІТКИ БАЗОВИХ КОНЦЕПТІВ

У статті розглянуто проблему корпусного підходу до вивчення концептів. Пропонуються принципи анотування семантичних реалізацій концепту в художньому тексті.Ключові слова: концепт, семантична анотація.

Корпусна лінгвістика сьогодні пропонує свої методи для дослідження різних аспектів мови: її морфологічної будови, синтаксичного устрою, лексичних значень тощо. Ми здійснюємо спробу застосування корпусних методів до вивчення базових концептів мови творів Валерія Шевчука. Наше дослідження передбачатиме створення авторського корпусу мови Валерія Шевчука із анотуванням формально-смислових параметрів базових концептів. Планований корпус творів Валерія Шевчука згідно з усталеними класифікаціями текстових корпусів [2: 25] буде ілюстративним, а отже створюватиметься після визначення й детального вивчення досліджуваних концептів.

Концепт — це складне абстрактне явище духовної культури, людської психіки, моральних принципів тощо, яке базується на осмисленні його тлумачень в усі періоди

* © *Т. Монахова, 2006*

історії людства відповідно до національної конотації, індивідуального досвіду людини. Це свідоме оперування культурною метаінформацією через використання загальнолюдських, національних і особистісних смислів. Концепт може презентуватися символами, образами, реалізується у фразеологічних одиницях, закріплюється в культурних традиціях. Його лінгвальна реалізація здійснюється через вживання назв концептів (імен концептів) у прямих і переносних значеннях.

Виходячи з такого розуміння концептів, їх аналіз є передовсім аналізом смислів, які нашаравуються в структурі самого концепту, а точніше їх реалізації в конкретному художньому тексті (корпусі художніх текстів). Процедуру семантичного аналізу смислів концепту можна продемонструвати на прикладі концепту *шлях* у творчості Валерія Шевчука, зокрема в романі «Дім на горі».

Формат розмітки становить чи не основну проблему корпусного анотування. Якщо схеми морфологічного маркування текстових даних розроблені досить чітко [1; 2], то принципи анотації концептів у лінгвоукраїністиці ще потребують з'ясування. Отже, спробуємо запропонувати певні підходи до реалізації розмітки формально-смислових параметрів базових концептів у корпусі текстів Валерія Шевчука на прикладі концепту ***шлях***.

Для здійснення так званої концептної анотації необхідно зробити таке:

1. У текстах В. Шевчука визначити всі лінгвальні реалізації концепту, наприклад: «*Йому здалося, що ця вулиця — не просто шлях, який так собі кладеться, весь простір наповнювався тими золотистими хвилями, і чи від блимотливих калюж серед дороги, чи від повітря видалося йому, що шлях, яким ішов, дивно запрозорів*».

2. Визначені контексти згрупувати за прямим і переносним значенням слова-імені концепту *(шлях, дім, дорога, стежка тощо)*, наприклад: а) «*Марія подумки проміряла той шлях, що пройшли його разом, тихий смуток заповнив її радість*» — переносне значення (шлях = подружнє життя); б) «*Вряди-годи озиралися вони на синю дорогу, по якій ішов Микола, приставляли долоні дашком і вдивлялися*» — пряме значення.

3. Слова-імена концептів з прямим значенням сортувати в такий спосіб: другорядні прямі значення — основні прямі значення. О***сновне*** (центральне) ***пряме значення*** — це таке пряме значення, яке організовує у тексті простір, є частиною хронотопу або конкретним образом твору. П*риклад: «Пізніше він пройде по цій дорозі ще раз»*. ***Другорядне пряме значення*** — це таке вживання слова в прямому значенні, яке не впливає на розвиток подій у творі; це супровідне, обставинне пряме значення. П*риклад: «Ген віддалік на далекій курній дорозі їхав віз, тягнучи за собою довгий білий шлейф»*.

4. Слова-імена концептів з переносним значенням розподілити таким чином: пасивне переносне значення — активне переносне значення. П***асивне переносне значення*** — вживання слова-імені концепту фігурально з метою створення художнього ефекту. П*риклад*: *«Саме туди вела синя, мерехтлива дорога, по якій ішов»* — метафора й епітет зі словом «дорога». ***Активне переносне значення*** — фігуральне слововживання, яке має програмний, лейтмотивний смисл. Активні переносні значення реалізують смисли концепту. П*риклад: «Він таки був тут, біля їхніх вікон, а зараз уже повільно відходить по синій дорозі»* — дорога в значенні «потойбічне життя, смерть»).

5. Окрему категорію становлять подвійні смисли, тобто коли у вислові пряме основне значення накладається на активне переносне. Таке значення можна назвати символічним. Власне, с***имволічні значення*** становлять смисли концепту. П*риклад: «Серед неба світив місяць, і був він у ту ніч такий яскравий, що дорога, по якій вони йшли, засвітилася синім вогнем. Микола обережно взяв її під руку, і вона довірливо подалася до нього: мали пройти по тій дорозі десять років. Через рік до них прилучиться на їхньому шляху ще один супутник, і наприкінці десятого року стане їх семеро»* — а) конкретна дорога, на якій зустрілися і покохалися парубок і дівчина; б) подружнє життя, народження дітей.

6. Символічні значення далі характеризувати за такими ознаками:

– іконічні (або візуальні) [3: 98], тобто всі зовнішні характеристики, виражені прикметниками, прислівниками тощо. П*риклади: мерехтлива дорога, синя дорога тощо.*

– функціональні ознаки [3: 98]— дійові характеристики, виражені найчастіше дієсловами. П*риклади: шлях кладеться, вела дорога, дорога засвітилася вогнем тощо.*

– етнічні ознаки — специфічні національні характеристики: *долати важку дорогу (за асоціацією з чумаками, бурлаками); шлях полотном стелиться тощо.*

Процедуру семантичного аналізу концептів можна подати схематично:

Таблиця 1.

<table>
<tr><td colspan="4">ІМ’Я КОНЦЕПТУ</td></tr>
<tr><td colspan="2">Пряме значення</td><td colspan="2">Переносне значення</td></tr>
<tr><td rowspan="2">Другорядне пряме значення</td><td>Основне пряме значення</td><td>Активне переносне значення</td><td rowspan="2">Пасивне переносне значення</td></tr>
<tr><td colspan="2">СИМВОЛІЧНЕ ЗНАЧЕННЯ</td></tr>
</table>

Для кодування смислів за пропонованими параметрами послуговуємося латинськими символами:

1. C — concept — концепт, ім’я концепту;
2. L — literal meaning — пряме значення;
3. F — figurative meaning — переносне значення;
4. M — minor — другорядне;
5. B — basic — основне;
6. A — active — активне;
7. P — passive — пасивне;
8. S — symbol — символічне;
9. Ab — активне переносне значення, що накладається на основне пряме;
10. Ba — основне пряме значення, на яке накладається активне переносне.

Пропонований мінімальний набір релевантних класифікаційних критеріїв та їх кореляція можуть бути представлені в такий спосіб:

Таблиця 2

№	I	II	III	IV	V
кри-терій	**Назва твору**	**Ім’я кон-цепту**	**Тип значення (прям./ непряме)**	**Основне/ другор. Активне/ пасивне**	**Лінгвальна реалізація концепту**
те-ги	Дім на горі **DNG**	**C**	**L / F**	**M / B A / P Ab/Ba**	Конкретна цитата з твору
н.п.	DNG	R (road)	L	B	Ніхто не знав про його хід цією дорогою…
н.п.	DNG	R	L	M	По тому вони побачили батька, що повертався з вулиці.
н.п.	DNG	R	F	P	… дорога, по якій вони йшли, засвітилася синім вогнем.
н.п.	DNG	R	F	A	Хто зна, може, в такий спосіб її дух долав отак кілометр за кілометром синю дорогу — час відлетілий та забутий…
н.п.	DNG	R	F	Ab	Вона подивилася далі, собі під ноги, і знову побачила там шматок синьої дороги, по якій ішла вона і простували й інші.

н.п.	DNG	R	L	Ba	Вже давно думала про дорогу. Ввижалася їй нескінченним білим змієм, бо простягалась у визорі звідсіля, від їхнього передмістя, туди, в далекий безмежний світ, думати про який було страшно й лячно.

Таким чином, спостерігаються кореляції F із L і, відповідно, A / P із M / B.

При цьому S = Ba, S = Ab, Ba = Ab = S.

Відповідно до зазначеного, корпусне анотування концептів має такий вигляд: ***Наша присутність — це і є та дорога, яку я так часто останнім часом бачу {DNG/R/F/A}*** («Дім на горі», концепт «Шлях», переносне значення, активне переносне значення).

Таким чином, концептні смисли реалізують значення S (Ba, Ab) та FA, тобто символічні значення і активні переносні значення. Можливо, символічні значення, визначені як поєднання активних переносних і основних прямих значень, є прикметою ідеостилю Валерія Шевчука, який свідомо накладає абстрактні глибинні смисли на конкретні образи у своїх творах. Тому ефективність пропонованої схеми аналізу на матеріалі творів інших авторів потребує перевірки.

Література

1. Автоматизация анализа научного текста / В. А. Вербицкий, Т. А. Грязнухина, Н. П. Дарчук и др. – К.: Наукова думка, 1984. – 256 с.
2. Демська-Кульчицька О. Основи національного корпусу української мови. – К., 2005. – 219 с.
3. Іващенко В. Концепт-символ «кобзар» у життєтворчості Т. Г. Шевченка // Материалы IV Международного семинара «Шевченковский Петербург». – СПб., 2005. – 186 с.
4. Никитин М. Курс лингвистической семантики: Учебное пособие для студентов, аспирантов и преподавателей лингвистических дисциплин в школах, лицеях, колледжах и вузах. – СПб., 1996. – 760 с.

Ł. Grabowski*
Institute of East-Slavonic Studies University of Opole (Opole, Poland)
УДК 81’322

COMPARATIVE CHARACTERISTICS OF POLISH AND RUSSIAN LANGUAGE CORPORA

The aim of this paper is to present and compare the most representative language corpora of Polish and Russian according to selected criteria determining both their potential as a source of linguistic data for various types of linguistic analyses and their availability for researchers. Moreover, the present paper indicates areas for improvement as far as the possibilities offered by the corpora and access to them are concerned.

1. Introduction

Corpus linguistics, which is both a branch of linguistics and a methodology of research, concentrates on the study of texts with the use of dedicated computer software [4:143]. Thus, any collection of texts stored in electronic form is called a corpus. The linguistic data stored in the corpus is the actual collection of electronic words, which are classified in terms of types and tokens [2:16]. This typology is important in that a token is every running word-form (segment) the corpus is composed of, whereas a type is a group of the same tokens (eg. the sentence ‘*Моя мама живёт в городе, а моя тётя живёт в деревне’* consists of 8 types and 11 tokens).

Since language corpora, unlike single and coherent texts, are objects of complex and multidimensional character, the answer to the question how similar or different they are will indubitably be complex and multidimensional itself. Thus, the study aimed at comparing corpora that represent two different languages will be even more challenging and complex. Corpora as objects of comparison will always be similar in some aspects and different in others. For the comparison to render objective observations, one has to adopt a framework that would serve as criteria for such a study. Although there have been multiple quantitative methods based on word frequencies and ngram frequencies [1], purely statistical approach to a comparison of different multilingual corpora falls beyond the scope of the current paper. Thus, its aim is to present possibilities offered by the most representative commercial corpora representing Polish and

* *© Ł. Grabowski, 2006*

Russian languages as far as the following criteria are concerned: a) the overall number of tokens collected in corpora; b) availability of meta-linguistic annotation and lemmatisation of a corpus data; c) types of queries offered by corpora; d) availability of corpora and access to them. The Polish corpora are represented by Korpus Języka Polskiego Wydawnictwa Naukowego PWN (henceforth PWN Corpus), Korpus Instytutu Podstaw Informatyki Polskiej Akademii Nauk (henceforth IPI PAN Corpus) and the PELCRA Corpus. The Russian ones are the following: the Russian National Corpus (henceforth the RNC), Comparable Corpus of English and Russian News Texts (henceforth CCERNT Corpus) and Computational Corpus of Russian Newspaper Texts at the End of the Twentieth Century (henceforth CCRNTE20 Corpus). Such a selection is not accidental since the aforementioned linguistic resources are the most representative corpora compiled for Polish and Russian, respectively.

2. Characteristics of Selected Polish Corpora

PWN Corpus is a synchronic and balanced corpus compiled by the commercial institution Wydawnictwo Naukowe PWN S. A., one of the largest publishing houses operating on the Polish market. The corpus in question is available on the CD-ROM, on the Internet [7] and at the headquarters of the compilers. The access differentiation has a bearing on the corpus sample which is available to researchers. The overall corpus available at the PWN S. A. headquarters consists of 100 million tokens, out of which 70 million tokens account for the balanced corpus, which includes modern Polish literature (41 %), press articles (45.5 %), dialogues, leaflets and manuals, Internet websites (13.5 %); the remaining part includes Polish literature and press archives. The opportunities provided to the user of the Internet are limited, however. When accessing the corpus on-line, one can choose either a web sample of the corpus (*wersja sieciowa*) or the demonstrative version (*wersja demonstracyjna*). The difference between the two is crucial since the web sample provides access to the corpus of 40 million tokens (out of which 22 million constitutes the balanced corpus), whereas the demonstrative version sports a corpus sample of 7.5 million whereby only 3.5 million of tokens is a balanced collection. Although both samples are available through a concordancer placed on the above website, the difference between the two is that the very access to the web sample is chargeable, whereas the demonstrative version is free-available. Moreover, the queries which may be executed in the demonstrative version disable to regulate the width of the left and right-hand context of the key-words subject to search. As for the annotation of the corpus data, it contains only meta-situational and meta-textual information in a format that conforms to the Text Encoding Initiative (TEI) guidelines. Although such annotation renders part-of-speech queries unavailable, the undeniable advantage of the concordancer is its ability to display all inflected forms of the key-word subject to search. Finally, the corpus sample available on the CD-ROM, which was distributed to all Institutes of Polish Studies at Polish universities free of charge is the same corpus sample as a demonstrative one available on the Internet.

The next corpus subject to this study is IPI PAN Corpus, which was developed at the Institute of Computer Science of the Polish Academy of Sciences. The full version of the corpus comprises almost 300 million segments available through the Poliqarp concordancer, which serves as an integral tool for browsing the corpus on-line or it can be downloaded as a compressed tar file. As for the subcorpora available free of charge from the http://korpus.pl website, three of them require further presentation. The source version (*próbka źródłowa*) of the IPI PAN Corpus comprises 100 million tokens which renders 286, 000 types. The preliminary sample (*próbka wstępna*) comprises 70 million segments which renders 364, 000 types. The sample of the corpus available on-line is composed of 15 million tokens which renders 217, 000 types and which further accounts for the opportunistic corpus whose 90 % fall into the category of modern Polish literature and socio-political journalism. All the above samples are downloadable from the corpus website in a form of tar archive files, which can be decompressed with the use of 7Zip freeware application. As far as the annotation is concerned, the corpus data are enriched with morphosyntactic tags whose common format is the slightly modified XML version of Corpus Encoding Initiative. As for the meta-textual and meta-situational information, it is still incomplete and subject to major improvements. The Poliqarp search engine and concordancer (available in three versions: on-line, graphical and GNU/Linux, where the latter one is the most advanced one) is equipped with a user-friendly interface, which enables researchers to request multifarious types of queries, which may contain standard regular expressions: the base-form query, grammatical class query (based on grammatical tags specifying the values of the part of speech), grammatical category query, query with

constraining matches to sentences or paragraphs as well as query with constraining matches to meta-situational or meta-textual information [5; 6].

The last Polish corpus subject to this description is the PELCRA Corpus, which is developed at the Institute of English at the University of Lodz in co-operation with the University of Lancaster. The Reference Corpus of Polish, as the major subcorpus within the framework of PELCRA project, comprises 93, 129, 588 tokens. Methodology used in the compilation of this corpus was very similar to the one adopted for the British National Corpus. It is a synchronic corpus of modern written and spoken Polish whereby the latter one comprises only 600, 000 tokens, which is the only major difference from the BNC. The target figure for spoken subcorpus, however, is 1 million tokens. As for the annotation of the corpus in question, its source form is the binary XML-annotated corpus text and its target form is a collection of data stored in the relational database MySQL. As a result, the meta-textual, meta-situational and meta-linguistic annotation that the corpus is equipped with is of hierarchical character and the corpus data are stored in the tables and columns of the MySQL database [3: 107]. Such a solution renders possible the extensive use of SQL (Simple Query Language) whose benefit is the opportunity to carry out multiple queries using the search tool available on the corpus website. The most important types of queries, which may again contain standard regular expressions, are the following: the base-form query, inflection query and phrase query, collocation query and the MI3 collocation query. The queries may be constrained by the meta-situational and meta-textual information, which enable researchers to specify the features of corpus samples subject to search. Moreover, the search tool offers a wide variety of statistical analyses concerning, in particular, the frequencies of words and collocations. However, the availability of the corpus for rank-and-file user is limited to executing a small number of queries and full access to the corpus data is chargeable. It has to be emphasized, however, that since the Reference Corpus of Polish described here is available on the World Wide Web, its interface is user-friendly and easy to manipulate.

3. Characteristics of Selected Russian Corpora

The description of commercially available corpora of Russian starts with the Russian National Corpus (*Национальный корпус русского языка*) which has been compiled at the V. Vinogradov Institute of Russian of Russian Academy of Sciences in Moscow. This corpus, which is available through the website http://ruscorpora.ru, includes 120 million tokens (recorded as of February 7, 2006) which make up the meta-situationally, meta-textually and meta-linguistically annotated representative collection of Russian texts in the electronic form. A particular emphasis shall be put on the meta-linguistic annotation, which is very detailed and comprehensive in that it comprises grammatical and semantic tagging. The former covers information concerning part of speech of the tokens, their case, gender, degrees of comparison, number, tense, aspect etc., whereas the semantic annotation is even more extended and covers semantic categories, taxonomy and axiology. As a result, the concordance search tool available on the website allows one to execute two types of queries: the 'exact word-forms query' (*Поиск точных форм*) and the 'lexico-grammatical query'(*Лексико-грамматический поиск*). The latter one enables researchers to constrain the queries to specified grammatical properties (which facilitates detailed searches by determining specific morphological properties of key-words subject to search) and semantic properties (by determining semantic properties of key-words) which may contain standard regular expressions. Moreover, the search tool offers an additional option which enables one to constrain the search to a specific genre or linguistic environments a given text functions in. This corpus has been equipped with the option of 'reduced homonymy' (*снятие омонимии*) or 'non-reduced homonymy' which allows one to display either the lemmatised or non-lemmatised results of the query. The former one reduces tokens (lemmatas) to specific types (lemmas) representing the same part of speech which enables to make a distinction between inflectional forms of the same lemma. The latter one allows inaccuracy in the case of multifunctional words, such as a Russian lemma *печь*, which may function either as a verb or a noun. Moreover, the sample of the corpus with reduced homonymy allows for a display of the concordances with marked accents, which is essential for many language teachers browsing the corpus for linguistic material to be used during their classes. Apart from that, one has access to the subcorpus of spoken Russian and to the parallel Russian-English and English-Russian subcorpora, invaluable collections of texts for translators and lexicographers comprising aligned text-units (sentences and paragraphs). These separate corpora with non-reduced homonymy are equipped with concordance search tool, which allows researchers to execute lexico-grammatical search therein.

The Comparable Corpus of English and Russian News Texts was compiled at the University of Leeds under the supervision of Sergei Sharoff. It consists of multiple subcorpora which are available on the website. Russian part consists of a morphologically-tagged (the grammatical properties include: part-of-speech, case, gender, animation, number, person, degree of comparison, aspect and tense) and lemmatised collection of articles from Izvestia daily newspaper (issued between 2000-2001), which add up to 14 million tokens. The search tools allow one to constrain the search to grammatical properties of the word-forms and to display rich statistical data on frequency and distribution of the key-words subject to search. The interactivity of the interface enables corpus-oriented researchers to execute searches in the Russian National Corpus (in addition to the Izvestia Corpus) and in the small corpus of modern Russian fiction (500, 000 tokens), of which the latter one is also compiled in Leeds. The search results are displayed as concordances (whose width can be extended) enriched with meta-textual and meta-situational information after double-clicking on the selected concordance. For the users who log in, having received the access code to the corpus, it is available free of charge.

The last Russian corpus subject to this description is Computational Corpus of Russian Newspaper Texts at the End of the Twentieth Century (*Компьютерный корпус текстов русских газет конца XX-ого века*). This synchronic and opportunistic corpus, which can be accessed through the http://www.philol.msu.ru/~lex/corpus website, is available either on-line (205, 000 tokens in 446 press articles) or it can be researched at the Institute of General and Computational Lexicography and Lexicology at Moscow State University, where it was compiled. The full version of the corpus comprises 11, 401, 479 tokens in 23, 110 press articles. Although the size of the corpus is limited, it contains meta-linguistic (morphosyntactic), meta-situational and meta-textual annotation. This collection of texts is also lemmatized and comprises samples of full issues of thirteen Russian newspapers dating from 1994 -- 1997. The search tool offers users two types of queries which may be constrained to the selected properties of annotation: the 'exact word-form query' (*Буквальное совпадение*) and the 'indirect query' (*Подстрока*), the latter allowing researchers to search for a sequence of characters which may either account for the full key-word or may be its integral part. The results of the queries are displayed as either concordances or lists of word-forms.

4. Conclusions

Having been familiarised with general characteristics of the Polish and Russian corpora and having laid down four criteria for above description, it is possible to arrive at the following observations.

As far as the overall number of tokens is concerned, three Polish corpora oscillate around the figure of 100 million, which may be the result of treating the British National Corpus as a reference point (the BNC has 100, 106, 008 tokens). Although IPI PAN Corpus is outstanding in that it includes 300 million tokens, its actual availability starts with a source sample which includes 100 million tokens. Moreover, the sheer number is not conclusive in that all Polish corpora have not been fully completed so far. What is crucial, however, are the issues concerning access to them for researchers working outside Poland and the question: which corpus to use if one wants to avoid charges and to access it on-line? In this respect IPI PAN Corpus comes to the fore as its demonstrative version available on-line includes 15 million tokens. Although PELCRA Corpus offers over 93 million tokens, the search is in practice limited to one query at a time. Another advantage of IPI PAN Corpus is that one can download all three samples (cf. the source, preliminary and on-line versions) together with Poliqarp search tool designed specifically to access the corpus. As a result, after decompressing the samples stored in tar files, one may execute searches directly on the hard drive. As for the Russian corpora, the biggest one is obviously the RNC, which comprises 120 million tokens available on-line and free of charge, which makes it a promising corpus for the size of data and access. If one conducts quantitative research, the Comparable Corpus of English and Russian News Texts compiled at Leeds makes up a good reference point as it provides extended statistical information about the key-words subject to search. The last Russian corpus, the Computational Corpus of Russian Newspaper Texts at the End of the Twentieth Century, is advantageous in that it is a specialist collection of texts but its overall size, which is relatively small (over 11 million tokens and only 200, 500 tokens available on-line), and opportunistic nature make it custom-designed for corpus-oriented researchers studying the language of modern Russian press.

As for the meta-linguistic annotation, the IPI PAN Corpus and PELCRA Corpus are the best annotated Polish corpora because they are equipped with all three types of annotation (meta-

situational, meta-textual and meta-linguistic) which extends the scope of executable queries. It is perfectly visible on the example of PELCRA Corpus, whose user-friendly interface in Polish and English allows one to execute multiple types of queries (as referred to above), which proves that designers and compilers of this corpus aptly took advantage of its rich annotation. Among Russian corpora, the RNC is the most promising one as its detailed meta-linguistic annotation includes morphological, semantic and axiological features, which widen the scope of executable queries. As a result, one may constrain any lexico-grammatical query to all grammatical and semantic properties allowing narrow and specified search to be executed. The two remaining corpora of Russian also use three major types of annotation, but the meta-linguistic one is limited to morphological properties.

Finally, access to both Polish and Russian corpora renders an interesting observation in that nearly all of them rely on the on-line access which is equivalent to browsing the corpus by means of a search engine, the fashion similar to looking for information on the Internet using such search engines as Google, Altavista, etc. In other words, the idea of developing a unique dedicated corpus client as a separate software application designed to be installed on any computer has not found much appeal for the analysed corpora. The exception to this rule is IPI PAN Corpus of Polish, which allows one to download Poliqarp, a dedicated search engine and concordancer to access and browse the corpus. This solution reflects the one adopted by the compilers of the British National Corpus, which can be accessed either on-line or from the hard drive by means of SARA client, a search engine and concordancer offering, among others, multiple options and queries to be executed (even building up a single query from multiple types of queries), printing the results straight away, manipulating the data and their format, etc.

The only corpus available on the CD-ROM is PWN Corpus of Polish. This is another interesting observation since the process of distributing corpora on CD-ROMs is both profitable for their compilers and convenient for armchair researchers, because it enables them to enjoy fast and efficient access to electronic collections of linguistic data, even if they do not have access to the Internet.

Summing up, the present comparison of selected Polish and Russian corpora shows that corpora vary and even using specific criteria does not guarantee that one arrives at objective and comprehensive results. Nevertheless, having been familiarised with above collections of texts, researchers interested in extensive linguistic analyses of Polish and Russian using corpus methodology are indubitably in a better position to choose the corpus which would meet their requirements and would correspond with goals of their research.

References

1. Kilgarriff, A. (2001). “Comparing Corpora”. Retrieved on 10 Jan. 2006 from: http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/publications.html
2. Mason, O. (2000). Programming for Corpus Linguistics. Edinburgh: University Press.
3. Pęzik, P., Uzar, R., Levin, E. (2005). Zastosowania baz danych w językoznawstwie. In: B. Lewandowska-Tomaszczyk, Podstawy językoznawstwa korpusowego (p. 95-115). Łódź: Wydawnictwo Uniwersytetu Łódzkiego.
4. Piotrowski, T. (2003). Językoznawstwo korpusowe: wprowadzenie do problematyki. In S. Gajda (Ed.), Językoznawstwo w Polsce. Stan i perspektywy (p.143-154). Opole: Wydawnictwo Uniwersytetu Opolskiego.
5. Przepiórkowski, A. (2004). “The IPI PAN Corpus in Numbers”. Retrieved on 11 Feb. 2006 from: http://nlp.ipipan.waw.pl/~adamp/Papers/2005-ltc-numbers/
6. Przepiórkowski, A. (2005). “The Potential of IPI PAN Corpus”. Retrieved on 11 Feb. 2006 from: http://nlp.ipipan.waw.pl/~adamp/Papers/2005-psicl-numbers/
7. Korpus Języka Polskiego Wydawnictwa Naukowego PWN. Retrieved on 12 Feb. 2006 from: http://www.korpus.pwn.pl
8. Korpus Języka Polskiego IPI PAN. Retrieved on 12 Feb. 2006 from: http://korpus.pl
9. The PELCRA Reference Corpus of Polish. Retrieved on 12 Feb. 2006 from: http://korpus.ia.uni.lodz.pl
10. Национальный корпус русского языка. Retrieved on Feb. 14 2006 from: http://ruscorpora.ru
11. Компьютерный корпус текстов русских газет конца XX-ого века. Retrieved on 14 Feb. 2006 from: http://www.philol.msu.ru/~lex/corpus
12. The Comparable Corpus of English and Russian News Texts. Retrieved on Feb. 14 2006 from: http://corpus.leeds.ac.uk

Н. Дарчук, к. філол. н., ***В. Сорокін****
Київський національний університет ім. Т. Шевченка (Київ)
УДК 161.2. 81'33

КОРПУС ТЕКСТІВ ЯК ДЖЕРЕЛО ДЛЯ МОВОЗНАВЧИХ І ЛІТЕРАТУРОЗНАВЧИХ ДОСЛІДЖЕНЬ

Стаття присвячена аналізу шляхів формування корпусів текстів різних стилів (поетичного, публіцистичного, наукового), вказується на засади створення параметризованих баз даних, характеризується програмне забезпечення таких одиниць.

Стан сучасної лінгвістики останніх десятиліть Ю. Д. Апресян [1] визначає красивою метафорою «золотий вік лексикографії» через безпрецедентний ріст числа й різноманітності словників, удосконалення методики лексикографування, принципи системного опису лексики і навіть переорієнтацію значної частини сучасної теоретичної лінгвістики на опис окремих лексем. Такий прорив у мікросвіт значення одного слова спонукає філологів до всебічного лінгвістичного «портретування», а значить інтегрального опису мови, що аж ніяк неможливо без досягнень комп'ютерної лінгвістики в цілому і корпусної зокрема.

Справді, фонетичні, просодичні, морфологічні, синтаксичні, семантичні, прагматичні, стилістичні, комунікативні, сполучуванісні властивості лексем є вивідними безпосередньо з текстів, тому стають невичерпним джерелом для словників і граматик.

Незаперечним є й той факт, що в останні десятиліття фахівціями галузей, що мають справу з комп'ютерним аналізом текстів, гостро відчувається потреба в точних функціональних характеристиках мовних одиниць у різних типах текстів. Для проведення теоретичних і прикладних (машинний переклад, аналіз і синтез мовлення, анотування і реферування тексту тощо) та дидактичних досліджень спеціалісти в галузі когнітивної лінгвістики, стилістики, семасіології, функціональної граматики, словотвору не мають узагальнених даних про закономірності функціонування мовних одиниць у мовленні. Отже, необхідність розбудови корпусної лінгвістики – ознака нашого часу, завдяки їй спеціалісти одержать усю необхідну лінгвістичну інформацію для подальшого її опрацювання у філологічних студіях.

На нашу думку, існує кілька шляхів розвитку корпусної лінгвістики, але спільним і неодмінним для них є накопичення текстів різних стилів і часових зрізів з відповідним зовнішнім (джерело, автор, рік видання тощо) і внутрішнім, пов'язаним зі структурацією тексту, його маркуванням (номер розділу, абзацу, речення, слова, початок і кінець тексту, абзацу, речення тощо).

Перший шлях полягає у програмному забезпеченні пошуку окремої словоформи у певному джерелі або всьому корпусі текстів та формуванні дослідницької ілюстративної текстової бази. Така система може функціонувати без використання лінгвістичних алгоритмів і словників або з використанням словників як допоміжного матеріалу для формування запитів, хоча сам корпус лінгвістично не розмічається.

Другий шлях пов'язаний з розробкою програмного забезпечення для самостійного маркування корпусу текстів дослідником. За допомогою спеціально розроблених програм та інтерфейсу користувач-дослідник може «портретувати» слово: визначати його морфологічну, синтаксичну, лексичну характеристику згідно з контекстом уживання. Цей шлях забезпечує можливість, з одного боку, одержання однозначної, точної у лінгвістичному смислі інформації, з іншого – створення лексикографічної картки за результатами розмітки.

Третій шлях передбачає наявність комп'ютерних інструментів – пакету програм для різнобічної лінгвістичної обробки текстового матеріалу. В цьому випадку корпус текстів лінгвістично розмічається в автоматичному режимі з подальшою автоматизованою обробкою для зняття омонімії. Завдяки цьому користувач одержує інформацію, передбачену програмним забезпеченням та базами даних (словниками з морфологічною, синтаксичною, лексичною, енциклопедичною інформацією), створює за текстами свої власні алфавітні й частотні словники, робить автоматично транслітерацію, фонематичний або фонетичний запис, сегментує текст на морфи, морфологічно індексує текст/слово, будує синтаксичні дерева залежностей речення тощо.

* *© Н. Дарчук., В. Сорокін, 2006*

Кожен з цих шляхів має свої позитивні риси і недоліки. До **позитивних** рис першого можна віднести швидкість створення текстових корпусів; другого – точність характеристики одиниць аналізу та при узгодженій роботі спільними зусиллями фахівців – створення сучасної автоматизованої лексикографічної картотеки; третього – ефективність і швидкість в одержанні різноманітної інформації. До **недоліків** – ресурсну обмеженість першого, повільність роботи другого та недостатню точність третього, тому що 100% бездоганної обробки тексту автоматично досягти практично неможливо.

Займаючись проблемами корпусної лінгвістики, викладачі Інституту філології Київського національного університету ім. Т. Шевченка обрали третій шлях, розробивши відповідну концепцію, суть якої полягає у параметризації текстових корпусів за якісними і кількісними ознаками, в розробленні методичних засад створення комп'ютерних інструментів для лінгвістів, а також пакетів програм для укладання електронних картотек, частотних словників, тезаурусів, словників морфемно-словотворчих гнізд. Це розвиває проект корпусного опрацювання вшир та вглиб.

Ми не будемо спинятися на проблемах і вимогах до корпусу текстів з точки зору користувача, розуміючи, що в його побудові є своя стратегія, свої особливості: репрезентативність, повнота, структуризація матеріалу. Дійсно, як показує лексикографічна практика, правильно сформований текстовий масив – запорука якісних мовознавчих досліджень, текстова база – джерело створення словників і полігон лінгвістичних досліджень функціонування мовних одиниць усіх рівнів. Тому підбір текстового матеріалу, організація його за стилями, зонами, хронологією є обов'язковою умовою організації корпусної системи.

Мета нашого проекту полягає у комплексній та аспектній розробці корпусу текстів таких стилів:

- поетичного (300 тис. слововживань);
- публіцистичного (300 + 300 тис. слововживань);
- наукового (тексти з екології – 300 тис. , біології – 150 тис., фізики – 150 тис., мовознавства – 500 тис., літературознавства – 150 тис., філософії – 150 тис., юридичної лексики – 150 тис.).

Корпуси текстів є **дослідницькими**, оскільки призначені для вивчення різних аспектів функціонування мовної системи і зорієнтовані на широкий клас лінгвістичних задач; **статичними**, оскільки відображають певний часовий стан мовної системи. Наприклад, авторські корпуси поетичної мови – це колекції текстів окремих поетів кінця XX століття, публицистичні тексти – репрезентують мовлення газет за короткий проміжок часу (2004 рік). Водночас ці тексти дають можливість дослідити мовні явища в динаміці, тому вони є **динамічними**, а також інтерактивними, тому що користувач у проведенні дослідження може виділити з генерального корпусу робочий корпус, який включає лише частину генерального корпусу, напр., тексти якоїсь зони (Політика, Суспільство, Економіка, Спорт) або якоїсь газети «Дзеркало тижня», «Україна молода», «Українська правда».;

Невеликий за всіма мірками корпус текстів (більше 2 млн. словоформ) побудований таким чином, щоб користувач міг одержати різноаспектну інформацію про мовну одиницю (морфема, слово, словосполучення, речення):

- зовнішню – анотування тексту (автор, джерело, рік видання тощо);
- внутрішню – структурація тексту (номер розділу, абзацу, речення, слова тощо);
- граматичну, лексико-граматичну, лексичну;
- кількісну – про функціонування її як інваріантної форми (лема, морфема) і варіантної (словоформа, морф) зі статистичними характеристиками (абсолютна та середня частота, міра коливання середньої частоти, коефіцієнт стабільності).

Така сукупність характеристик дозволила створити **параметризовану базу даних,** під якою укладачі розуміють багатоаспектну і багатофункціональну систему, яка включає:

1. корпус текстів – джерело різного роду словників;
2. серію алфавітно-частотних словників з усією інформацією про слово: граматичною, лексико-граматичною, стилістичною та статистичною (абсолютна та середня частота, міра коливання середньої частоти, коефіцієнт стабільності);
3. алфавітно-частотні словники слів та слововживань спільної лексики.
4. словники неолексем;
5. словники синтаксичних моделей керування: дієслівних, іменникових, атрибутивних;

6. серію морфемних та словотвірних словників з частотними характеристиками морфа/морфеми, за якими можна вивчати комбінаторно-дистрибутивну будову, словотвірне значення кожної афіксальної морфеми в текстах;
7. словники тропів: епітетів, метафор, метонімії, порівнянь, синекдохи, оксюморонів, гіпербол;
8. словники синонімів, антонімів, фразеологізмів, тезауруси;
9. словник-конкорданс як допоміжний інструмент для формування лексико-семантичної, синтаксичної та стилістичної характеристики слова.

Ми розглядаємо параметризовану базу як реальну можливість для створення універсального (багатоцільового) словника. Але цілком закономірним постає питання: що є параметром у структурі бази? У зв'язку з тим, що термін цей неоднозначний і досить розповсюджений, ми визначаємо його як особливе словникове представлення мовних характристик одиниці. Ю. М. Караулов визначає цей термін так: «під параметром розуміємо деякий квант інформації про мовну структуру, який ...може... виступати у сполученні з іншими квантами (параметрами) і знаходити специфічне вираження у словниках» [2].

У базі передбачено параметри:

- граматичні (частина мови і категоріальні значення, напр., рід, число, відмінок, особа тощо);
- структурні (напр., моделі морфної структури слів різних частин мови; моделі керування дієслівні, іменні, атрибутивні тощо);
- лексико-семантичні, які віддзеркалюють системні відношення (синонімія, антонімія, омонімія, паронімія, ідеографія, тропіка);
- статистичні.

З цього неповного переліку видно, що:

1) деякі параметри є синкретичними (напр., частиномовний пов'язаний з категоріальним або структурним);

2) параметр завжди відноситься до лексеми/словоформи в цілому;

3) параметр може описувати окрему одиницю (напр., структуру або граматичну категорію);

4) параметр взагалі може відображати кількісні (статистичні) характеристики мовних одиниць.

5) параметри представляють багатовимірну класифікацію .

Комп'ютерна підтримка функціонування параметризованої бази забезпечується пакетом алгоритмів і програм автоматичного морфологічного і синтаксичного аналізу української мови, морфного сегментування тексту, укладання повних і часткових словників за частотою, алфавітом, а також пакетом статистичної обробки лінгвістичних даних. Створення алфавітного словника здійснюється в двох режимах: автоматичному (без зняття омонімії) і автоматизованому (зі зняттям лексичної і лексико-граматичної омонімії в деяких випадках ця функція також передбачена пакетом програм). Для виконання таких завдань, як створення словників неолексем, синонімічних груп, тезауруса, поетичних тропів та ін. передбачене програмне забезпечення в автоматизованому режимі, оскільки їх здійснення потребує контроль за значенням. При цьому обов'язковою є функція конкордансної ілюстрації з корпусу текстів. Компонентами програмного забезпечення є такі словники української мови: орфографічний, граматичний, морфемний, синонімічний, антонімів, фразеологічний, тлумачний, кожен з яких має комп'ютерне представлення, в розробці якого брали участь автори цього проекту. Програмно передбачене автоматичне порівняння авторських текстових словників зі словниками мови скорочує роботу лінгвіста у десятки разів.

Отже,

- Потреба в точних функціональних характеристиках мовних одиниць у різних стилях та жанрах гостро відчувалася в останні десятиліття фахівцями ряду галузей, що мають справу з комп'ютерним аналізом текстів, з укладанням навчальних словників, з теорією та практикою функціонування мови. Пропонована параметризована база даних є одним із перших кроків у заповненні цієї лакуни.
- Систематизована у серіях словників і списках інформація є важливою для граматичних, стилістичних, літературознавчих і семантичних досліджень лексичного фонду української мови в його статиці та динаміці.

- Запропонована методика опрацювання лінгвістичних даних в електронній базі є узагальненням комплексу теоретичних і прикладних ідей сучасного мовознавства. Технологія конструювання бази робить її надзвичайно ефективним та раціональним інструментом (вона зберігає багато часу та людських ресурсів) для спеціалістів-філологів різного профілю.
- Електронна база даних зі своєю методологією і технологією допомагає ефективно й оперативно здійснювати масштабні комплексні філологічні дослідження на рівні сучасної наукометрії.
- Ця база даних використовується у навчальному процесі на філологічних факультетах університетів, забезпечуючи необхідний теоретичний і технологічний рівень підготовки фахівців.
- Створена нами база даних починає використовуватися як готовий продукт і як модель для конструювання аналогічних баз університетами України.

Література

1. Апресян Ю. Д. О толковом словаре управления и сочетаемости русского глагола// Словарь. Грамматика. Текст. – М., – 1998. С.13.
2. Караулов Ю. Н. Лингвистическое конструирование и тезаурус литературного языка. М., 1981. С.51.

О. Зубань *
Київський національний університет ім. Т. Шевченка (Київ)
УДК 81’33

ПАРАМЕТРИЗОВАНА БАЗА ДАНИХ ЯК ІНСТРУМЕНТ ДОСЛІДЖЕННЯ КОРПУСУ ТЕКСТІВ

Розвиток теорії і практики прикладної лінгвістики, зокрема методів комп'ютерного моделювання, дозволили по-новому сформулювати лексикографічне завдання, а саме – створення автоматизованої системи лінгвістичного аналізу тексту. Структура цієї системи базується на модульно-словниковій ідеології: системно-структурні відношення одиниць кожного рівня мовної системи представлені в окремому модулі: фонетичному, морфемному, словотвірному, морфологічному, синтаксичному, семантичному.

У традиційній лінгвістиці наукова значущість, вагомість результатів дослідження визначається перш за все репрезентабельністю матеріалу дослідження: чим більше мовних фактів, тим достовірніші спостережувані закономірності, але традиційна форма збирання і систематизації інформації (переважно паперова картотека та використання різноманітних паперових словників) сьогодні не задовольняє потреб дослідників-філологів. Потрібні нові інформаційні технології, які б оптимізували роботу дослідника. Тому в українському мовознавстві на сьогодні нагальною є проблема укладання електронних лінгвістичних словників, які мають формат параметризованих електронних баз даних, оснащених пошуково-класифікаційними програмними аналізаторами, що забезпечують:

- ефективне та оперативне проведення лінгвістичного аналізу;
- можливість аналізу великих лексичних масивів;
- отримання точних формальних характеристик мовних одиниць різних рівнів як у системі мови, так і реляційно-функціональних особливостей їх у тексті.

У сучасній комп’ютерній лінгвістиці досить плідно розвивається лексикографічний напрямок. Зокрема в українському мовознавстві сьогодні створено чимало різногалузевих електронних словників, з різноманітними системами навігацій та інтерфейсами, частина з яких стала вже лінгвістичним комерційним продуктом і доступна для широкого кола користувачів. Такі бази даних покликані виконувати функцію своєрідних довідників для лінгвіста-дослідника і, без сумніву, є надзвичайно важливими для організації повномасштабного дослідження мови, але вони є **статичними**, їх не можна використовувати в режимі автоматизованого аналізу тексту. Тому особливої уваги сьогодні заслуговують ті електронні лінгвістичні продукти, які спрямовані на аналіз тексту і мають статус **динамічних** пошукових систем, які здатні в автоматичному або автоматизованому режимі вилучати інформацію про мовні одиниці з будь-якого параметризованого тексту.

Текст як основна форма збереження і передачі інформації, будучи результатом мовленнєвого акту, є інвентарем мовних одиниць, які комбінуються в ньому за законами

* © *О. Зубань, 2006*

мовної норми. А оскільки в мові немає нічого, чого б не було в мовленні, то текст водночас і реалізує, і формує систему мови. Тобто, правомірним і перспективним є опосередкований шлях вивчення мовних явищ через текст, оскільки аналізуються всі одиниці мовних рівнів як цілісна реальність у спостереженні.

В останні десятиріччя фахівцями різних сфер, дослідження яких дотичні з комп'ютерним аналізом текстів, гостро відчувається необхідність у точних функціональних характеристиках мовних одиниць у різних типах текстів. Таку лінгвостатистичну інформацію можна отримати за допомогою електронної бази даних, досліджувальні функціональні можливості якої забезпечуються взаємопов'язаною роботою різних модулів – фонетичного, морфемного, словотвірного, морфологічного, синтаксичного, семантичного.

Працюючи над проблемами конструювання параметризованих лінгвістичних баз даних, викладачі Інституту філології Київського національного університету імені Тараса Шевченка розробили методику формалізованого опису мовних одиниць різних рівнів, на основі якої створили комп'ютерні інструменти – пакети програм, що забезпечують вилучення інформації із тексту і створення різноманітних словників: частотних, граматичних, морфних структур, кореневих, афіксальних, словників кореневих гнізд, синтаксичних структур словосполучення та речення, різнотипних електронних картотек, спрямованих у різні галузі лінгвістичних досліджень.

Розвиток теорії і практики прикладної лінгвістики, зокрема методів комп'ютерного моделювання, дозволили по-новому сформулювати лексикографічне завдання, а саме – *створення автоматизованої системи лінгвістичного аналізу тексту*.

Структура цієї системи базується на модульно-словниковій ідеології: системно-структурні відношення одиниць кожного рівня мовної системи представлені в окремому модулі: фонетичному, морфемному, словотвірному, морфологічному, синтаксичному, семантичному.

Робочою одиницею кожного модуля є слово (вихідна форма), що грунтується на лінгвістичному постулаті – слово – центральна одиниця мови, що на різних рівнях мовної системи характеризується різними типами структур, або ж як структурна одиниця входить до одиниць вищих мовних рівнів – словосполучення та речення; а також зумовлено текстовою спрямованістю створюваної автоматизованої системи аналізу, оскільки в графемному представленні будь-якого тексту визначається дискретна одиниця від пробілу до пробілу – слово. Отже, базовою одиницею кожного модуля виступає слово (вихідна форма) як формальна одиниця: задана послідовність графем, що через посередництво морфологічного модуля має постійний зв'язок із текстовою словоформою – послідовністю графем між пробілами тексту. У кожному модулі ця робоча одиниця формалізується через представлення її у термінах програмних процедур, що дозволяє описати різноманітні лінгвістичні структури слова (фонетичну, морфемну, словотвірну, морфологічну), словосполучення, речення. Формалізація опису цих структур здійснюється за допомогою єдиної метамови, що забезпечує зв'язок слова із структурами слів в усіх модулях, і зокрема з морфологічним модулем, який дозволяє ідентифікувати кожну словоформу тексту із її словниковою формою (вихідною) у кожному модулі.

Кожен модуль складається з двох блоків: 1) блок-словник: інвентар слів кожного модуля формується на основі 11-томного тлумачного словника української мови і в процесі роботи доповнюється новою лексикою (іншомовними словами та термінами різних галузей науки), загалом на сьогодні словникова база системи автоматизованого лінгвістичного аналізу складає ≈ 170 тис. слів української мови; 2) блок-аналізатор, що розглядається як надійний лінгвістичний інструмент мовознавчих досліджень, здатний працювати в режимах пошуку, класифікації, лінгвістичного аналізу на великих лексичних словникових та текстових масивах.

Інформаційний макет словника кожного модуля та його ідеологія визначалися специфікою електронного характеру та перспективою створення на їх основі автоматизованої системи лінгвістичного аналізу тексту, тому лінгвістичні аналізатори кожного модуля дозволяють в автоматизованому режимі приписувати слову формалізовану інформацію про його структуру залежно від рівня аналізу, а через ідентифікацію словоформи тексту із словом модуля отримати автоматичний аналіз словоформи.

Кожен модуль автоматизованої системи лінгвістичного аналізу спрямований на вирішення своїх завдань, що зумовлює структуру його словника, формалізацію запису структури слова та програмування аналізу.

Фонетичний модуль (на сьогодні розроблений лише проект модуля). У цьому модулі ставиться завдання представити графемний запис слів у двох типах структур: фонетичній та фонемній. Визначення цих структур забезпечується двома типами алгоритмів, оскільки процеси чергування та асиміляції чітко визначаються фонетичною позицією, що дозволяє їх алгоритмічно описати, наголос як необхідний знак транскрипції приписується лінгвістом. Одиницею словників (фонетичного та фонемного) у цьому модулі є слово у фонетичному та фонемному записі. Автоматизований фонетичний аналіз забезпечує аналізатор, який дозволяє отримувати інформацію про:

- фонемну будову слова/словоформи:
- кількість голосних: за характеристикою місця творення, ступеня підняття, лабіалізації, наголосу;
- кількість приголосних: за класифікаційними ознаками сонорна, шумна (дзвінка, глуха), губна, язикова (передньоязикова, середньоязикова, задньоязикова), глоткова, зімкнена, щілинна, африката, свистяча, шипляча, носова;
- фонемні моделі слів/словоформ, що дозволить розробити алгоритм складоподілу;
- складову структуру слова/словоформи: кількість складів (закритих, відкритих), їх моделі;
- коефіцієнт консонантності та вокалізму тексту.

Морфемний модуль. Словник цього модуля формують слова із визначеними морфними структурами. Формалізація опису морфних структур слів полягає у моделюванні структурно-функціональних зв'язків морфів у слові: представлення кожної морфної структури слова у вигляді лінгвістичної моделі, яка визначає межі і тип кожного морфа і дозволяє автоматично описати кожну морфну структуру через програмну процедуру, наприклад, *заледеніти* / P2R5I7S8F10. Автоматично сформована програмна процедура моделі морфної структури, пов’язана з формою вираження кожного слова морфемної бази даних, дає повну лінгвістичну інформацію про морф, його структурні відношення з іншими морфами і визначається як робоча одиниця морфемної бази даних.

Укладання морфемного словника та використання морфемного модуля в режимі лінгвістичного аналізу забезпечується зручним автоматизованим інструментом, який дозволяє доповнювати морфемну базу новою лексикою.

Формалізація морфних структур слів через їх опис у термінах програмних процедур МБД дозволила створити на основі цієї бази даних автоматизовану систему лінгвістичного аналізу, здатну виконувати цілий ряд завдань в автоматичному режимі:

- групувати лексику у спільнокореневі та спільноафіксальні класи;
- класифікувати лексику за кількісно-морфними моделями;
- створювати алфавітно-частотні кореневі та афіксальні словники;
- проводити морфемний аналіз вихідних словоформ;
- проводити морфемну сегментацію словоформ тексту.

Робота морфемного модуля із словоформами тексту забезпечується не тільки словником морфемного модуля, а й програмою автоматичного морфемного сегментування словоформ тексту, якою був доповнений модуль у процесі дослідження корпусу текстів поетичного стилю.

Формалізація опису морфних структур у морфемному модулі дозволяє швидко і ефективно створювати нові лексикографічні лінгвістичні описи, що спрямовані на вирішення різноманітних проблем у галузі морфеміки. Сьогодні ведеться робота над створенням словника морфем, що репрезентує морфему як знакову парадигматичну одиницю з приписаним квантом значення, реалізовану в морфах у визначених дистрибуціях.

Словотвірний модуль. Укладання електронного словотвірного словника здійснюється за допомогою аналізатора морфемного модуля системи, що застосовується для виконання двох завдань: 1) автоматизованого групування лексики у спільнокореневі вибірки; 2) автоматизованої класифікації лексики у межах кожного спільнокореневого класу за кількісно-морфними моделями.

Кожна вибірка спільнокореневої лексики формує у межах словотвірного модуля окреме поле, в якому морфні структури слів класифікуються за кількісно-морфними

моделями. За принципами словотвірної похідності було розроблено формалізовані методики опису словотвірних зв'язків між мотивуючим та мотивованим словами, що дозволяє автоматично побудувати робочу гіпотезу-модель словотвірного гнізда.

Класифікація слів спільнокореневої вибірки за словотвірними тактами є лише лінгвістичною гіпотезою, яка вимагає перевірки, що здійснюється в автоматизованому режимі: 1) перевірка правильності групування слів у спільнокореневі вибірки та розташування їх за словотвірними тактами; 2) визначення у кожному похідному слові словотвірної основи та словотвірного форманта; 3) додавання інформації про морфонологічні процеси, що відбуваються на кожному такті словотвору.

Словотвірний модуль дозволяє аналізувати текстові слововживання за такими параметрами:

- спосіб творення слова та його словотвірна структура;
- місце слова у словотвірному гнізді;
- класифікація лексики за ознаками способу творення, словотвірними значеннями та словотвірними категоріями,

а також:

- моделювати потенційно можливі слова української мови;
- визначати неологізми та оказіоналізми в тексті;
- доповнювати електронний словотвірний словник новою лексикою.

Морфологічний модуль. Словник цього модуля представляє вихідні форми слів, яким приписано код граматичної інформації: частина мови, нерегулярні граматичні значення (наприклад, для іменника – рід, для дієслова – вид), код парадигматичного класу, за допомогою якого словоформі приписуються регулярні граматичні значення (наприклад, для іменника – відмінок, для дієслова – час, особа).

Морфологічний модуль у системі автоматизованого аналізу тексту є стрижневим, тому що всі інші модулі працюють із текстовою словоформою через ідентифікацію її з граматичним кодом морфологічного модуля. Аналізатор морфологічного модуля в автоматизованому режимі (із зняттям омонімії) здійснює параметризацію корпусу текстів.

Морфологічний аналізатор дозволяє:

- проводити автоматичний морфологічний аналіз словоформ тексту;
- групувати текстові слововживання за класифікаційними ознаками:
  - частина мови;
  - у межах іменних частин мови: рід, число, відмінок, лексико-граматичний розряд;
  - у межах дієслова: вид, спосіб, час, перехідність/неперехідність, стан, особа, число / рід, число;
  - у межах прислівника: лексико-граматичні розряди;
  - у межах прийменника: участь у керуванні відмінками іменників;
  - у межах сполучника: розряди за функцією та значенням;
  - у межах частки: розряди за функцією;
  - у межах вигука: розряди за значенням.
- створювати словник-конкорданс визначених словоформ або словоформ визначеного граматичного значення.
- отримувати статистичні дані за різними параметрами граматичної інформації словоформ.

Синтаксичний модуль. Цей модуль не має свого електронного словника, тому його аналізатор має тільки текстову ідеологію і дозволяє описувати синтаксичні структури речень, використовуючи морфологічну базу даних та автоматизований інструмент побудови дерева залежностей, що репрезентує формальну синтаксичну структуру речення через моделювання зв'язків між словоформами та визначення їхнього виду: підрядний (керування, прилягання, узгодження), сурядний та координація.

Застосування цього модуля в автоматизованому аналізі тексту дозволяє:

- визначати кількісні параметри та структуру надфразових єдностей: кількість абзаців, простих та складних речень в абзаці; словосполучень в абзаці та реченні, довжину речення у слововживаннях;
- визначати функціональні параметри: структуру речення через модель дерева залежностей, характер підрядного зв'язку словосполучення (керування,

прилягання, узгодження), морфологічну модель словосполучення у межах кожного типу підрядного зв'язку;
- на базі визначених моделей укладати словники структур словосполучень та речень з текстовою ілюстрацією цих структур.

Семантичний модуль. Цей модуль будується на словниках двох типів:
- Електронного словника синонімічних рядів;
- Словника-тезауруса.

Електронний словник синонімічних рядів, який укладено на базі лексики "Словника синонімів української мови" (Т. 1 – 1999, Т. 2 – 2000), систематизує слова за їх лексичними значеннями і групує лексику у 12 тис. синонімічних рядів, що покривають 50 тис. слів української мови;

Семантичний аналізатор цього словника дозволяє використовувати його в процесі створення словника-тезауруса.

Словник-тезаурус систематизує лексику не за лексичним значенням, а за ідеографічним принципом представлення поняттєвої спорідненості лексики – від концепту (поняття) до лексичних засобів його вираження, що дозволяє групувати семантично близькі лексичні одиниці, які з певними змістовими варіаціями передають одне поняття, у лексико-семантичні групи. Ідеографічна методика вибрана не випадково, тому що саме така класифікація дозволяє:
- формалізувати опис структурних відношень лексичної системи мови, а саме найтиповіші семантичні зв'язки у лексиці – синонімічні, антонімічні та полісемічні, через визначення місця кожної лексико-семантичної групи, а отже, кожної лексеми, у межах ідеографічної класифікації;
- отримати формалізований опис семантики слова через суму поняттєвих концептів;

Лінгвістичний аналізатор цього модуля дозволяє в автоматизованому режимі заповнювати синоптичну карту словника із врахуванням тлумачення слів за 11-томним тлумачним словником.

В укладанні обох типів словників семантичного модуля використовуються словники морфемного і словотвірного модулів, які дозволяють глибше зрозуміти і правильно описати семантику слова.

У результаті роботи семантичного модуля лінгвіст може отримувати інформацію про:
- семантику слова за логіко-концептуальною схемою синоптичної карти;
- класифікувати слова/словоформи за лексико-семантичними групами;
- визначати синонімічні, антонімічні та полісемічні слова в тексті;
- визначати та описувати різноманітні переносні значення слів, тропіку художніх текстів.
- укладати словники поетичних образів.

Створені різногалузеві лінгвістичні бази даних були використані у дослідженні корпусу текстів, до якого ввійшли тексти поетичного, публіцистичного та наукового стилів обсягом майже 2 млн. словоформ.

Корпус текстів був об'єднаний із модульною системою лінгвістичного аналізу, в результаті чого створюється параметризована база даних, що структурується на:

1. **Модуль-текст**, який репрезентує корпус текстів;

2. Модуль-словник, в якому систематизується різнотипна лінгвістична інформація в результаті аналізу текстів;

3. **Модуль-аналізатор** – інструмент лінгвістичних досліджень великих текстових масивів, що включає пакети програм, які можуть виконувати такі функції:
- забезпечують зв'язок корпусу текстів з лінгвістичними базами даних: морфологічною, морфемно-словотвірною, синтаксичною, семантичною;
- забезпечують роботу в автоматичному режимі пошуку і класифікації лексики за різними параметрами, а також морфологічного, морфемного, статистичного, синтаксичного та семантичного аналізів;
- формують словник-конкорданс мінімальних контекстових слововживань.

У логіці конструювання параметризованої бази даних корпус текстів є джерелом автоматизованого укладання різноманітних словників (модуль-словник) за допомогою модуля-аналізатора. У результаті лінгвістичних досліджень за допомогою параметризованої лінгвістичної бази були укладені такі словники:

1. **Інтегрований алфавітно-частотний словник поетичного мовлення** обсягом 31.000 слів і 69.000 словоформ з інформацією про слово: граматичною, лексико-граматичною і статистичною (середня частота словоформи(x_{ser}), среднє квадратичне відхилення середньої частоти словоформи (sigma), середня частота лексеми($x_{serbases}$), среднє квадратичне відхилення середньої частоти лексеми (sigmabases), абсолютна частота і коефіцієнт стабільності);

Рисунок 1. Фрагмент словника для дієслова “брати”

Лек-сема	Слово-форма	Гр. інф.	x_{ser}	sigma	$x_{serbase}$	sigma-base	abs	Abs-base	de	De-base
Брати	Брати	Інф	0,02	0,14	0,33	0,6385661	6	99	5	8
Брати	Беру	1 ос.т.	5,333333E-02	0,2654974	0,33	0,6385661	16	99	7	8
Брати	Береш	2 ос.т.	2,333333E-02	0,1716262	0,33	0,6385661	7	99	5	8
Брати	Бере	3 ос.т.	0,07	0,2800595	0,33	0,6385661	21	99	7	8
Брати	Берем	1ос.мн.т.	6,666667E-03	8,137704E-02	0,33	0,6385661	2	99	2	8
Брати	Беремо	1ос.мн.т.	3,333333E-03	5,763872E-02	0,33	0,6385661	1	99	0	8
Брати	Беруть	3ос.мн.т.	0,03	0,2215852	0,33	0,6385661	9	99	5	8
Брати	Брав	Ч.р. мин.	2,333333E-02	0,1716262	0,33	0,6385661	7	99	5	8
Брати	Брала	Ж.р.мин.	0,01	9,949874E-02	0,33	0,6385661	3	99	4	8
Брати	Брали	Мн.мин.	0,02	0,14	0,33	0,6385661	6	99	5	8
Брати	Бери	Нак.одн.	4,666667E-02	0,2404625	0,33	0,6385661	14	99	7	8
Брати	Беріть	Нак.мн	2,333333E-02	0,1509599	0,33	0,6385661	7	99	6	8

2. **Словник поета,** в якому лексика, параметризована і систематизована в таких словниках: *Алфавітно-частотний словник слів і слововживань*; *Морфемні словники різного лінгвістичного спрямування*; *Лексичні словники різного лінгвістичного спрямування;Словники тропів.*

Ці словники формують: окрему базу даних для кожного з 15 поетів поетичного стилю корпусу текстів (вибірка 20 тис. сововживань, загальний обсяг корпусу поетичних текстів 300.000 слововживань), а також зіставно-порівняльну базу спільної лексики поетів (105 списків для кожної пари поетів),

Укладання алфавітно-частотного словника слів і слововживань та морфемних словників із частотними характеристиками для кожного морфа (морфеми для коренів) забезпечується пакетом алгоритмів і програм автоматичного морфемного і морфологічного аналізу, а також пакетом программ автоматичної статистичної обробки лінгвістичних даних. Укладання алфавітно-частотного словника здійснюється у двох режимах: а) автоматичному – без зняття омонімії; б) автоматизованому – за умови зняття лексичної, а в деяких випадках лексико-граматичної омонімії.

Алфавітно-частотний словник представляє точні характеристики частоти, стабільності вживання слова/словоформи, його лексико-граматичні характеристики. Укладання за цими параметрами лексичних реєстрів, спільних для двох авторів, дозволяє здійснювати порівняльну характеристику лексики, а також досліджувати закономірності організації ідіостилю кожного поета, зокрема аналізувати в автоматизованому режимі такі групи слів, як неологізми, архаїзми, діалектизми – ця інформація систематизується у відповідних словниках.

Укладання морфемних словників здійснювалось після проведення автоматичного морфемного сегментування словоформ тексту.

У результаті роботи автоматичного морфемного сегментатора на базі корпусу текстів поетичного мовлення було укладено цілу низку морфемних словників, які характеризують ідіостиль поета на морфемному рівні і дозволяють провести зіставно-порівняльний аналіз морфних структур різних поетів, а також різних мовних стилів (рис. 2).

Рисунок 2. Фрагмент результатів автоматичного морфемного сегментування словоформ поетичного тексту Л. Костенко

id	cls	morfem	morfema	comm
378422	ПТ	R	за	за | RC | RC
378423	ЙИ	R	щит	щитом | RDFE | RDFF
378424	ЙИ	F	ом	щитом | RDFE | RDFF
378425	АИ	R	смарагд	смарагдових | RHSJFL | RHSJFL
378426	АИ	S	ов	смарагдових | RHSJFL | RHSJFL
378427	АИ	F	их	смарагдових | RHSJFL | RHSJFL
378428	ЙИ	R	ліс	лісів | RDFE | RDFF
378429	ЙИ	F	ів	лісів | RDFE | RDFF
378430	X	N	.	

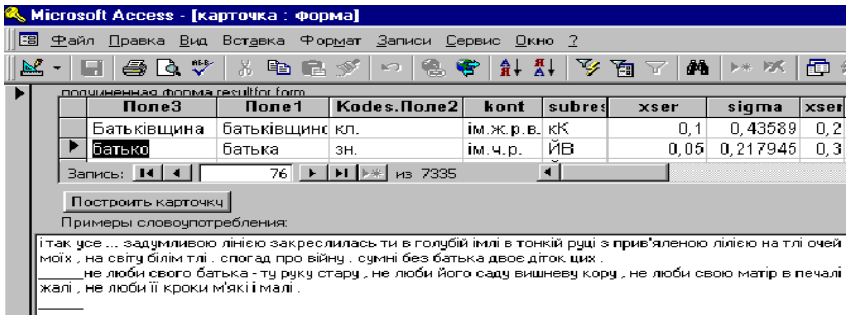
Словник лексики кожного поета через словник-конкорданс модуля-аналізатора має постійний зв'язок із корпусом текстів, що дозволяє автоматично ілюструвати текстове вживання кожної словоформи і будувати електронну картку (рис. 3).

Така картка дозволяє швидко та ефективно створювати різноманітні лексичні словники: *Словники синонімів, омонімів, неологізмів, діалектизмів, архаїзмів, фразеологізмів; Словники тропів: епітетів, метафор, метонімії, синекдох, порівнянь, оксюморонів, гіпербол.*

Ці словники укладаються в автоматизованому режимі, оскільки така інформація не може бути повністю формалізованою і вимагає участі лінгвіста. Наприклад, укладання словника неологізмів, словника синонімічних гнізд, ідеографічного словника супроводжується автоматичним морфемним та словотвірним аналізом, але вимагає тлумачення значення слова, введення його до відповідної лексико-семантичної групи синоптичної схеми або передбачає контроль за значенням слів при зіставленні їх із тією чи іншою тематичною групою електронного словника синонімічних гнізд.

Методика формалізованого лінгвістичного опису мовних одиниць у створенні системи автоматизованого лінгвістичного аналізу є узагальненням теоретичних і прикладних ідей сучасного мовознавства, що робить цю систему надзвичайно ефективним і раціональним інструментом лінгвістичних досліджень.

Рисунок 3. Фрагмент Частотного словника поетичного мовлення М. Вінграновського з текстовою ілюстрацією



Ця методика може бути використана як еталонна при укладанні різноманітних електронних словників, картотек, автоматизованій класифікації лінгвістичного матеріалу,побудові навчальних комп'ютерних тренажерів, тестів тощо. З іншого боку, отримана параметризована інформація про організацію українських текстів на різних рівнях мовної системи дає можливість вивчити закономірності функціонування системи мови в різних стилях, комплексно досліджувати мовні особливості ідіостилів українських поетів та письменників.

В. Старко, к. філол. н.*
Волинський державний університет імені Лесі Українки (Луцьк)
УДК 161.2.81'373.374.322

ФРАЗЕОЛОГІЯ, ЛЕКСИКОГРАФІЯ І КОРПУС

The article deals with the problems of multiaspect investigation and lexicographic description of phraseology based on corpus data. The author discusses the frequency of idioms and proverbs and the demand it places on the size and content of corpora and compares three different corpora with respect to their suitability for phraseology research.

Корпусні дослідження фразеології є відносно новою галуззю лінгвістики. Чи не найзначніший внесок в її розвиток внесла Розамунд Мун, чия робота [3] була першою

* © *В. Старко, 2006*

монографією на цю тему. І до сьогодні, за нашими даними, ця робота залишається найґрунтовнішим корпусним дослідженням фразеології. У своїй книзі авторка заторкнула низку важливих проблем, пов'язаних з дослідженням фразеології, що спирається на корпусні дані. Серед цих проблем варто виокремити такі: типологія фразеологічних одиниць (ФО), їхня частотність та розподіл частотності за типами ФО, вимоги до обсягу, наповнення й репрезентативності корпусу з огляду на специфіку ФО, варіативність ФО та її типи, дискурсивні функції ФО, прагматика й стилістика фразеології, роль ФО в забезпеченні зв'язності тексту тощо. Дальший розвиток ідеї Р. Мун набули у статті [4], де було розглянуто деякі проблеми корпусної фразеографії.

Предметом розгляду цієї статті є питання частотності ФО, зокрема прислів'їв, та обсягу корпусу, потрібного для повноцінного лексикографічного опису фразеології. Як матеріал для аналізу ми використаємо англійські фразеологізми. Наше дослідження є частиною проекту укладення «Англійсько-українського фразеологічного словника».[1]

Почнімо з деяких спостережень, що їх зробила Р. Мун у своїх роботах. Вона зауважує, що до 1990-го року обсяг наявних корпусів (до 20 млн. слововживань) був замалим для надійного дослідження фразеології [4]. Навіть її робота 1998 року [3] спирається лише на 18-мільйонний корпус англійської мови. Тому багато важливих висновків, за визнанням самої авторки, потребують підтвердження на ширшому матеріалі. Поява Британського національного корпусу (БНК, 100 млн. слововживань) та Банку англійської мови (БАМ, 323 млн. слововживань у 1997р., 450 млн. у 2005р., на 2006р. заплановано 530 млн.) означає, за словами Р. Мун, що «ідіоми й подібні до них вислови мають більшу статистичну ймовірність появи [в корпусі], однак цих даних все одно може забракнути, аби адекватно їх описати» [4: 265]. Однією з причин цього є те, що переважна більшість ФО мають набагато нижчу частотність, ніж окремі слова мови. Інша причина полягає в тому, що для опису ФО потрібно мати в наявності суттєво більше контекстів, ніж цього вимагає опис однослівної одиниці. Це зрозуміло, якщо взяти до уваги наявність різних типів фразеологізмів та, що важливіше, комплексну природу ФО, яка виявляється в їхній несистематичній й часто непередбачуваній варіативності, трансформаційному потенціалі, різноманітних дискурсивних й прагматичних функціях тощо. Тому 10 слововживань може бути достатньо, аби дати лексикографічний опис, наприклад, певному терміну, але замало для ФО, що має таку саму частотність вживання. Аби дати приблизну ідею про частотність ФО, використаймо влучний приклад з [4: 266]: англійська ідіома *looking for a needle in a haystack* має (з усіма варіантами) частотність 25 на кожні 100 млн. слововживань. Таку саму частотність мають слова *etiology, disequilibrium, granddaddy, hollowness, igneous, methylated, strychnine, timorous* та *ungentlemanly*.

Оскільки на теперішньому етапі проекту ми працюємо з англійськими прислів'ями, нас цікавить співвідношення частоти ФО взагалі і прислів'їв в англійській мові. Аби продемонструвати це співвідношення проаналізуємо дані, наведені в [3] – їх наведено в таблиці 1.

Таблиця 1. Співвідношення частотності ФО й прислів'їв в англійській мові.

Вживань на 100млн.	ФО	Прислів'я
< 6	8%	5%
6–22	32%	11%
22–100	32%	3%
100–200	12%	<0, 1%
200–500	9%	<0, 1%
500–1000	4%	0
1000–5000	3%	0
5000–10000	<1%	0
> 10000	<1%	0

[1] Проект був підтриманий ґрантом від Американської ради наукових товариств (American Council of Learned Societies) та стипендією ім. В. Фулбрайта від уряду США.

У першому стовпчику наведено кількість вживань на 100 млн. одиниць корпусу. Ці дані було отримано шляхом заокруглення даних, що їх дослідниця наводить для використовуваного нею 18-мільйонного корпусу під назвою «Оксфорський експериментальний корпус Гектор» (Oxford Hector Pilot Corpus). У другому стовпчику подано відсоток ФО (від загальної кількості ФО, проаналізованих в роботі Р. Мун, що складає 6776 одиниць), які належать до кожної смуги частотності. В третьому стовпчику в такий самий спосіб наведено відсоток тих ФО, що є прислів'ями (знову ж таки, від загальної кількості ФО). Як бачимо, через невеликий обсяг корпусу частотність вживання прислів'їв фактично перебуває на межі випадку. Тим не менше навіть за таких умов можна простежити тенденцію вживання ФО й прислів'їв.

Слід завважити, що корпус «Гектор» не є збалансованим. Майже на 60% він складається з газетних текстів, приблизно на 11% з художньої літератури і на 18% зі спеціальної (нехудожньої) літератури. Його незбалансованість має і негативні, і певні позитивні наслідки. До негативних слід зарахувати неможливість дійти остаточних висновків щодо відносної частотності вживання мовних одиниць в різних жанрах й їхніх відмінних характеристик у кожному окремому жанрі. Одним з позитивних наслідків є те, що переважання публіцистики в корпусі, як це не парадоксально, сприяє дослідженню ФО, адже, як показують наукові розвідки, зокрема й [3], [4], з усіх жанрів прислів'я і чимало інших типів ФО найчастіше вживають саме в публіцистичному жанрі. У збалансованому корпусі частотність ФО буде нижчою. Ці міркування приводять нас до логічного висновку, що для ґрунтовного обстеження й лексикографічного опису прислів'їв досліднику необхідно мати корпус максимального розміру, щонайменше на порядок більший, ніж для дослідження інших ФО. Якщо не стоїть завдання порівняльного міжжанрового аналізу ФО, то бажаною є щонайвища частка публіцистичних текстів.

Працюючи над лексикографічним описом англійських прислів'їв, ми маємо змогу користуватися саме таким корпусом. Йдеться про корпус газетних та журнальних публікацій англійською мовою під назвою «ЛексисНексис Академік» (LexisNexis Academic), що до нього надає доступ компанія «Рід Елсевір Інк.»[1] [2]. Нас цікавитиме кілька питань: 1) обсяг цього корпусу; 2) його співвідношення з обсягом БАМ-у; 3) розподіл частоти вживання ФО в корпусі за часовими періодами; 4) достатність/недостатність корпусу для повноцінного дослідження англійських ФО, зокрема прислів'їв.

Обсяг корпусу «ЛексисНексис Академік» точно встановити неможливо з тієї причини, що він щодня поповнюється новими текстами. Однак приблизне уявлення сформувати все ж таки можна, і в цьому нам допоможе порівняння його з БАМ-ом. Ми скористалися даними Р. Мун, наведеними в [4], щодо частоти вживання семи ідіом в БАМ-і 1997р. (323 млн. слововживань) і порівняли їх з вживаністю цих самих ідіом в «ЛексисНексис» станом на 27.02.2006р. Результати порівняння подано в таблиці 2.

У першому стовпчику наведено ідіоми в тій формі, що її було задано в пошуковій команді в роботі з БАМ-ом, а в другому – кількість їх вживань (абсолютні дані). Запис «needle w/5 haystack» означає, що було задано пошук лемми «haystack» в межах п'яти позицій від лемми «needle» (це ключові елементи ідіоми *looking for a needle in a haystack* та її варіантів), дужки позначають факультативні елементи, скісна риска – альтернативні елементи. В наступних чотирьох стовпчиках наведено кількість вживань в абсолютному вимірі в корпусі «ЛексисНексис» за часовими періодами. П'ятий стовпчик подає сумарну кількість, а шостий – відношення даних в другому стовпчику до даних в п'ятому.

Таблиця 2. Порівняння БАМ-у й корпусу «ЛексисНексис».

Ідіома	**БАМ**	**«ЛексисНексис» станом на 27.02.2006р.**					**Про-порція**
		1976–89	**1990–94**	**1995–99**	**2000–06**	**Всього**	
give smb.	64	141	288	443	792	1664	1:26

[1] Ми висловлюємо вдячність керівництву Нью-Йоркського державного університету в Олбані (США) за люб'язний дозвіл користуватися доступом до згаданого корпусу через університетську бібліотеку.

the cold shoulder							
(let) the cat out of the bag	91	302	525	1003	1503	3333	1:36
needle w/5 haystack	85	250	538	1144	1874	3806	1:45
beat about/around the bush	109	271	554	865	1407	3097	1:28
bury the hatchet	110	506	977	1620	2115	5218	1:47
bite the dust	145	636	1537	2573	3220	7088	1:49
(skate) on thin ice	241	400	890	1500	2320	5110	1:21

Ми обмежили пошук в «ЛексисНексис» лише газетними публікаціями, залишивши поза увагою журнальні статті. Підрахунок кількості вживань перших двох ідіом було проведено вручну, причому ми з'ясували, що через різні чинники (головно через повторення тих самих статей та джерела, що їх ми ігнорували, наприклад, англійськомовні газети Японії чи Китаю) нам довелося відкинули 12% вживань ідіом, переважно в посиланнях від 1990-х років і пізніше. Для решти ідіом ми використали формулу: загальна кількість вживань (що її подає пошукова система) х 0, 87. У такий спосіб ми відкинули 13% вживань за всі роки й отримані приблизні дані подали в таблиці.

Як бачимо, співвідношення між частотою вживання вибраних ідіом у БАМ-і й корпусі «ЛексисНексис» коливається в межах від 1:21 до 1:49. Якщо припустити, що «ЛексисНексис» лише втридцятеро більший за БАМ, то його обсяг має становити понад 9 млрд. слововживань. Такий обсяг більш ніж достатній для дослідження й лексикографічного опрацювання ідіом. БАМ має зрости в 1, 6 рази проти 1997 року, але й в такому обсязі він значно поступатиметься корпусу «ЛексисНексис», а через деякий час, цілком ймовірно, ще більше відстане від нього. Тисячі прикладів вживання ФО, що їх дослідник дістає в розпорядження, дають змогу, як ми переконалися на власному досвіді, відтворити їхній повний «лексикографічний портрет» і є унікальним джерелом багатоаспектного дослідження англійської фразеології. З іншого боку, менше 100 контекстів часто не вистачає, аби простежити тенденції вживання, і в цьому полягає обмеженість БАМ-у. «ЛексисНексис» містить тексти не тільки американських і британських газет, а й канадських, австралійських і новозеландських, що уможливлює докладну перевірку географічного поширення тієї чи іншої одиниці англійської мови. З таблиці видно, що кількість вживань ідіом зростає майже в геометричній прогресії від 1976 до 2006 року, а в проміжку 2000–2006рр. складає замалим 50% від загальної. Це створює сприятливі умови для дослідження найновітнішого зрізу мови.

Заради порівняння наведімо дані з роботи [1], де проаналізовано вживання 103 «основних» англійських ідіом у БНК й показано, що найчастотніші з них трапилися в корпусі не більше, ніж 500 разів, переважна більшість – менше, ніж 100 разів, з них чимало – менше, ніж 20 разів.

Слід зауважити, що «ЛексисНексис» є корпусом в тому сенсі, що це збірка задокументованих автентичних електронних текстів із пошуковими засобами. Він цілком складається із публіцистичних текстів (тобто є незбалансованим), і його вміст не було проаналізовано морфологічно чи синтаксично автоматичними чи іншими засобами. В цих аспектах він відрізняється від корпусів на кшалт БНК і БАМ-у і, безумовно, поступається їм. Однак у плані знаходження контекстів вживань ФО його гнучка пошукова система, зручний інтерфейс і величезний обсяг дають йому незрівнянну перевагу над іншими корпусами.

Що стосується придатності корпусу «ЛексисНексис» для дослідження англійських прислів'їв, то на його користь говорить, звичайно, його величезний обсяг. Нижче, в таблиці 3, наведено порівняльні дані частоти вживання 10 англійських прислів'їв у БНК та корпусі «ЛексисНексис»[1].

[1] Оскільки ми не мали змоги провести пошук у більшому за обсягом БАМ-і, ми вдалися до БНК, використавши пошуковий інтерфейс за адресою http://view.byu.edu/. Важливою особливістю цього інтерфейсу є те, що він не має мінімальних обмежень щодо кількості видобутих з корпусу рядків.

У першому стовпчику наведено прислів'я (в основній формі), у другому і третьому – кількість знайдених вживань у БНК і «ЛексисНексис» відповідно.

Таблиця 3. Частота вживання прислів'їв у БНК та корпусі «ЛексисНексис».

Прислів'я	БНК	«ЛексисНексис»
Necessity is the mother of invention	7	1353
(The love of) money is the root of all evil	7	910
Absence makes the heart grow fonder	7	498
An apple a day keeps the doctor away	2	348
tation is the sincerest form of flattery	1	1493
ke while the iron is hot	1	762
k of all trades and master of none	1	482
apple doesn't fall <never falls> far from the tree	0	576
A good name is better <rather be chosen> than great riches	0	23
Appetite comes with eating	0	10

Для перших восьми прислів'їв ми використали згадану вище формулу обчислення кількості вживань ФО в «ЛексисНексис», а для решти підрахунок було здійснено вручну. Як засвідчує таблиця, БНК значно поступається «ЛексисНексис» і не дає змоги сформувати реальної картини функціонування прислів'їв. Наприклад, маючи в розпорядженні лише БНК, можна було би зробити хибний висновок, що прислів'я *Absence makes the heart grow fonder* є досить вживаним, *Jack of all trades and master of none* – вживаним рідко, а *An apple doesn't fall <never falls> far from the tree* – взагалі вийшло з ужитку. «ЛексисНексис» показує, що ці три прислів'я мають приблизно ту саму – досить високу – частотність і жодне з них не виходить з ужитку. Щодо останніх двох прислів'їв у таблиці, то корпус засвідчує, що вони є досить рідко вживаними, але все ж наявні в колективній свідомості носіїв мови, хай навіть у пасивному запасі.

Отже, для повноцінного, багатоаспектного дослідження фразеології та її лексикографічного опису необхідно мати в розпорядженні збалансований корпус обсягом на порядок більшим, ніж БАМ чи БНК. Таких корпусів англійської мови, за нашими даними, наразі немає, однак корпус «ЛексисНексис», попри свою незбалансованість, має більш ніж достатній обсяг і дає хороші можливості дослідження англійської фразеології, в тому числі й прислів'їв, що мають значно нижчу частотність вживання проти інших фразеологізмів.

Література

1. Grant L. E. Frequency of the 'core idioms' in the British National Corpus (BNC) // International Journal of Corpus Linguistics. – 2005. – С. 429–451.
2. LexisNexis Academic, http://www.lexisnexis.com/academic/universe/academic/, 27.02.2006р.
3. Moon R. Fixed Expressions and Idioms in English. A Corpus-Based Approach. – Oxford: Clarendon Press, 1998. – 338 с.
4. Moon R. Needles and Haystacks, idioms and corpora: Gaining insights into idioms, using corpus analysis // The Perfect Learners' Dictionary. – Tübingen: Max Niemeyer Verlag, 1999. – С. 265–281.

О. Тищенко, к. філол. н.*
Інститут української мови НАН України (Київ)
УДК 161.2.81'322.221.24

ЗАСАДИ СТВОРЕННЯ КОРПУСУ УКРАЇНСЬКОЇ ЖЕСТОВОЇ МОВИ ГЛУХИХ

У статті обґрунтовано актуальність і мету створення корпусу української жестової мови глухих як автономної комунікативної системи. Означено змістові, структурні, інструментальні й технічні принципи такої праці.

Актуальність. Уходження в лінгвоукраїністику такого напряму, як корпусна лінгвістика [3; 11], спонукає до розвідок зі створення корпусу не лише природної словесної мови, а й української жестової мови глухих (УЖМГ). Це зумовлено активізацією досліджень, зокрема й лінгвістичних, національних жестових мов як за кордоном [1; 2; 3; 7; 8; 19; 20], так і в Україні [10; 12; 13; 17] у зв'язку з гуманізацією та соціалізацією всіх галузей науки [9].

* © *О.Тищенко, 2006*

Жестова мова глухих (ЖМГ і ЖМ) – це автономна природна комунікативна система, зовнішня сторона якої має мануально-мімічну форму. Це мова, втілена у жестах, сприймана зоровим аналізатором, основний орган її творення – руки. Основною цільнооформленою відтворюваною одиницею ЖМГ є **жест** для позначення різних фактів дійсності (предметів, процесів, відношень). Внутнішня сторона жесту пов'язана з відображенням цих фактів дійсності у людській свідомості. Жести – це, як правило, умотивовані матеріальні знаки ідеальних уявлень, вони мають власну структуру, особливі джерела номінації. Автономність ЖМГ стосовно природної словесної мови визначається її власною морфологією, синтаксисом, семантикою тощо.

Новизна. Вітчизняні студії лінгвістики УЖМГ перебувають, на жаль, у зародковому стані й репрезентовані фрагментарними свідченнями в межах педагогічної літератури як елементи методики вивчення жестової мови. Відсутність її цілісної лінгвістичної картини, зокрема усталеної метамови та понятійної бази, звичайно, ускладнює вироблення принципів корпусу, однак не робить його неможливим.

Передбачаємо, що становлення нового напряму науки – української сурдолінгвістики відбуватиметься паралельно зі створенням такої електронної бази ЖМ. Так, якщо для дослідження вербальних мов корпусна лінгвістика, окрім нових даних, дає змогу *уточнити*, *перевірити* зроблені раніше результати й висновки, то для ЖМ корпус буде першоджерелом для наукових висновків.

У дослідженні виходимо з визначення **корпусу мови** як певним чином (фізично, логічно) організованої, параметризованої *множини* елементів (текстів), репрезентативних для певної мови, діалекту чи соціолекту, *призначеної* для лінгвістичних студій чи програмних застосувань [3; 15]. Організація корпусу може бути різною залежно від прагматичних цілей як його укладачів, так і користувачів. Відповідно наявні на сьогодні бази даних жестової мови – словники, добірки навчально-методичного характеру (на паперових чи електронних носіях) [1; 4; 5; 10; 12; 19] – не є корпусами. Наприклад, відеокурс із навчання жестової мови [5] – репрезентований відеофайлом (демонстровані „диктором" повторювані двічі окремі жести – еквіваленти слів, розміщені алфавітно за першою буквою слова-еквівалента: АВТОБУС, АПТЕКА...), який супроводжено графічним (писемна форма слова-еквівалента).

У зв'язку з цим, а також з відсутністю будь-яких теоретичних розвідок у цій галузі, сьогодні некоректно говорити про українську корпусну сурдолінгвістику, доцільно лише означити **мету** створення корпусу ЖМГ – *забезпечити українську сурдолінгвістику репрезентативною вибіркою текстів жестової комунікації для наукових спостережень над сучасною ЖМГ* та **завдання**, розв'язання яких наблизить до реалізації корпусу – *вироблення ідеології корпусу*, *теоретичне обґрунтування механізмів його створення*, *формування концепції предметного наповнення* тощо.

Як відомо, сучасні корпусні розроблення охоплюють сферу словесної мови – письмової та усної її форм. Для розвитку лінгвістичної науки загалом та корпусної лінгвістики, зокрема, важливим є системний підхід, який передбачає вивчення *об'єкта* як специфічної *цілісності. У* такому разі до експериментального опрацювання слід залучати максимально репрезентативний обсяг даних про об'єкт вивчення, оскільки деякі коректні лінгвістичні порівняння можна зробити лише на великому обсязі фактичного матеріалу [15].

Довільні вибірки мають відповідати універсально-статистичним критеріям достатності, надійності та спеціальним критеріям якісної репрезентативності об'єкта (у часі й просторі) [16]. Будь-яка національна жестова мова глухих, як і будь-яка інша національна мова, представлена розмаїттям узусу – текстами мануальної (жестової) форми у різних сферах, ситуаціях, різної мотивації та цілей. Тому корпус – це вибірка текстів за *інтегральними* (жанровими) характеристиками та специфічними характеристиками *одиниць* (для вербальної мови – лексеми, слофоформи тощо, для ЖМГ – жест, жестова фраза та ін.). Оптимальна форма представлення такої вибірки на сьогодні – це електронна база даних.

ЖМГ не має закріпленої традицією писемної форми, крім деяких спроб окремих дослідників та користувачів ЖМГ створити спосіб фіксації кінетичних форм у двовимірній площині [4; 20]. Отож, прообразом корпусу ЖМГ може бути корпус *усної словесної мови* [14] – мови, втіленої у звуках, сприйманої на слух, зовнішню форму якої реалізує артикуляційний апарат. ЖМГ, як і усна форма вербальної мови, має переважно спонтанний, ситуативний характер, є первинною відносно будь-яких засобів фіксації (для

ЖМГ – писемні нотації, транскрибування, для вербальної мови – традиційна писемна форма), вона невідтворювана і не може бути перенесеною в просторі або збереженою в часі.

Оптимальний спосіб фіксації ЖМ у дії – це **відеозапис** (як для усної – аудіозапис або озвучене відео), який, окрім статичної форми рук, тіла, міміки, артикуляції (як на фото чи малюнку), відбиває *характер руху*: напрям, інтенсивність, швидкість, повторюваність та ін., а також *прагматичні особливості* певного жестового тексту, привнесені завдяки особистісним характеристикам мовців: „почерк", „вимова".

Одним із перших репрезентативних корпусів усної вербальної мови є TIMIT − корпус американського варіанту англійської мови (1998–1990 рр.), усну англійську дитячу мову репрезентують корпуси CHILDERS, POW та ін. Принципи їх створення лягли в основу першого корпусу для російської мови в Інституті системного аналізу РАН (1996–1998 рр.) – ISABASE [14]. Українські досягнення у цій галузі представлений лише електронними базами даних без спеціальної розмітки, маркування тощо, напр. українські говірки Донеччини не мають структури корпусу [18].

Аналізуючи досвід створення корпусів усної мови, констатуємо, що для формування корпусу УЖМГ передусім необхідно вирішити низку таких питань.

А. Щодо **змісту** слід визначитися з принципами:

1. Добору жестового текстового матеріалу: із ситуації живого, спонтанного спілкування, репрезентативного для різних ситуацій; інтерпретації зразків вербальної форми (напр., корпус усної мови TIMIT становлять 2342 окремі речення). Жестові тексти мають „покривати" усі сфери функціонування ЖМ, при цьому оперувати максимальною кількістю жестів, виявляти різні граматичні значення, відношення.

Слід зазначити, що серед засобів комунікації глухих виокремлюють природне жестове мовлення (ЖМл) як зовнішню форму жестової мови (ЖМ) та кальковане жестове мовлення як мануально-мімічну форму вербальної мови (КЖМл). Наше завдання – зібрати матеріал передусім ЖМ як репрезентації оригінальної комунікативної системи (говоритимемо про ЖМ і як про систему, і як про діяльність з її використання).

Вибір ЖМ чи КЖМл як засобу спілкування зумовлено багатьма чинниками, зокрема ситуацією – *формальною* (ширше застосовують КЖМл) або *неформальною* (надають перевагу ЖМ). Спеціальних досліджень, які б означили функціональну диференціацію ЖМ в Україні, на жаль, немає[1]. Однак носії природної ЖМ свідчать, що вона спроможна забезпечити обидва дискурси. Щодо сфер вживання ЖМ можна провести паралелі зі сферами, які обслуговує вербальна мова: це побут, сімейні стосунки (батьківські та інтимні), дозвілля, освіта, наука, виробництво, спорт, політика, культура, мистецтво, релігія тощо. Відповідно, майже в кожній з цих сфер ЖМ репрезентована у *неформальній, неофіційній* формі та *формальній, офіційній* (коли можна так сказати – в *стихійно стандартизованій* формі ЖМ, яка часто переходить у форму КЖМл).

Отже, предметна сфера жестових текстів має репрезентувати названі дискурси (кількісно відповідно до їхнього представлення у житті глухих).

Як і будь-яка національна вербальна мова, українська жестова мова має територіальні й соціальні діалекти, що засвідчено завдяки емпіричному досвіду глухих та сурдоперекладачів, проте не зафіксовано в жодному із наявних словників жестової мови. Виявити ці особливості (поки що хоча б на рівні „лексичному") також є важливим завданням корпусу.

Змістову складову для створення корпусу ЖМ можна через добір зразків а) монологічного мовлення; б) діалогічного мовлення носіїв природної жестової мови у невимушеній обстановці (під час налаштовування, спрямування комунікантів до необхідної теми, форми мови слід ураховувати психологічну, соціальну специфіку людей із функціональними обмеженнями). Пропонуємо використовувати а) навідні питання; б) спеціально підготовлені вербальні тексти, які відбивають розмаїття ситуацій, притаманних життю глухої людини у різних сферах (зазначених раніше). Ці тексти призначені не для *перекладу* калькованим жестовим мовленням, а для *інтерпретації* жестовою мовою. Порівняльний аналіз, сподіваємося, виявить територіальну, соціальну, прагматичну специфіку жестових текстів та сприятиме підготовці до введення певних ремарок: *діал., сленг, жарг., проф. т*ощо.

---

[1] Предметом майбутніх ґрунтовних розвідок у галузі ЖМ мають стати поняття сфери мовного спілкування, суспільні функції ЖМ, стилю, жанрів, адресності ЖМ, тип мислення та ін.

2. Кількісного та якісного добору комунікантів із урахуванням регіональної належності, місця проживання (місто чи сільська місцевість) віку, статі, освіти, захоплень, професії, релігійних переконань, знання/незнання вербальної мови, калькованого жестового мовлення тощо. Важливим параметром є характер фізичних вад: чи це глухоніма людина, чи глуха із навичками вербальної усної та письмової мови, чи німа (перев. безгортанна) зі збереженим слухом, чи сурдоперекладач (перекладач із жестової й на жестову мову) без вад слуху.

Комунікантів має бути не менше 100 осіб. Відповідно, для кожного слід заповнити *анкету,* дані якої в електронному вигляді супроводжуватимуть файл із жестовим текстом щоб забезпечити можливість наступних різнопланових досліджень ЖМ, зокрема соціальних, гендерних, вікових тощо:

АНКЕТА

Прізвище, ім'я, по батькові
Місце проживання: *область, місто (село)*
Стать
Вік
Освіта *(немає, середня, вища, науковий ступінь тощо)*
Професія
Місце роботи
Захоплення
Коло постійного спілкування *(глухі – ...%, чуючі – ...%)*
Знання й навички вербальної мови:
Ступінь розуміння вербальної мови
– української *(високий, середній, низький, відсутній)*
– російської *(високий, середній, низький, відсутній)*
Ступінь володіння вербальною мовою
– українською *(високий, середній, низький, відсутній)*
– російською *(високий, середній, низький, відсутній)*
Знання й навички жестової мови (ЖМ)
Ступінь розуміння ЖМ
– української *(високий, середній, низький, відсутній)*
– російської *(високий, середній, низький, відсутній)*
Ступінь володіння ЖМ
– українською *(високий, середній, низький, відсутній)*
– російською *(високий, середній, низький, відсутній)*
Ступінь відмінності української ЖМ та російської ЖМ:
– відмінні на ....%
– немає відмінності між українською ЖМ та російською ЖМ
Знання й навички калькованого жестового мовлення (КЖМл):
Ступінь розуміння КЖМл *(високий, середній, низький, відсутній)*
Ступінь володіння КЖМл *(високий, середній, низький, відсутній)*
Характер фізичної вади:
– глухота повна вроджена чи набута (вік),
– глухота часткова вроджена чи набута (вік),
– німота зі збереженим слухом вроджена чи набута (вік)
– вад немає
Умови зростання:
– у сім'ї глухих
– у сім'ї чуючих
– у сім'ї глухих та чуючих
Умови навчання:
– *„усний" метод (вербальна мова)*
– білінгвальний метод (жестова й вербальна мова)
Чи розрізнює *формальну* ЖМ та *неформальну* ЖМ *(так, ні)*
Який засіб (формальну, неформальну жестову мову, вербальну мову, кальковане жестове мовлення, дактиль) використовує для спілкування
– у побуті:
– у навчальній (виробничій) сфері:
– під час дозвілля:

– інші сфери:

3. Розподілу текстового матеріалу за комунікантами.

4. Добору типів інформації, асоційованої з кожним відеофайлом: писемний вербальний еквівалент, писемний жестовий еквівалент (напр., вербальна фраза „поговоримо після заняття" може бути передана жестами, писемний еквівалент яких є таким: ЗАРАЗ ЗАНЯТТЯ ГОВОРИТИ НІ або ЗАНЯТТЯ КІНЕЦЬ МИ ГОВОРИТИ та ін.); транскрипція жестової фрази – запис кожного жесту за допомогою спеціальних графічних знаків на означення складових жеста – кінем (фонем) – конфігурації, локалізації, характера руху тощо; усний вербальний еквівалент (аудійований).

Б. С**труктурні** проблеми пов'язані з організацією інформації в корпусі: визначення одиниці зберігання (жест, жестовий текст), вибір формату, зручного для розміщення, збереження, пошуку й використання необхідної інформації.

В. Т**ехнічні** питання потребують з'ясування умов відео- й аудіозапису жестового мовного матеріалу – це стосується вибору типів і кількості відеокамер, мікрофонів, звукових та відеокарт комп'ютера, режиму цифрового кодування, формату файлів тощо. Так, тривимірність жестової мови потребує *двох камер* для фіксації *монологу* (анфас і профіль) і *трьох камер* для зйомки *діалогу* (два анфаси й профіль обох комунікантів).

Г. Вирішення **інструментальних** питань забезпечить автоматизацію і стандартизацію різних етапів створення жестового корпусу: транскрибування, вербалізацію, аудіювання, сегментацію тощо. Це потребує створення спеціальних робочих інструкцій і, відповідно, спеціального комп'ютерного інструментарію для забезпечення зручності, швидкості й надійності роботи – анотування відеофайлів, тобто супроводження акустичною, письмовою та іншою інформацією.

Так, із кожним відеофайлом буде асоційовано ще кілька файлів: звуковий, текстові (засвідчуватимуть різні рівні розмітки: писемний вербальний еквівалент, писемний жестовий еквівалент, транскрипція та ін.). Сегментація жестового тексту, репрезентованого відеоматеріалом, певно, здійснюватиметься відповідно до елементарних дискурсних одиниць.

Документація, якою буде супроводжено корпус УЖМГ, забезпечить користувача вичерпною інформацією з питань його предмета, концепції, технічного використання.

Отже, потреба у корпусі української жестової мови зумовлена активізацією її лінгвістичного дослідження, необхідністю формування анотованої певним чином бази емпіричних даних ЖМ для можливого створення на її основі словників, підручників та ін. Реалізація такого корпусу потребує ґрунтовних теоретичних досліджень та технічного забезпечення.

Література

1. Гейльман И. Ф. Ваш друг – жестика. – СПб., 2003. – 173 с.
2. Гейльман И. Ф. Знакомтесь: ручная речь. – 2001. – 171с.
3. Демська-Кульчицька О. М. Основи національного корпусу української мови – Київ, 2005. – 219 с.
4. Димскис Л. С. Изучаем жестовый язык. – Минск: НМЦентр, 1998. – 135с.
5. Жест. Відео. – ТОВ „Республіканський навчально-науковий центр ДІНІТ"– CD. – Ч. 1, 2.
6. Жестовые языки // Энциклопедия «Кругосвет». – Central European Uniyersity Regents, 2001. – www.rol.ru/
7. Зайцева Г. Л. Жестовая речь. Дактилологія. – М., 2000. – С. 37–38.
8. Зайцева Г. Л., Фрумкина Р. М. Психолингвистические аспекты изучения жестового языка. – Дефектология. – 1981. – №1. – С. 7.
9. Закон України Про внесення змін до Закону України «Про основи соціальної захищеності інвалідів в Україні». – №1773–IV, стаття 23. – К., 2004. – С. 1.
10. Иванюшева Н. В. Мимико-жестикуляционная речь. – К.: Реклама, 1968. – 290с.
11. Карпіловська Є. А. Вступ до комп'ютерної лінгвістики. – Донецк: Юго-Восток, 2003. – 183 с.
12. Краєвський Р. Г. Мова жестів глухих. – Київ: Радянська школа, 1964. – 219 с.
13. Кульбіда С. В. Стан використання і вивчення жестової мови в Україні / Матеріали міжнародної науково-практичної конференції «Проблеми сучасної спеціальної освіти (українсько-канадський досвід)» – Київ–Едмонтон, 2004. – С. 63–67.
14. Речевые корпусы (опыт разработки и использование). Кривнова О. Ф., Захаров Л. М., Строкин Г. С. – М., МГУ. Електронний ресурс: http:/www.dialog.ru.
15. Рыков В. В. Прагматически ориентированный корпус текстов / Сб. Тверской лингвистический меридиан. – М., 1990.
16. Системный подход и корпусная лингвистика. – Електронний ресурс: www.philol.msu.ru.

17. Тищенко О. М., Кульбіда С. В. Українська жестова мова глухих як об'єкт лексикографічної параметризації / Лексикографічний бюлетень. – Київ, 2005. – Вип. 12. – С. 37–43.
18. Українські говірки Донеччини. Фонотека. – Донецький національний університет. – Донецьк., 2000. – CD, Вип. 1.
19. Фрадкина Р. Н. Говорящие руки. Тематический словарь жестового языка глухих России. – М., 2001. – 597с.
20. Цыпин Р. Жестовый язык и письменность / Наше життя, 2002. – 16 серпня.

Л. Тименко*
Інститут української мови НАН України (Київ)
УДК 161.2.81'373.46.322

НАУКОВА ТЕРМІНОЛОГІЯ В НАЦІОНАЛЬНОМУ КОРПУСІ УКРАЇНСЬКОЇ МОВИ

Питання, порушені в статті, пов'язані з проблемою створення національного корпусу української мови, яку опрацьовують у відділі лексикології, лексикографії та національного корпусу Інституту української мови НАН України. У статті висвітлені засади формування підкорпусу наукової термінології, обґрунтовано принципи добору текстів, а також хронологічні межі.

Для української мови завдання створення різноманітних корпусів текстів є дуже актуальним. Без таких корпусів і баз, побудованих за різними параметрами класифікації текстів та вміщеної в них інформації, неможливе успішне розв'язання багатьох важливих теоретичних і практичних проблем сучасної україністики, зокрема вироблення критеріїв унормування та кодифікації лексикону й граматичного ладу сучасної української мови, вивчення властивих їм динамічних процесів, проблем мовного планування й прогнозування, створення надійного підґрунтя для розв'язання проблем варіантності та конкурентоздатності тих чи інших одиниць, питомої ваги в мові нових тенденцій, вивчення співвідношення інновацій і сталих явищ у різних підсистемах мови, вивчення тенденцій розбудови текстових структур і вироблення надійних рекомендацій для формування структури текстів нових типів. Особливо це стосується таких функціонально-стильових різновидів української мови, як ділова мова, законодавчі та інші державні документи (наприклад, мова дипломатії), реклама, мова засобів масової інформації, тексти з тих фахових галузей, які ще донедавна обслуговувалися російською мовою.

Однією із особливостей сучасної української мови є динаміка її лексико-семантичної системи, частиною якої є термінологічна лексика. Тому введення наукової термінології до національного корпусу української мови сприятиме глибшому й детальнішому опису історії української мови, точнішій хронологічній атрибуції лексичних і семантичних новацій, деталізованій демонстрації семантичного, стилістичного й фразеологічного порядку, адже, на думку В. Виноградова, «…історія слова повинна проводити весь зміст і ланцюжок змістових «приростів», усі «метаморфози». Вона намагається розкрити певні умови вживання слова в різні періоди його мовного життя» [2: 19].

У лінгвоукраїністиці від 60-х років минулого століття і до сьогодні існують значні досягнення в суміжних з корпусною лінвгістикою напрямах – прикладній (комп'ютерній) лінгвістиці, а саме праці В. С. Перебийніс, Т. О. Грязнухіної, Н. Ф. Клименко, М. М. Пещак, Є. А. Карпіловської, Л. І. Комарової, О. М. Демської-Кульчицької та створені комп'ютерні текстові ресурси, зокрема Комп'ютерна лабораторія КНУ ім. Т. Шевченка та лексикографічний процесор «СЛОВО», який розробили львівські дослідники О. М. Коссак, О. Коруд, Л. Хвищун, який допомагає створювати одно- та двомовні термінологічні словники. На сьогодні теоретичні засади одномільйонного корпусу української мови розроблено, а практичне його створення тільки починається [3]. Цю роботу проводить група національного корпусу Інституту української мови НАН України. Під час визначення аспектів створення термінологічного підкорпусу української мови виникають проблеми, які ми спробуємо вирішити. Це питання про жанри й джерела підкорпусу наукової термінології, про хронологічні межі, про метатекстову інформацію

* © *Л.Тименко, 2006*

цього підкорпусу, про розроблення системи кодування на основі стандартних мов кодування.

Основою нашого підкорпусу буде репрезентативна вибірка текстів 60-х років XIX століття – до сучасного періоду, зокрема: а) роботи видатних науковців України; б) публіцистика; в) науково-популярна література; г) ділові, офіційні документи. Ці джерела відповідатимуть наступним вимогам:

- видання українською мовою (в Україні чи за її межами);
- надання переваги сучаснішим перевиданням, які максимально зберегли орфографію й пунктуацію джерела.

При складанні корпусу, як ми вже говорили, важливим є питання про хронологічні межі, тобто тексти/уривки текстів якого періоду потрібно включати до науково-термінологічного підкорпусу. Системний підхід до творення нової, збирання та перегляду наявної народної термінології починається з другої половини XIX століття переважно в Галичині. Оскільки саме з II половини XIX ст. було закладено підґрунтя, на якому розвивалася, збагачувалася й удосконалювалася українська термінологія. Вперше проблема терміна в українській мові постала в 60-ті рр. XIX ст. на сторінках журналу «Основа», а також вченими-членами Наукового товариства імені Шевченка. Будемо вважати 60-ті роки XIX століття нижньою хронологічною межею опису термінології. Не залишаться поза увагою роботи науковців 20 – поч. 30-х років XX ст. – періоду так званого «золотого десятиріччя», тобто до нашого підкорпусу будуть внесені тексти/уривки текстів, починаючи з другої половини XIX ст. і закінчуючи XXI ст.

Не менш важливим є питання про жанри й джерела корпусу. Як зазначав А. М. Баранов, описуючи проблеми корпусу текстів сучасної російської публіцистики (90-і рр. XX ст.), «… текстів повинно було бути достатньо багато, щоб відобразити всі релевантні властивості проблемної галузі» [1: 131–132], у нашому разі – властивості термінології як окремого функціонально-стильового різновиду сучасної української літературної мови. Українська мова, зокрема наукова термінологія – це великий масив різних за походженням і семантичною/стилістичною віднесеністю слів і словосполучень: це вся без обмежень термінологія (як власне українська, так і запозичена, особливо терміни, які трапляються в декількох термінологічних системах, або терміни, які набули широкого (переносного) значення). Постає питання, скільки місця в загальному корпусі має займати термінологія? На нашу думку, це приблизно 10% всієї лексики.

Наступна проблема – це метатекстова інформація, завданням якої є забезпечення багатофункціонального використання корпусу різними дослідниками. Тексти будуть уводитися до корпусу й перетворюватимуться у корпусний формат TEI, де буде створено електронний заголовок до тексту, а полотно тексту розмічене на III рівні CSS. Модель електронного заголовка повинна містити дані про:

1) повну й коротку назву тексту;
2) графіку тексту (книжно-слов'янська; орфографія певного хронологічного періоду; використання латинської та інших видів графіки), наявність іноземних елементів у тексті, наявність формул, малюнків і т. ін.;
3) атрибуцію текстів його автором;
4) прізвище, ім'я, по-батькові автора/авторів текстів;
5) науковий/літературний псевдонім автора;
6) дата народження/дата смерті автора;
7) датування створення/датування першої публікації;
8) вказівку на першу публікацію (зібрання творів, окреме видання, наукове видання; журнал, газета, архівний матеріал і т. д.);
9) вказівку на наявність авторських (редакторських) коментарів, приміток, виправлень; спеціальних зауважень щодо підготування тексту до видання;
10) вказівку на спосіб уведення тексту в корпус (ручне введення, СД-версія, сканування, Інтернет, оригінал-макет видання, авторська копія);
11) вказівку на осіб, які вводили текст до корпусу, робили розмітку й філологічну перевірку тексту; осіб, які підготували формуляр тексту;
12) вказівку на місце зберігання тексту (вказати на заклад, куди текст відправлено після обробки – адреса тексту).

Загалом, дослідницькі можливості сучасних електронних текстових корпусів/підкорпусів досить різноаспектні. Але найважливішим аспектом аналізованого дослідного підходу є те, що він «змушує подивитися на багато проблем з погляду,

відмінного від традиційного опису мови» [5]. Як справедливо стверджує В. Риков, корпус повинен «мати власну концепцію, бути чітко детермінованою щодо призначення, а фактичний матеріал повинен репрезентувати реальні тексти національної мови, враховуючи весь історичний період її існування, варіантність мовної реалізації, стилістичну диференціацію і теорію поширення» [6]. Отже, упровадження корпусних методик в українське термінознавство повинно обов'язково враховувати, що незалежно від типу здійснюваних корпусних досліджень, їхня якість залежатиме передовсім від коректності побудови самого термінокорпусу, яка має відповідати національній філологічній традиції.

Література

1. Баранов А. Н. Введение в прикладную лингвистику. – М., 2001.
2. Виноградов. В. «Чтение древнерусского текста и историко-этимологические каламбуры». // Вопросы языкознания. – 1968. – №1 – С. 19.
3. Демська-Кульчицька О. М. Основи національного корпусу української мови. – К., 2005.
4. Коссак О., Коруд О., Хвищун Л. Система комп'ютерної підтримки словникових баз даних СЛОВО // Проблеми українізації комп'ютерів. – К., 1993.
5. Мустайоки А., Копотов М. К вопросу о статусе эквивалентов слова типа потому что, в зависимости от, к сожалению // Вопросы языка. – 2004. – №3. – С. 88–107.
6. Рыков В. В. Корпус текстов как отражение состояния русского языка. – http// rykov-cl.narod.ru. – 2002.

А. Кузнєцова*
Інститут філології КНУ ім. Т. Шевченка (Київ)
УДК 161.2.81'373.322

ПАРАМЕТРИЗАЦІЯ ПУБЛІЦИСТИЧНИХ НЕОЛЕКСЕМ У КОРПУСНІЙ ЛІНГВІСТИЦІ

У статті аналізуються публіцистичні неолексеми з погляду їхньої параметризації. Так, проаналізовано корпус Частотного словника українського публіцистичного стилю, запропоновано етапи роботи зі створення словника неологізмів, визначено перспективи дослідження.

Характерною особливістю словникового складу української мови, як і будь-якої іншої, є його безперервний розвиток. Актуалізація в мові застарілих слів або значень, поява нових слів або розширення значень наявних слів спричиняється потребою людини називати нові предмети, явища, поняття, що з'являються у суспільно-економічному, політичному й культурному житті народу.

Пропонована робота присвячена дослідженню неолексем у публіцистичних текстах. Орієнтовані на пересічного читача, такі тексти є одним з найбільших джерел нових лексичних одиниць і саме в них ми маємо нагоду зафіксувати нові слова, які постійно виникають у мовленні. Але на сьогодні закономірності їх виникнення й функціонування ще не стали предметом фронтальних лінгвістичних досліджень.

Матеріалом нашого дослідження є корпус Частотного словника українського публіцистичного стилю – 2004, укладений студентами спеціалізації "Комп'ютерна лінгвістика". Лексичний склад словника групувався за рубриками: Політика (67 тис. слововживань), Суспільство (56 тис.), „Помаранчева революція" (54 тис.), Культура (46 тис.), Закон (29 тис.), Економіка (20 тис.), Спорт (12 тис.), Здоров'я (6 тис.), Релігія (6 тис.). Загальний обсяг вибірки становить 300 тис. слововживань. Словник укладений за матеріалами центральних та обласних газет „Українська Правда", „Україна молода", „Голос України", „Дзеркало тижня", „Наша Газета+", „Львівська газета", „Рівне вечірнє", „ПРАВДА", „Критика", „Контекст", „Коментар", „Церква і молодь", виданих протягом кінця 2003 та всього 2004 року. Дослідження матеріалу словника має на меті створення параметризованого словника неологізмів, здійснення опису їх словотвірних особливостей і розподіл нової лексики за функціональними групами. Досягнути цієї мети можна завдяки вирішенню наступних завдань:

1. Створення корпусу текстів публіцистики, об'єднаних за хронолічним і тематичним принципом. Форма подачі – електронна. На цьому етапі ми отримуємо вже готовий матеріал, поділений у словнику за вищезгаданими рубриками і обмежений, відповідно, виданнями кінця 2003 та 2004 року.

* © *А. Кузнєцова, 2006*

2. Автоматизовано добираються неологізми, які відмічаються відповідними маркерами.

3. Наступним кроком вручну відкидаються всі квазінеологізми або слова з помилками. На цьому етапі відбір неологізмів здійснюються за такими критеріями:

- відсутність слова у Великому зведеному орфографічному словнику української лексики;
- відсутність слова у 2-ому виданні ВТС СУМ (видавництво „Перун”) 2005 року;
- відсутність слова у словниках діалектів української мови;
- відсутність помилок у слові.

Але відсутність слова у зазначених словниках не була достатньою причиною для вміщення його у словник, деякі слова не добирались ще й у таких випадках:

- якщо це складні прикметники з першим компонентом числівником і опорним компонентом, що містить основу іменника на позначення одиниці виміру, наприклад: *тридцятисорокарічний, чотирьохсотграмовий, п’ятидесятидвохміліметровий* і т.ін.;
- якщо це складні прикметники, обидва компоненти яких утворені від назв країн, наприклад: *українсько-французький, візантійсько-російський, польсько-румунський;*
- прикметники, утворені від прізвищ пересічних громадян: *базилинський, брошковський, іванченківський.*

Усі вони утворені за одним словотвірним зразком, який допускає сотні й тисячі різних конкретних його реалізацій. Із цього випливає, що весь масив таких словоутворень не може й не повинен реєструватись.

4. Створення конкордансів (контекстів речень для однозначного тлумачення неоодиниць). На цій стадії роботи ми також використовуємо ресурси Частотного словника, адже кожна реєстрова одиниця в ньому шляхом натиснення миші подає всі контексти вживання та відповідне джерело цитування.

5. Після відбору неологізмів укладається параметризований словник неолексем. Нові слова в ньому розподіляємо такими кроками:

- кожне слово ділимо на морфеми;
- визначаємо частину мови;
- записуємо морфемну формулу;
- подаємо мінімальну словотвірну пару, якщо слово є похідним;
- визначаємо мотивуючу основу слова;
- визначаємо словотвірний формант у похідному слові, за допомогою якого воно утворилось ;
- визначаємо спосіб словотвору кожної нової лексеми;
- визначаємо словотвірне значення, яке сформувалось на базі семантичних відношень між твірною базою і похідним словом;
- наступним пунктом подаємо контекст у якому зустрілась ця лексична одиниця;
- до слів подаємо дані про їхню частоту;
- останнє – відомості про джерело, з якого взято лексему.

Вже на цьому етапі роботи над словником можна зробити деякі висновки про словотворчі особливості неологізмів:

- найбільш продуктивною групою виявилися неолексеми, утворені способом суфіксації, зокрема іменники на позначення осіб *(кучманоїд, мажоритарник, оранжист, піарник, толкініст)* і абстрактних понять *(маскулінність, поліцистика, поттеріана, постіндустріальщина, дефініціоналізм);*
- друге місце за продуктивністю посідає група, утворена способом префіксації, зокрема такі лексеми, як *квазі-інавгурація, антивідмивальний, інтерконтинетальний, затусуватися, мегахіт, проєвропейський*;
- наступними за кількістю є слова-абревіатури, що належать головно до лексики політико-економічної спрямованості *(адмінсправа, політреформа, політтехнолог, інвестфонд, спецрубрика, телемарафон, укрнет, гм-продукт);*
- далі йдуть неолексеми, утворені способами основоскладання і словоскладання, це такі слова, як *вузькосировинний, грантоїд, металодетектор, інтернет-агенство, кінопортал, конкурентноздатність;*
- наступною групою за ступенем продуктивності виявилися неологізми, утворені способом транслітерації, що свідчить про активний процес проникнення в українську мову слів іншомовного походження. Це такі лексеми, як *екзит-пол, бейдж, індастріал-фолк, лайт-бокс, меседж, реаліті-шоу, стем;*

- низькою кількістю характеризуються слова, що були створені префіксально-суфіксальним способом. Одиничними випадками представлені слова, що творились за допомогою постфіксації та усічення як засобу творення сучасного молодіжного сленгу *(синхрон, фест, відкат).*

6. Після завершення роботи над словником відбувається систематизація неологізмів за словотвірними типами й словотвірними категоріями. Дослідження неолексем публіцистики свідчить, що реєстрові одиниці словника були сформовані за допомогою традиційних способів словотвору.

7. Далі здійснюється поділ лексики на функціональні групи. У результаті аналізу неологізмів за семантико-стилістичними ознаками встановлено такі функціональні групи неолексем:

- жаргонна лексика (молодіжний жаргон, арго): *намисто-фєнєчки, скарфер, толкініст, драйвово.*
- розмовна лексика: *кучміст, майданувати, наївняк, вернадка, попсовий.*
- професіоналізми й терміни: *реадмісія, екстермінація, міні-соціум, мікрорівень, термочутливий.*

За нашими підрахунками у корпусі із 300 тис. слововживань виявилось 475 неологізмів. Очевидно, далеко не всі вони можуть увійти до активного вжитку або закріпитись у мові. Найбільші шанси мають терміни, створені відповідно до найпродуктивніших словотвірних типів і за допомогою високопродуктивних засобів деривації *(іпотерапія, дактилокартотека, поліцистика, сателітарни).* Крім того, неологізм має бути однозначним і потенційно бути конституентом широкого дериваційного гнізда .

Не претендують на закріплення передусім багатокомпонентні композити на зразок *фантастично-футуристичний, неймовірнобрудноподаний* або *демократ-неоющенко*, а також деривати з яскраво вираженою експресією, наприклад: *толкінутий, панкуючий, супермарафон.*

Матеріал цієї роботи цікавий для дослідження основних тенденцій словотвору української мови, функціонально-стилістичних розвідок, а також експлікації отриманих знань на теорію і практику лексикографії (створення тлумачних словників, словників неологізмів, термінологічних словників). Подібні дослідження корисні для простеження основних причин появи нових лексичних одиниць та їхньої перспективності у мовному вжитку.

Література

1. Берков. В. П., Павлов В. М. Новые слова? // Нове слова и словари нових слов. – Л., 1983.
2. Горпинич В. О. Словотвір і словотворення української мови. – К., 1995.
3. Корпусная лингвистика // Баранов А. Н. Введение в прикладную лингвистику: учебное пособие. – М., 2001.
4. Сучасна українська літературна мова / за ред. І. К. Білодіда. – Т. 1. – К., 1973.
5. Ярцева Н. В. Неологизмы // Лингвистический энциклопедический словарь. – М., 1990.

R. Garabík, Ph.D*
Jazykovedný ústav Ľ. Štúra SAV (Bratislava, Slovakia) УДК 81’322

COMPUTER(IZED) LINGUISTIC RESOURCES AT THE Ľ. ŠTÚR INSTITUTE OF LINGUISTICS

Abstract. *Ľ. Štúr Institute of Linguistics, as the main linguistic institution in the Slovak Republic, publishes a lot of information either in the form of dictionaries, scientific publications or periodicals, targeted both to scientific community and general audience. With increasing amount of information produced, information retrieval and effective searching can be greatly assisted by organizing and indexing electronic versions of the documents. As a public service, the Institute offers some of the data to be accessed and searched on the Internet.*

Introduction

Ľ. Štúr Institute of Linguistics is the main linguistic institution in the Slovak Republic. Among its main areas of research and activities is the research of the contemporary Slovak language, Slovak dialectology, study of history of the Slovak language, and also connections

* © *R. Garabík, 2006*

with general and comparative linguistics. Recently, the Institute started to work in the fields of etymology and corpus linguistics. The sociolinguistic and culturologycal approaches are gaining momentum across various research topics. The Institute has a long-standing tradition of producing dictionaries either for academic or for general environment, linguistic atlases and specialized publications presenting different aspects of linguistic research. "Flagships" of Institute's publication activities are three specialized journals:

Kultúra slova is a scientific popular journal, serving mostly as an outpost for prescriptive activities, promoting the usage of cultural, codified literary Slovak language, introducing new professional terminology into the common use and a providing general discussion of language culture in both professional literature and fiction.

Slovenská reč is the oldest scientific journal focused on slovakistics. It is the base scientific journal dedicated to a generic research of the Slovak language, including grammar, morphology, syntax, onomastics, theoretical description of orthography, and also dialectology and phraseology. The journal serves as a place for professional discussion for Slovak linguists, brings information about oncoming events in Slovak linguistics, informs about new important publications and publishes concise lists of works of Slovak linguists.

Jazykovedný časopis is a scientific journal oriented towards general linguistics research, not limited to Slovak language, bringing also reviews of important linguistic publications and information about current events in linguistics.

Rather important publication is a tetralogy *Slovenskí jazykovedci – Súborná personálna bibliografia slovenských slovakistov a slavistov*, containing detailed list of publications of Slovak linguists, covering the years 1925 – 2000, with fifth volume in preparation [1].

Two publications targeted towards general public audience of Slovak language users are:

- *Krátky slovník slovenského jazyka* (KSSJ) – *the* dictionary of contemporary Slovak language, containing definitions of 63818 headwords [2].
- *Pravidlá slovenského pravopisu* (PSP) – a detailed description (definition) of Slovak orthography, together with dictionary-like section describing (defining) inflectional paradigms of most common Slovak words [3].

Importance of these two publications stems from the Slovak language law, which defines a codified version of Slovak language as that described in above mentioned publications.

Dictionary interface

KSSJ is by far the most demanded publication, because non-linguistically oriented people see the language codification mostly through the prism of "proper" and "improper" vocabulary items. Searchable version of KSSJ is accessible on-line at the address http://kssj.juls.savba.sk. Backend is based on homegrown version of DICT server[4] software (modified[1] serpento server, written in Python). Frontend is a simple cgi script, with a possibility to either directly find the definition of the word entered, or to search for entries beginning with given prefix, suffix or containing given substring. The maximum number of obtained results is limited to 30 (for various reasons, most important one being the fear of unauthorized copying of the dictionary data). If the query has been given without diacritics (i.e. containing only 7-bit ASCII characters), the search is performed on the index version built with diacritics removed – this is for the benefit of users who either cannot enter 8-bit characters in proper encoding (such as older or embedded browsers), or if entering proper diacritics is cumbersome (e.g. systems without Slovak keyboard layout), and also taking into account widespread custom to drop diacritics in internet context (reflecting the past era. when using non ASCII characters was often met with technical problems and encoding incompatibilities). The output – i.e. entry definition – is with full diacritics in any case. The script works fully in UTF-8 encoding, but it is possible to force the input and output encoding to be one of ISO 8859-1, ISO 8859-2, CP1250, CP1252, CP850, CP852, to facilitate the access with older or restricted browsers.

Similar interface is planned for the dictionary part of PSP, once all the necessary copyright issues are resolved.

Publishing Non-dictionary Information

There are many different ways how to effectively publish information on the WWW. Ideally, the system should not put much demand on system and human resources, should be easy to install and use, information stored inside should be easily modified. Plain

1 In fact, this will become next version of serpento in due time

static html files provide by far the best performance, however they fail short in terms of easy modification (that is, preferably by untrained users). In addition, an ability to provide an easy feedback from the users about the information is desirable.

Relatively recent phenomenon is a wiki collaboration software, a system that allows users to edit the information presented via a WWW interface [5]. Overwhelming success of a wiki-based collaborative approach can be seen on the development of Wikipedia [6].

Our system accessible from the Institute's main WWW page (http://www.juls.savba.sk) uses wiki engine written in the Python programming language, using the *Karrigell* application framework[1]. Thanks to the user access control list support, it is easy to implement the possibility to modify existing data for selected users from the Institute, and have a safe read-only access from anonymous users from the Internet. While at the first glance most of the data included is frozen (coming from already published printed information), there is a need to modify the data occasionally – especially the texts obtained from scanned printed sources, with scanning error remaining even after proofreading (the number of errors approaches zero as the amount of proofreading man-hours approaches infinity, and it is often desirable to make the information available immediately after the first basic proofreading, with further corrections carried out while the data are already accessible). As the texts are edited mostly by untrained persons (often students), the classic wikipedia syntax is deemed too difficult, with a steep learning curve. Therefore we opted for a much simpler and clearer reStructuredText syntax [7], backed up by an excellent *docutils* parsing module [8]. Effectively, from the point of view of proofreaders, the text data is stored as a plain text, with reStructuredText markers perceived only as a decoration of the text.

Table 1: Example of reStructuredText format (left) and the final rendered output (right)

MÁRIA ČIŽMÁROVÁ, rod. DZIAKOVÁ **----------** **Nar. 22. 6. 1969, Hrabová Roztoka, okr. Snina. Študovala v r. 1978 – 1983 na Filozofickej fakulte v Prešove Univerzity P. J. Šafárika v Košiciach (ukrajinský jazyk – slovenský jazyk). 1993 PhDr., 1995 CSc. V r. 1983 – 1985 pôsobila na Základnej škole v Humennom (učiteľka), v r. 1985 – 1997 pracovníčka Katedry ukrajinského jazyka a literatúry Filozofickej fakulty v Prešove Univerzity P. J. Šafárika v Košiciach, od r. 1997 Prešovskej univerzity (odborná pracovníčka, odborná asistentka).** **Venuje sa konfrontačnému výskumu ukrajinčiny a slovenčiny, pracuje aj v oblasti dialektológie.** **1996** **1. Konfrontačný výskum frazeologizmov v ukrajinčine a slovenčine. – In: Studia Philologica. 3. Acta Facultatis Paedagogicae Universitatis Safarikanae. Red. Z. Stanislavová et al. Prešov, Univerzita P. J. Šafárika v Košiciach, Pedagogická fakulta v Prešove 1996, s. 143 – 145, nem. res. s. 145.** **2. Do vlasnoho bereha. – Dukľa, 44, 1996, č. 2, s. 104 – 105.** **1997** **3. Uvaha botaničnij terminolohiji. – Dukľa, 45, 1997, č. 2, s. 110 – 111.**	**MÁRIA ČIŽMÁROVÁ, rod. DZIAKOVÁ** Nar. 22. 6. 1969, Hrabová Roztoka, okr. Snina. Študovala v r. 1978 – 1983 na Filozofickej fakulte v Prešove Univerzity P. J. Šafárika v Košiciach (ukrajinský jazyk – slovenský jazyk). 1993 PhDr., 1995 CSc. V r. 1983 – 1985 pôsobila na Základnej škole v Humennom (učiteľka), v r. 1985 – 1997 pracovníčka Katedry ukrajinského jazyka a literatúry Filozofickej fakulty v Prešove Univerzity P. J. Šafárika v Košiciach, od r. 1997 Prešovskej univerzity (odborná pracovníčka, odborná asistentka). Venuje sa konfrontačnému výskumu ukrajinčiny a slovenčiny, pracuje aj v oblasti dialektológie. **1996** 1. Konfrontačný výskum frazeologizmov v ukrajinčine a slovenčine. – In: Studia Philologica. 3. Acta Facultatis Paedagogicae Universitatis Safarikanae. Red. Z. Stanislavová et al. Prešov, Univerzita P. J. Šafárika v Košiciach, Pedagogická fakulta v Prešove 1996, s. 143 – 145, nem. res. s. 145. 2. Do vlasnoho bereha. – Dukľa, 44, 1996, č. 2, s. 104 – 105. **1997** 3. Uvaha botaničnij terminolohiji. – Dukľa, 45, 1997, č. 2, s. 110– 111.

1 http://karrigell.sourceforge.net

Fulltext Search

In addition to providing access to "raw" information and texts, we need to effectively search in the data. Fulltext search is provided by a Xapian library (http://www.xapian.org). As already mentioned above, users often enter queries without diacritics, even if it is otherwise available. Therefore we indexed the texts with stripped diacritics, and the query interface strips it from the query before passing it over to xapian – obtaining the effect of ignoring diacritics in the query system. In the future, our plans envisage to use a simple Slovak language stemmer when indexing the texts. In this context, lemmatization is less suitable then stemming, because of the need to disambiguate the lemmatized texts, while stemming (often) does not care about preserving grammar information, thus increasing recall at the expense of precision.

What is Included

Since the first volume of [1] has been typeset before the onset of computer typography, using traditional methods[1], it had to be scanned and proofread, the other three volumes were readily available as digital texts. The text was split into small units, each unit corresponding to one author mentioned in the volume, and then converted into reStructuredText format and placed into the wiki. All the four volumes have the size of 7.9 MB and contain together 1.17 million words.

At the time of writing, the archives of *Kultúra slova* from 1994 and *Slovenská reč* from 1997 onwards are processed and made available (with the exception of the latest year, which is available only to the Institute's employees, due to copyright reasons). The total amount of texts in these two journals is 17 MB, containing 2.28 million words.

Slovak National Corpus

Separate from the resources mentioned above, Slovak National Corpus is a project to build an annotated representative corpus of contemporary written Slovak language. Now, it contains about 256 million words, covering the language as published during 1955 – 2005 (obviously, with a huge bias towards the later date). The texts are automatically morphologically annotated and lemmatized, containing detailed bibliographic annotation. Smaller, balanced subcorpus containing 60% journalistic texts, 20% professional (scientific) texts and 30% fiction has the size of 56 million words. Another subcorpus contains manually lemmatized and annotated texts that could be used to train morphological analysers and other NLP tools. The size of this subcorpus is 322 600 words.

Access to the whole corpus and all the subcorpora (excluding texts with too prohibitive copyright licenses) is publicly available free of charge on internet[2], requiring a simple registration procedure.

Literature

1. Dvonč, L.: Slovenskí jazykovedci – Súborná personálna bibliografia slovenských slovakistov a slavistov, Veda, Bratislava, 1987, 1997, 1998, 2003.
2. Krátky slovník slovenského jazyka.Veda, Bratislava, 2003.
3. Pravidlá slovenského pravopisu. Veda, Bratislava, 2000.
4. Faith, R., Martin, B.: "A Dictionary Server Protocol", Request for Comments 2229, Network Working Group, October 1997, ftp://ftp.isi.edu/in-notes/rfc2229.txt
5. Ebersbach, A, Glaser, M.: Towards Emancipatory Use of a Medium: The Wiki. In: International Journal of Information Ethics, 2 (2004).
6. Ciffolilli, A.: Phantom Authority, Self-selective Recruitment and Retention of Members in Virtual Communities: The case of Wikipedia. In: First Monday, 12 (2003).
7. Goodger, D.: Python Enhancement Proposal 287: reStructuredText Docstring Format, http://www.python.org/peps/pep-0287.html
8. Goodger, D.: Python Enhancement Proposal 258: Docutils Design Specification, http://www.python.org/peps/pep-0258.html

1 i.e. phototypesetting
2 http://korpus.juls.savba.sk

B. Radić-Bojanić*
Faculty of Philosophy
Zorana Đinđića, Novi Sad (Serbia and Montenegro)
УДК 81'322

METHODOLOGY OF REGISTER ANALYSIS: COMPUTER-MEDIATED COMMUNICATION

The paper lists arguments in favour of compiling an annotated CMC corpus by illustrating how it could be used in the description of the methodology of register analysis, where such a corpus would be immensely useful. In addition, the paper lists some problems which will have to be dealt with during the compilation of such a corpus and states that they will have to be overcome if linguists want to conduct a thorough and detailed analysis.

1. Introduction

In the age of computers corpus linguistics has been dealing with the description and analysis of spoken and written texts in their electronic form. Surprisingly, communication in the electronic medium, which has made all this possible, has been largely neglected as regards corpus linguistics. Namely, despite the vast quantity of the spoken/transcribed and written texts which have found their place in electronic corpora, there have been very few attempts at including new modes of electronic communication (electronic chatting, e-mails, discussion group texts, etc.) into these corpora. This paper is meant to draw attention to this problem by showing how such corpora of computer-mediated communication (henceforth: CMC) could be used in linguistic research.

The current corpora greatly rely on the speech/writing dichotomy, which automatically excludes the new modes of electronic communication since they lie halfway between the two aforementioned poles. For instance, electronic chatting, which is formally written communication, displays a large number of oral characteristics. If CMC is to be understood fully and investigated in detail, the research will have to rely on an electronic corpus of texts, which would, in turn, be used in various types of analyses at different linguistic levels. Therefore, "it would seem well worth the effort to devise and gather well-sampled corpora typifying the new modes of communication, so as to understand their central and typical aspects and work towards the larger goal of unravelling the complex relationship between traditional forms of writing and speech, as well as between these traditional forms and the newer modes of communication, such as computer-mediated communication on the Web" [9: 92]. The process of compiling a CMC corpus needs to include two complimentary approaches: qualitative and quantitative. Whereas the qualitative aspect would deal with the circumstances in which communication takes place, the quantitative one would involve analyzing an electronic corpus with "existing part-of-speech software, in order to identify any unknown lexico-grammatical features, and [training] the software for future analysis of such modes of communication"[9: 92].

This paper deals with one form of CMC, namely electronic chatting. Through the description of the methodology of register analysis, the author hopes to show and prove the necessity of compiling a CMC corpus. After describing what is involved in register analysis, the paper goes on to consider the quantitative and qualitative aspects of electronic chatting and finishes with arguments which speak in favour of compiling CMC electronic corpora in various languages, not just of English.

2. Methodology of Register Analysis

"A communication situation that recurs regularly in a society (in terms of participants, setting, communicative functions, etc.) will tend over time to develop identifying markers of language structure and language use, different from the language of other communication situations" [6: 20]. Language structure is understood here as similar vocabulary, similar intonation features, similar syntax, etc. Register analysis, besides the typical linguistic characteristics of a register, must describe the extent of linguistic variability, e.g. the range of syntactic structures or intonations which occur in a communication situation. Studies of register have three major components [2: 10]:

- description of the situation in which the register is used
- description of the linguistic features of the register

* © *B. Radić-Bojanić, 2006*

- analysis of the functional or conventional associations between the situational and linguistic features.

These relationships are bidirectional: situational characteristics influence the choice of linguistic forms, while the choice of linguistic features helps to create the situation [2: 10].

Register analysis is conducted in several stages, which are described in detail in Biber [1: 64]. As is stated there, the preliminary analysis considers previous research in order to identify potentially relevant linguistic features. After that, a corpus of texts, which is to be used in the analysis, is compiled. Alongside with the compilation of a corpus, note must be made of all relevant characteristics of the situation in which the texts were produced. At the end of the preliminary phase, various linguistic features of the text are singled out and counted. With already annotated corpora this part of analysis is done fairly quickly, but the problem arises with the texts which are not annotated. This is the case with CMC, which is a paradox. On the one hand, all texts of CMC are readily available because of the very nature of the communication process – it takes place exclusively via computers. On the other hand, they are not annotated, which present a great problem in any kind of thorough linguistic research.

The very annotation of an electronic chatting corpus will encounter certain problems, connected with CMC-specific features, such as abbreviations, emoticons, typos, respellings, etc., illustrated in the text below:

<Shania> I just d/led at 405 k/s rofl !!!!!!!!!!!!!!! – abbreviations
<Siren> OMG! – abbreviation
<Rom|Breakfast> wb Shania sorry for being a smart idiot b4 – abbreviation, respelling
<Rom|Breakfast> omg – abbreviation
<Sphinx30> woohoo
<Siren> cool
<ICU2> Oh my I missed you so much.....NOT :P – emoticon
<Waiter|v|> ;) – emoticon
<Shania> Sphinx30 i wanna send u soemthing – respelling, abbreviation, typo
<Shania> ti see – typo
<Shania> and then u send me soemthing – abbreviation, typo
<Sphinx30> hmm ok
<Shania> 22 both ways for us Sphinx30
<Sphinx30> looks like it
<Sphinx30> i wish it would be faster
<Shania> try this Sphinx30
<Sphinx30> but its fast
<Shania> type /pdcc 20000
<Shania> why isnt it fast to Sphinx30 yet it is elsewhere
<Shania> Sphinx30 is egtting it at 22 k/s – typo
<Sphinx30> you did shoot it by mirc elsewhere Shania
<RED^EviL> HIIIIIIIIIIIIIIIIIIIIIIIIIIIII – emphasis
<HomeAlone> me to.....
<HomeAlone> :))))))))))) – emoticon

Besides that, a CMC corpus should contain information about punctuation or the lack of it, since that also contributes to determining the features of this register. As can be seen in the text above, punctuation is not used in the traditional way. Rather, it becomes part of the new CMC conventions, where punctuation marks function as a means of emphasis or building blocks for emoticons. In addition, capital letters have lost their traditional roles (capitalizing the personal pronoun "I", personal names, sentence beginnings, etc.) and they are also used in new ways, again mainly to achieve emphasis.

In the next phase of register analysis the linguistic features are grouped according to their frequency of occurrence in the texts and then are interpreted by assessing their communicative functions. In other words, a frequent occurrence of the passive voice points to a more formal style, usually found in scientific texts. Moreover, a frequent occurrence of personal and demonstrative pronouns speaks of the importance of the connection between the text and the extralinguistic context, especially in interpreting the reference of the pronouns.

The last phase, which is particularly important in the analysis of different kinds of texts and registers, compares statistical data for linguistic features and interprets their link with the texts and situations in which they are found more or less frequently.

In order to arrive at a description of the system of registers in a language, as well as to correctly explain the occurrence of linguistic features in some types of texts, a comprehensive register analysis must be based on three equally important factors: the description of a situation, representative samples of texts and a detailed description of linguistic features occurring in a

register. Biber [1: 70] states that texts differ on the basis of topics, purposes, rhetoric structures and styles, as well as situational parameters, such as the relationship among the participants in the communication process, the attitude of participants towards the extralinguistic content and the attitude of participants towards the text itself. Therefore, one of the mistakes that can occur in corpus compilation is not making notes of the participants in the communication process and the circumstances of text production [1: 70]. This causes mistakes in register analysis, since some linguistic features, without the information on extralinguistic factors, can be interpreted in a wrong way. That is why it is necessary for researchers to include information about the extralinguistic elements of the communication situation while compiling a corpus. To conclude, every thorough register analysis must be based on two very important methodologies: statistics, which testifies of the frequency of occurrence of linguistic features, and the ethnography of communication, which interprets the results of the investigation taking into consideration the extralinguistic and situational factors.

The quantitative analysis, conducted by way of statistics, interprets the differences among registers on the basis of differences in the distribution of linguistic features, which means that a relative frequency of several features identifies a register. The statistical processing of the data has a few phases [1: 75–78]. First of all, a choice is made of linguistic features which should receive special attention in a particular corpus or a particular kind of research. The researcher relies on the previous studies while making the choice of the features, which is part of the preliminary phase of the register analysis. In order to get as many details in the analysis as possible, a large number of potentially important features should be included. The counting is based on a text of 1000 words or on the texts of the same length so that the comparison is correct. In case the comparison is done on the texts of different lengths, the frequency data will not be correct. Therefore, according to Biber [1: 76], the formula for calculating the frequency of a certain linguistic feature is as follows:

number of occurrences of one feature/number of words in a text · 1000 =
number of occurrences of the feature in a text which is 1000 words long,
i.e. the frequency in 1000 words

Besides the quantitative analysis, there is a need for a non-quantitative one. While the former supplies an empirical background, the latter is necessary in order to interpret the results correctly. In other words, single linguistic features are neither enough to establish the differences among registers nor to describe one register in detail. On the contrary, the co-occurrences of several features and their interrelations are the core of the register description. Biber [1: 70] and Biber & Finegan [3: 7] stress the importance of the knowledge of the context in which the text was produced, since registers correspond to a series of activities characteristic of members of a community. That is why the ethnography of communication is a good starting point, since it studies the immediate use of language in a context. A social community in this case is the context, so its communicative activities are studied in the whole [7: 15–16]. The ethnographic practice is based on observing the communicators, on stressing the importance of the specificity of social life and points of view of other participants in the communication process [7: 23], which means that the researcher must spend a longer period in the field, must become part of the environment he/she is investigating, must become acquainted with the connections and activities and finally start understanding the environment and the participants [8: 4–5]. The researcher, on the one hand, must be close enough to the culture he/she studies in order to understand how it functions, while, on the other, must be capable of distancing him/herself enough in order to be able to write and report about it. Considering the involvement of the researcher in the situation, the problem of objectivity arises immediately and the only way to present the results in an objective way is to rely on the quantitative analysis, which in turn substantiates or refutes ethnographic conclusions.

The ethnography of the Internet is not necessarily connected with a physical journey, as was the case with earlier ethnographic research, where valid results could not be achieved with a certain amount of field work. On the contrary, ethnographic Internet studies focus on an experiential, not physical journey [8: 45], since one travels on the Internet by looking, reading, imagining and experiencing. The very lack of physical presence in the situation which is investigated introduces the element of unreliability of the source of information and the data which the researcher can gather and later analyze. This unreliability occurs due to the fact that it is impossible to check the information, that the real identities of communicators are not known to researchers and that the membership of a group or community is not constant.

3. Aims of Register Analysis

The aim of register analysis is to investigate the link between the linguistic expression and the social situation, with a view towards explaining that link [3: 3–12]. In the same manner, the analysis of the register of electronic chatting aims at exploring and explaining the following:

- how the characteristics of the situation in which CMC takes place change the language and to what extent;
- what linguistic and extralinguistic features make this register different from other registers;
- is there a connection between extralinguistic features (e.g. lack of face to face contact, keyboard and computer screen as the only ways of producing and receiving the information) and linguistic features (e.g. typos, innovativeness and brevity of expression).

Finally, the aim of register analysis is to identify and interpret the generalizations concerning register variations, as well as to systematically list linguistic and extralinguistic features of the register [2: 29–30]. A point of importance here is that the linguistic features must be interpreted not only in relation to the extralinguistic features of text production, but also in relation to other linguistic features, because the isolated analysis of a linguistic feature may lead to wrong conclusions. Thus, the use of personal pronouns of the first and the second persons singular, direct questions and imperatives indicate interactivity, without disclosing if they are found in a written or a spoken text, while abbreviated forms and self-corrections point to a spontaneous spoken discourse.

4. Conclusion

After presenting a number of reasons for compiling an electronic annotated CMC corpus, it must be said that this is a task that should be done in as many languages as possible. This is because the interlinguistic comparison and contrasting reveals explicit similarities and differences between two or more languages and may draw attention to phenomena that would otherwise be missed. Also, that kind of analysis would reveal any linguistic and cultural differences in one register across two or more languages, thus contributing to a better understanding of other languages and cultures.

References

1. Biber, D. (1988). Variation across Speech and Writing. Cambridge University Press. Cambridge.
2. Biber, D. (1995). Dimensions of Register Variation. A Cross-Linguistic Comparison. Cambridge University Press. Cambridge.
3. Biber, D. & Finegan, E. (1994). "Situating Register in Sociolinguistics". In: Biber, D. & Finegan, E. ed. Sociolingustic Perspectives on Register. Oxford University Press. Oxford. (3-12).
4. Crystal, D. (2001). Language and the Internet. CUP. Cambridge.
5. Danet, B.(2001). Cyberpl@y. Berg. Oxford, New York.
6. Ferguson, Ch. E. (1994). "Dialect, Register and Genre: Working Assumptions about Conventionalization". In: Biber, D. & Finegan, E. ed. Sociolingustic Perspectives on Register. Oxford University Press. Oxford. (15–30).
7. Hajmz, D. (1980). Etnografija komunikacije. [Foundations in Sociolinguistics. An Ethnographic Approach.] Beogradski izdavačko-grafički zavod, XX vek. Beograd.
8. Hine, C. (2000). Virtual Ethnography. SAGE Publications. London, Thousand Oaks, New Delhi.
9. Ooi, Vincent B. Y. (2000). "Aspects of Computer-Mediated Communication for Research in Corpus Linguistics". In: Peters, P., Collins, P. & Smith, A. ed. Language and Computers, New Frontiers of Corpus Research. Papers from the Twenty First International Conference on English Language Research on Computerized Corpora Sydney 2000. (91–104).

Р. Ющенко, В. Гудзенко*
Институт кибернетики им. В. М. Глушкова НАН Украины (Киев)
УДК 81’322.33

ПРИМЕНЕНИЕ ПАРАЛЛЕЛЬНЫХ КОМПЬЮТЕРОВ ДЛЯ АВТОМАТИЗАЦИИ МОРФОЛОГИЧЕСКОЙ РАЗМЕТКИ

Морфологическое аннотирование Национального корпуса предполагает значительные усилия, учитывая объем текстов, который составляет сотни миллионов слов. Большую часть рутины, однако, можно избежать, подключив к работе современные компьютерные технологии. Задача морфологической разметки поддается частичной формализации, а следовательно, процесс тегирования можно автоматизировать. В рамках этих работ Институтом украинского языка были

* © *Р.Ющенко, В.Гудзенко, 2006*

разработаны электронные морфологические словари словоформ украинского языка. Используя их, для большинства слов можно поставить в соответствие его морфологический разбор.

Каждый словарь образуется комбинацией части речи с ее словоформами [1]. Далее в таблицах приведены возможные словоформы частей речи. Здесь использованы условные обозначения для словоформ: род (m – мужской, f – женский, n – средний), число (p – единственное, s – множественное, t – pluralia tantum для существительных только множественного числа, d – двойственное), склонение (n – именительный, g – родительный, d – дательный, a – винительный, i – творительный, l – предложный, v – призывной), степень сравнения (1 – высшая, 2 – наивысшая), вид (f – совершенный, c – несовершенный, d – двувидовая форма), наклонение (a – изъявительное, m – повелительное, j – сослагательное), время (p – настоящее, t – прошедшее, u – будущее), лицо (1 – первое, 2 – второе, 3 – третье), состояние (a – активное, v – пассивное). Дефис в ячейке таблицы указывает на неопределенную форму соответствующей части речи.

Словоформа	Существительное	Прилагательное	Числительное
Род	m, f, n, -	m, f, n, -	m, f, n, -
Число	s, p, t, d	s, p, -	s, p, -
Склонение	n, g, d, a, i, l, v	n, g, d, a, i, l, v	n, g, d, a, i, l, -
Степень сравнения		1, 2	

Словоформа	Порядковое числительное и местоимение	Глагол	Инфинитив
Род	m, f, n, -	m, f, n, -	
Число	n, g, d, a, i, l, -		
Вид		r, c	r, c, d
Наклонение		a, j, m, z	
Время		p, t, u, c	
Лицо		1, 2, 3, -	

Словоформа	Причастие	Деепричастие	Наречие
Род	m, f, n, -	m, f, n, -	
Число	s, p, -		
Склонение	n, g, d, a, i, l, -		
Степень сравнения			1, 2
Вид	r, c	r, c	
Время	p, t		
Состояние	A, v		

Кроме вышеуказанных, в состав комплекса включены отдельные словари для союзов, предлогов и междометий.

Использование словарей для автоматический морфологической разметки приводит к проблеме эффективности разметки. Объем словарей составляет порядка миллиона словоформ. Учитывая, что для построения корпуса необходимо обработать сотни миллионов слов, актуальной становится скорость разметки каждого слова.

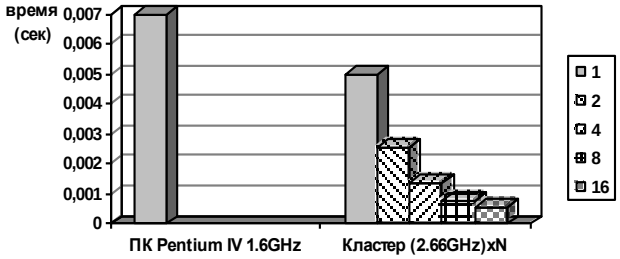
В Институте кибернетики по проекту Академии наук были разработаны кластеры СКИТ-1, СКИТ-2. Они представляют собой вычислительную систему из 16 и 32 компьютеров, называемых узлами кластера [5]. Высокая производительность достигается за счет того, что вычисления разбиваются на независимые блоки (декомпозиция задачи), которые считаются параллельно на разных узлах. Кластеры СКИТ-1 и СКИТ-2 являются открытыми ресурсами для институтов НАНУ, на них решалось множество научных задач, включая экономические, физические, химические. Доступ к кластерам СКИТ осуществляется с помощью Интернета, позволяя использовать его ресурсы находясь на расстоянии от территориального расположения Института кибернетики. Вычислительный ресурс кластеров СКИТ был применен для повышения эффективности автоматической разметки текстов для Национального корпуса украинского языка.

Для того чтобы извлечь пользу от кластеров, необходимо выполнить декомпозицию задачи автоматической разметки – определить, как эта задача сможет решаться

параллельно. Суть решения состоит в том, что входной текст разбивается на части, а каждый узел кластера обрабатывает (размечает) свою часть. После вычислений кусочки размеченного текста «склеиваются» в один размеченный текст. Для организации поиска словоформы в словаре использованы индексы в виде бинарного дерева поиска. Индексы хранятся в оперативной памяти, обеспечивая существенное ускорение обработки.

В результате применения такого алгоритма, была достигнута такая производительность:

Рисунок 1. Зависимость времени разметки одного слова от числа процессоров



Текст, полученный в результате разметки, представляет собой промежуточный материал для создания корпуса. Существуют словоформы, для которых нет соответствия словоформ в словаре (например, имена собственные), не все слова можно автоматически определить (например, сокращения), для некоторых словоформ существует сразу несколько соответствий (слова с одинаковым написанием используются в разном качестве). Сотрудниками Института кибернетики разработано программное обеспечение, позволяющее просматривать и править размеченные тексты, выставлять разметку для слов, разметка которых не существует в словаре, выбирать нужную разметку при наличии нескольких. В результате создается текст в специальной разметке, из который принципиально возможен импорт в любую разметку, которая определяется создателями корпуса.

На данный момент единственным правилом выставления соответствия словоформы ее морфологическому разбору является словарь словоформ. Это позволило корректно разметить 33% текста, взятого в качестве тестовой выборки. 45% слов содержали неоднозначности (для них существовало несколько вариантов разметки). Соответственно 22% слов не были найдены в словаре. Дальнейшим направлением исследований видится нам применением обучающих и статистических методов для формирования правил, используя которые можно уменьшить неоднозначность автоматической разметки.

Литература

1. Демская-Кульчицкая О. М., Семеренко В. Р., Ющенко Р. Р. Методы автоматической разметки текстов национального корпуса языка // Компьютерная математика. – К. – 2004. – С. 70–76.
2. Демська-Кульчицька О. М. Базові поняття корпусної лінгвістики // Українська мова. – 2003. – № 1. – С. 40–46.
3. Brill E. A Simple Rule-Based Part of Speech Tagger //In Proceedings of the DARPA Speech and Natural Language Workshop. – San Mateo, California: Morgan Kauffman, 1997. – P. 112–116.
4. Brill E. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of speech Tagging. //Computational Linguistics. – 1995. – v.21. – No 4. – P.122–132.
5. Суперкомпьютеры ИК НАН Украины // http://cluster.icyb.kiev.ua.

А. Загнітко, д. філол. н., ***Г. Ситар,*** к. філол. н.,
І. Данилюк, І. Щукіна*
Донецький національний університет (Донецьк)
УДК 811.161.2'344

ПРИНЦИПИ І ЗАКОНОМІРНОСТІ УКЛАДАННЯ БАЗИ "УКРАЇНСЬКІ ПРИЙМЕННИКИ ТА ЇХНІ ЕКВІВАЛЕНТИ"

The method of Ukrainian prepositions electronic processing, ways of differentiating between the proper-prepositions, the words, which are used in the value of prepositions, the words, which appear in the function of prepositions, the words which come only by analogy like prepositions, and the words which are used occasionally as prepositions, are reflected. The procedure of queries creation in a base is represented.

У рамках міжнародних проектів "Слов'янські прийменники в синхронії і діахронії: морфологія і синтаксис" і "Граматика категорії прийменника в російській та українській мовах. Загальнолітературна мова і діалекти" проблемною групою лінгвістів Донецького національного університету опрацьована концепція створення бази українських прийменників, що максимально охоплює їхній сучасний склад з урахуванням різних напрямів еволюції, динаміки й статики.

База даних прийменників української мови реалізована за допомогою програмних продуктів компанії Microsoft – Microsoft Word і Microsoft Access. У її сьогоднішньому вигляді база даних становить таблицю, що складається з 32 стовпців (полів), кожний з яких містить певну інформацію про прийменник, і 2028 рядків (записів), що відповідає кількості виявлених на сьогоднішній день прийменників та їхніх еквівалентів в українській мові. Для зручнішої роботи з базою була сконструйована так звана форма, що складається з текстових полів, полів зі звичайним списком і таким, що розкривається кнопками, іншими елементами управління і деяких вбудованих процедур автоматичної обробки даних.

Реєстр прийменникових одиниць різного типу формувався в кілька етапів. На першому був вивчений реєстр прийменників, зафіксованих в 11-томному Словнику української мови, у зіставленні з реєстромприйменників Словаря української мови за редакцією Б. Грінченка і реєстром прийменників Великого тлумачного словника української мови. Кістяк бази даних прийменників української мови побудований на основі трьох словників: "Словарь української мови" (за ред. Б. Грінченка, 4 т., 1907 р.; перевид.: К., 1996–1997), "Словник української мови" (11 т., 1970–1980 р.), "Великий тлумачний словник сучасної української мови" (2001). Насамперед були відібрані одиниці, позначені в словниках як прийменники. Потім були виділені лексеми, позначені як інші частини мови, але такі, які функціонують як прийменники, що простежується з наведених наприкінці словникової статті прикладів. На останньому етапі враховувалися одиниці неприйменникових частин мови, які потенційно можуть вживатися зі значенням або у функції прийменника. Пізніше ця гіпотетична здатність підтверджувалася відібраними прикладами такого вживання в спеціально створених корпусах художніх, наукових та офіційно-ділових текстів XX ст. (близько 1, 5 млн. слів у кожному корпусі). Крім того, максимально був врахований матеріал списків прийменників російської і білоруської мов, люб'язно надані проф. М. В. Всеволодовою (Москва) і проф. М. Й. Конюшкевич (Гродно).

Отже, кожен запис бази даних, що відповідає певному прийменнику, складається з таких полів (див. мал. 1.):

1. "Прийменник" – тут наводиться сам прийменник або прийменникова конструкція.
2. "Керування". У цьому полі відображається інформація про те, з якими відмінковими формами іменних частин мови або ж з інфінітивом поєднується прийменник. Якщо прийменник поєднується з декількома відмінковими формами (напр., *О* зі знахідним і місцевим відмінками = О + N_4/N_6), то кожній такій конструкції присвячений окремий запис (у базі це O_1 *і* O_2). Потреба створення окремого запису зумовлена нетотожними семантико-синтаксичними відношеннями, різною синонімією, антонімією, частотністю вживання і под.

* © *А. Загнітко, Г. Ситар, І. Данилюк, І. Щукіна, 2006*

3. "№" привласнюється значення 1, 2 або 3. Якщо прийменник не має розгалуженої сполучуваності, то поле одержує значення 0;

4. "Тип". У цьому полі визначається тип прийменника – первинний (*без, в, з-під, між, навпроти, під* і под.) або вторинний (*попереду, коло, край, у ролі* та ін.).

5. "Статус". Відповідно до класифікації можливих статусів прийменника, запропонованої проф. М. Й. Конюшкевич (Гродно), поле може набувати одного з п'яти значень.

5.1. "Прийменник". До власне-прийменників належать усі традиційно кваліфіковані прийменники типу *у, за, про* тощо.

5.2. "Зі значенням прийменника". Коло слів, вжитих зі значенням прийменник, утворюють такі, які мають відповідники (дублети) з-поміж власне-прийменників, але містять додаткові (з узагальненою конкретною семантикою, без образності) елементи: *у напрямку до = до, у процесі слухання = протягом, у стані облоги = в облозі, під облогою, під час засідання = на засіданні. Х*арактерною особливістю таких прийменників постає рівнорядність постпозитивної іменниково-відмінкової форми: *під час засідання кафедри = на засіданні кафедри* та ін.

5.3. "У функції прийменника" – це такі елементи, які можуть бути опущені (редуковані) без відповідної заміни, але значення виразу при цьому не змінюється: *розмовляв з дівчиною на ім'я Настя = розмовляв з дівчиною Настею*; *перебувають на грані зникнення = зникають*; *виступив у ролі миротворця = виступив миротворцем* і под. Сюди ж слід віднести конструкції, в яких є можливим додавання узгодженого означення типу *такий (а, е, і), цей (я, е, і)*: *у результаті обрання мене депутатом – у такому результаті, у ролі миротворця – у такій ролі, у разі вибуху – у цьому разі, у випадку розбіжностей – у такому випадку* тощо.

5.4. "Вживається за аналогією". Слова, що тільки за аналогією виступають прийменниками, – це досить обмежена група слів, основний масив якої складають структури з кількісним значенням тих чи інших вимірів (довжина, глибина, розмір, товщина, ширина, об'єм та ін.): *завдовжки понад, завдовжки від; завтовшки понад, завтовшки від; завширшки понад, завширшки від, з відстані (не) більш(е); з відстані (не) більш(е), (а)ніж; з відстані (не) більш(е), (а)ніж у/в; з відстані (не) менш(е); з відстані (не) менш(е), (а)ніж; з відстані (не) менш(е), (а)ніж у/в; з відстані від; з відстані від ... до; з відстані до; з відстані понад; з відстані у/в* та ін.

5.5. "Вживається оказіонально" – утворення, в яких супровідною постає певна образність і які використовуються в ролі прийменника в одиничних випадках: *на світанку людської історії, на горизонті сучасної науки* і под.

6. "Позиція". У цьому полі вказується, якою може бути прислівна позиція прийменникової одиниці: препозиція (*у сфері науки*); постпозиція (*рік тому, тиждень по тому*); препозиція / постпозиція (*заради слави / слави заради*); інтерпозиція (*один до одного*); рамка (*далеко я живу від університету*); препозиція / рамка (*заввишки у три метри / у три метри заввишки*).

У наступних полях репрезентована інформація про відображення прийменника в трьох зазначених словниках, у корпусі текстів і узагальнені відомості, повний обсяг яких буде отриманий на фінальному етапі створення бази. Зокрема, вказується, чи наявна прийменникова одиниця в словнику або тексті взагалі (відповідне поле може набувати значення "так" або "ні", тобто ставиться галочка), якщо так, то з якими значеннями, варіантами і синонімами, до якої з частин мови зараховують аналізовану прийменникову одиницю укладачі словника. Наприклад, у словнику Б. Грінченка слово *навпаки* визначається як прислівник, хоча з прикладу видно, що це безсумнівний прийменник: *Покотила паляницю навпаки сонця* (Г. Квітка-Основ'яненко) (Словарь, II: 472). Кожне значення ілюструється прикладом. Список значень наводиться в спеціальному полі, яке знаходиться над кнопкою "Додати". На сьогодні виділено 16 основних значень, що уявляється достатнім для реалізації завдання створення відносно вичерпного реєстру українських прийменників та переліку їхніх значень:

а) об'єктне (*без, в, в боротьбі з, в боротьбу за, до, з, під керівництвом, про* та ін.): *Скіфське плем'я арімаспів вступило з грифами в боротьбу за золото й коштовні камені, то були безстрашні варвари* (П. Загребельний. Диво);

б) адресатне (*до, для, на ім'я* тощо): *Що це важить супроти приємності бути поголеним! Хоч процедура часом і бувала болючою (особливо для тих, хто мав цупку*

рослинність), зате закінчення було радісним, а найбільш приємним був результат (І. Багряний. Сад Гетсиманський);

в) інструментальне (*в, за, на* і под.): *Дай заграю я на дудку, а то давно вже грав* (М. Драгоманов);

г) локативне *(біля, близько, ближче, в, в бік, в глиб, від, до, за, коло, край, на, над, під, після, при, у межах, у місці, у напрямі, через та ін.): Рано-вранці біля сільської Ради юртувалося десятків два підвід* (Г. Тютюнник. Вир);

ґ) міри і ступеня (*більше, заввишки, коло, мало не, менше, на, на... більше, навбільшки, навглибшки, об, по, у міру* тощо): *Сушені груші, сливи та яблука свої. Господи Боже – мало не два карбованці готовими грішми видалось, аж страх подумати – таке нечуване марнотратство* (У. Самчук. Волинь);

д) причини (*від, внаслідок, з, з огляду на, за, завдяки, зважаючи на, під впливом, у зв'язку з, у нагороду за, у надії на, у силу, через* і под.): *Таж, якщо матуся витримала стільки горя, що звалилося на її слабенькі плечі, то тільки завдяки своїй вірі і завдяки моральній підтримці* (І. Багряний. Сад Гетсиманський);

е) способу дії (*без, з, без допомоги, до, за допомогою, на, у розрізі, у супроводі, у такт, у тон, через* та ін.): *Найвідважніші з гуляк, розгойдуючись, злітали аж під покрівлю і, вигинаючись, без допомоги рук, намагалися відкусити від граната, калача або від баранини* (П. Загребельний. Роксолана);

ж) наслідку (*внаслідок, під дією, під владою, під впливом, у результаті,* тощо): *Маленька піщинка завжди залишається піщинкою, тоді як навіть могутні гори руйнуються під дією стихій, і що вищі вони, то тяжчі й страхітливіші руйнування на їхньому велетенському тілі* (П. Загребельний. Роксолана);

з) порівняльне (*подібно до, порівняно з, порівняно до, проти, у порівнянні з* і под.): *І вже вона не здається важкою в порівнянні з «Красою», навпаки, здається, що вони обидві мають щось надзвичайно спільне й рідне між собою* (В. Винниченко. Сонячна машина);

і) темпоральне (*в еру, в епоху, до, з ...до, за, напередодні, одночасно з, під час, під час перебування, пізніше, після, по, при, протягом, у добу, у процесі, у ході* тощо): *Мистецтво як форму духовно-релігійного діалогу людини з Богом (на основі концепцій активності мистецтва, які сформувалися в епоху середньовіччя в лоні християнського віровчення) розглядає М. Бровко* (Квасюк Л. В. Іконопис і словесність у культурі України XVI–XVIII століть: естетичний аспект: Автореф. дис. ... канд. філос. наук. – К., 2000);

к) умови (*за умови, при, при нагоді, у випадку, у разі* та ін.): *Не діждавшись нічого від своїх в'язнів, комендант так само байдуже дав час для роздумів до обіду, пригрозивши, що в разі мовчанки він розстріляє кожного десятого* (П. Загребельний. Диво);

л) допустове (*без огляду на, всупереч, наперекір, на противагу, незважаючи на, незалежно від, попри* і под.): *Він одривається від грат і ходить по темній камері, не звертаючи уваги ані на те, що камера, всупереч тюремним законам, зовсім не освітлена, ані на тяжке, насичене духом людського поту, цвіллю й смородом блощиць повітря* (І. Багряний. Сад Гетсиманський);

м) мети (*в ім'я, в інтересах, для, для блага, з метою, за, задля, заради, на, на честь, по, у пошуках, у справі* тощо): *Всякий акт, що робиться в інтересах цієї грандіозної мети, є акт святий* (В. Винниченко. Сонячна машина);

н) атрибутивне (*у/в, від під, після* та ін. – за умови вживання в присубстантивній позиції): *Маруся Чурай (історичний роман у віршах)* (Л. Костенко);

о) кількісно-атрибутивне (*від ... до, о, об, об'ємом, об'ємом (не) більш(е), об'ємом (не) менш(е), об'ємом (не) менш(е) за, об'ємом (не) менш(е) (а)ніж, об'ємом (не) менш(е) як, об'ємом від, об'ємом від ... до, об'ємом до, об'ємом з/із/зі, об'ємом понад, об'ємом у/в, обсягом, обсягом (не) більш(е), обсягом від, обсягом від ... до, обсягом до, обсягом з/із/зі, обсягом понад, обсягом у/в*): *Порушення цієї норми тягне за собою відповідальність у вигляді штрафу від 100 до 1000 неоподатковуваних мінімумів та компенсацію оператору ТМ понесених збитків та недоотриманого прибутку* (Про кабельне, ефірно-кабельне телебачення та телеінформаційні мережі. Закон України);

п) суб'єктне значення, яке у межах прийменникових одиниць є вторинним, оскільки відповідні конструкції є трансформами ядерних речень: *Прелюдія до скандалу пролунала у виконанні Адама Прайса, члена британського парламенту від націоналістичної партії Уельсу Plaid Cymru, – він роздобув копію листа Тоні Блера й запідозрив його в лобіюванні інтересів сталевого магната* (Ю. Загоруйко. "Стальгейт" чи "сміттєгейт"? // Дзеркало

тижня. – 2002. – 2–8 березня (№ 8 (383)) (*сп. дії+суб.*) ← *Адам Прайс виконанав прелюдію ...*

Окреме місце посідають предикативна й структурна функції. Щодо першої, то йдеться про регулярність вживання прийменника в присудковій позиції: *в ізоляції від, в опозиції до, в опозиції з/із/зі, у гонитві за, у залежність від, у зв'язок з/із/зі, у полоні, у претензії на, у ранзі.* Структурна функція є асемантичною, її роль зведена до поєднання лексем. Вживання прийменникових одиниць у структурній функції є поодиноким і складає периферію: *По матері вона черкешенка, тому намірилася розбудити в синові мужчину, знайшовши для нього гідну жону з свого племені* (П. Загребельний. Роксолана).

Стосовно поля "Тексти", то тут відображені значення і приклади вживання тих прийменників, які не були виявлені в словниках, але знайдені в корпусах текстів трьох функціональних стилів української мови, а також у мережі Інтернет.

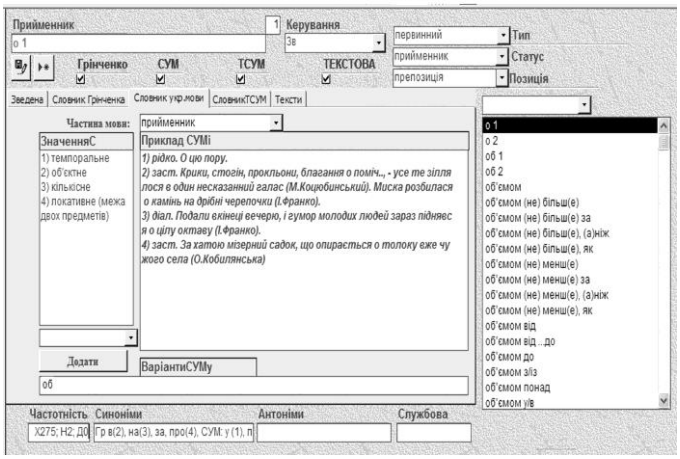
У полі "Варіанти" зазначена варіативність прийменникових одиниць у такому вигляді, як її подають словники: напр., прийменник *У1* (з Р.в.) має тільки один варіант – *В,* а *В2* (зі З.в.) – *В, ВІ, ВО, УВ*.

Поля "Синоніми" та "Антоніми" заповнювалися не тільки з урахуванням лексикографічних відомостей, але й на підставі спостережень проблемної групи над функціонуванням прийменників та їхніх еквівалентів у текстах різних функціональних стилів української мови: прийменникова сполука *аналогічно до* має синоніми *аналогічно (з/із/зі), відповідно (до / з/із/зі), за аналогією до, подібно (до)* та антоніми *на відміну від, на противагу (до), на протилежність.*

Запис типу «Х5; Н13; Д17» у відповідному полі відображає абсолютну частотність у корпусах художніх текстів (твори класиків української літератури XIX і XX століть, а також сучасних письменників), наукових текстів (переважно авторефератиі дисертацій з різних галузей знань), а також текстів сучасного офіційно-ділового стилю (кодекси, закони, постанови, накази і под.) загальним обсягом біля 4500000. У випадку відсутності ілюстративного матеріалу в текстовій базі (тоді абсолютна частотність дорівнює нулю) були використані приклади з інших джерел, зокрема, сучасної української періодики, а також ресурси Інтернету (як джерело вказується повна адреса відповідної веб-сторінки).

Рисунок 1.

Загальний вигляд вікна електронної бази даних українських прийменників



Стосовно поля "Тексти", то тут відображені значення і приклади вживання тих прийменників, які не були виявлені в словниках, але знайдені в корпусах текстів трьох функціональних стилів української мови, а також у мережі Інтернет.

У полі "Варіанти" зазначена варіативність прийменникових одиниць у такому вигляді, як її подають словники: напр., прийменник *У1* (з Р.в.) має тільки один варіант – *В,* а *В2* (зі З.в.) – *В, ВІ, ВО, УВ*.

Поля "Синоніми" та "Антоніми" заповнювалися не тільки з урахуванням лексикографічних відомостей, але й на підставі спостережень проблемної групи над функціонуванням прийменників та їхніх еквівалентів у текстах різних функціональних стилів української мови: прийменникова сполука *аналогічно до* має синоніми *аналогічно (з/із/зі), відповідно (до / з/із/зі), за аналогією до, подібно (до)* та антоніми *на відміну від, на противагу (до), на протилежність.*

Запис типу «Х5; Н13; Д17» у відповідному полі відображає абсолютну частотність у корпусах художніх текстів (твори класиків української літератури XIX і XX століть, а також сучасних письменників), наукових текстів (переважно авторефератидисертацій з різних галузей знань), а також текстів сучасного офіційно-ділового стилю (кодекси, закони, постанови, накази і под.) загальним обсягом біля 4500000. У випадку відсутності ілюстративного матеріалу в текстовій базі (тоді абсолютна частотність дорівнює нулю) були використані приклади з інших джерел, зокрема, сучасної української періодики, а також ресурси Інтернету (як джерело вказується повна адреса відповідної веб-сторінки).

База супроводжується певним набором запитів, що дозволяють одержувати узагальнену інформацію, а також елементи навігації у вигляді активного списку (див. рис. 2 і 3).

Рисунок 2. Відсортований за абеткою активний навігаційний список для пошуку прийменникової одиниці у базі

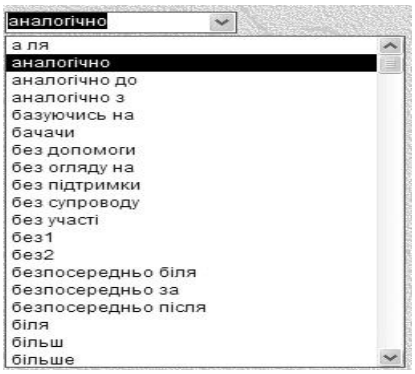
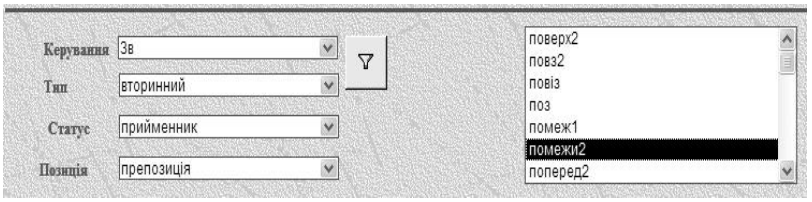


Рисунок 3. Частина макету форми для роботи із запитами. Праворуч – відсортований за абеткою активний навігаційний список із результатами запиту.



При укладанні бази “Українські прийменники та їхні еквіваленти” визначальними є структурна типологія, врахування парадигматичних ресурсів прийменника і простеження його синтагматики. Урахування цього дає підстави для цілісного відображення прийменника зі встановленням відповідних функціональних виявів. Особливу вагу має

встановлення граничності прийменника, умов його внутрішнього членування і впливу на семантику як внутрішніх, так і зовнішніх показників.

Література

1. Берко А., Верес О. Організація баз даних: практичний курс. – Львів: Національний університет «Львівська політехніка», 2004. – 152 с.
2. Вихованець І. Р. Прийменникова система української мови. – К.: Наукова думка, 1980. – 288 с.
3. Всеволодова М. В., Клобуков Е. В., Кукушкина О. В., Поликарпов А. А. К основаниям функционально-коммуникативной грамматики русского предлога // Вестник Московского университета. – Серия 9. Филология. – М.: Изд-во МГУ. – 2003. – № 2.– С.17–59.
4. Карпіловська Є. А. Вступ до комп'ютерної лінгвістики. – Донецьк: ТОВ "Юго-Восток, Лтд", 2003. – 184 с.
5. Українські прийменники: Синхронія і діахронія: пробний зошит / За ред. проф. А. П. Загнітка. – Донецьк: ДонНУ, 2003. – 162 с.

А. Карчова, М. Шимкова*

Інститут мовознавства імені Л. Штура Словацька академія наук
(Братислава, Словацька республіка)

УДК 81'322.3

МОРФОЛОГІЧНА АНОТАЦІЯ ТЕКСТІВ СЛОВАЦЬКОГО НАЦІОНАЛЬНОГО КОРПУСУ

Abstract. *From 2004 to 2005, four different kind of texts have been manually disambiguated and morphologically annotated: Orwell's 1984 novel, selected Slovak Wikipedia texts, Internet magazine InZine and SME daily newspaper articles. Empirically estimated precision of manual annotation is presented, together with description of the annotation process. The result is high quality manually annotated corpus of Slovak language texts, containing over 300 000 words.*

Одночасно із першими кроками побудови Словацького національного корпусу (СНК) в 2002 – 2003 рр. розпочалася підготовка морфологічної анотації текстів. Паралельно зі збиранням текстів та їх первісним опрацюванням формувалася внутрішня структура корпусу, йдеться про технічний аспект, та здійснювалася стильово-жанрове опрацювання текстів. Важливою справою було придбання придатного корпусного програмного засобу чи менеджера. Не менш важливим було також запитання: розпочинати морфологічну анотацію чи ні, і якщо так, то яку саме?

На момент початку робіт над створенням СНК у Словаччині комп'ютерна або корпусна лінгвістика як галузь науки не були достатньо розбудованими і ще не існували відповідні до морфологічної анотації програмні забезпечення, не кажучи вже про лінгвістичні описи, які можна було б використати для комп'ютерного опрацювання словацької мови. У той же час у світі ті, хто створювали та використовували електронні корпуси, щораз більше орієнтувалися на лінгвістичні анотації корпусів, на відміну від минулого, коли корпуси будували на взірець комп'ютерних баз даних слів чи текстів із майже виключно лексикографічним призначенням, а лінгвістичні інформації, які вносилися в корпуси, вважалися зайвими через змінність граматичних категорій. Лінгвістичне анотування корпусів стимулювало появу знаряддя для морфологічної анотації флективних мов. Таке ж знаряддя безпосередньо для словацької мови було створено на Фізико-математичному факультеті Карлового університету в Празі, і виникло як другорядний продукт знаряддя на лематизацію, аналіз та зняття неоднозначності для чеської мови, яке застосовували в Чеському національному корпусі.

На першому етапі побудови Словацького національного корпусу – етапі підготовки морфологічної анотації текстів, – було здійснено об'єднання вже наявних підходів до морфологічних анотацій корпусних текстів [3], проведено тестове перевіряння чеського знаряддя, яке для нашого відділу СНК віддала на підставі договору вже згадувана інституція, а також вирішено, яким чином здійснюватиметься анотація текстів СНК.

На підставі тестового аналізу празького програмного забезпечення було виявлено, що:

а) програмне забезпечення, створене на початку 90-х років у межах проекту Чеського національного корпусу SYN 2000, було доволі часто критиковане за надто строгий технічний підхід; за відсутність мітки для категорії виду у дієслова, яка є важливою специфічною ознакою більшості слов'янських дієслів; під критику підпав також

* © *А. Карчова, М. Шимкова, 2006*

позиційний спосіб запису морфологічних знаків, який є важкий для читання та розшифровування, і наявність помилок при анотуванні за статистичним принципом, наприклад, часто вживане чеське слово, яке функціонує як займенник або сполучник *jak* було переважно визначуване як іменник зі значенням «назва тибетського тура» тощо;

б) словацьке знаряддя для анотації, яке по суті виникло як другорядний продукт, базувалося на основі словника, який не був цілком коректний, а саме тому помітно знижувалася успішність анотації;

в) словацька мовна система та словацька граматична теорія в порівнянні із чеською мовою має кілька суттєвих розбіжностей.

Виходячи зі сказаного, ми дійшли до таких висновків: чеське знаряддя для анотації буде використане в першій фазі на автоматизоване попереднє анотування словацьких текстів та для внутрішніх досліджень, поки не буде створено власного анотаційного знаряддя, а також вручну анотований корпус достатнього розміру. Таке рішення означало підготовку власного таґсету (правил токенізації, лематизації та маркування), який розпочали розробляти приблизно в середині 2003 р.

Засадничим обрано формально-морфологічний підхід, скорочений – атрибутивний запис, у якому кожний тип мітки має свою власну стабільну позицію, і звичайний поділ слів на окремі частини мови. Словацький теґсет, який був захищений в кінці 2003 р. нашими лінгвістами в рамках Словаччини, а також на міжнародній конференції SLOVKO 2003 [3; 4; 5], містить дев'ятнадцять класів слів, з яких десять відповідає звичайним частинам мови а дев'ять інших класів слів позначає специфічні мовні елементи, наприклад інтерпункцію, скорочення та ремарки, цитатні звороти запозичені з інших мов, цифри, елементи, які не входять до категорії слів. Окремо відділяються формальні дієприкметники типу *napísaný, prijatý*, що з одного боку ускладнює анотацію на морфологічному рівні, але з іншого боку є вдалою вихідною точкою для подальшого опрацювання текстів. Навпаки певним полегшенням є спосіб інтерпретації слів *sa*/*si*, які мають різні значення і різні функції в системі морфології. За нашими правилами вони завжди вважаються зворотними займенниками і мають спеціальну мітку «R». Специфічною рисою теґа в іменних частинах мови є встановлювання типу парадигми на другій позиції мітки: окремо так відзначаються іменники з прикметниковою парадигмою (*obžalovaný, vedúci* – частина мітки «SA»), прикметники із не повною парадигмою (*super* – частина мітки «AF») тощо. Наш теґсет можна знайти за адресою http://korpus.juls.savba.sk/.

Другий етап (лютий 2004 р. – жовтень 2005 р.) був етапом ручної анотації з використанням словацького теґсету, розвитку та модифікації анотаційних знарядь. Паралельно відбувалося удосконалення в теґсету, а саме, вносилися до нього необхідні зміни та доповнення.

Першим текстом обраним для анотації став переклад новели Джорджа Орвелла «1984». Таке рішення мотивовано двома чинниками, по-перше, ця новела була частиною міжнародного проекту Multex-East, до якого ми долучилися, і, по-друге, існував чеський сегмент проекту Multex-East з вдало зробленою ручною анотацією, який за необхідності можна було б використати для порівняння і тестової перевірки можливостей автоматизованої анотації [8].

Анотацію вручну виконували студенти факультетів словацької мови університетів в містах Пряшів, Ружомберок та Братислава. Але анотації зроблені під час навчання не дали достатніх результатів, вдалішим та успішнішим виявилося індивідуальне анотування обраними студентами у відділенні СНК поза навчанням, тобто студенти працювали на підставі угоди про виконану роботу і за свою роботу одержували відповідну платню. Анотовані тексти ще проходили фазу остаточного виправлення, тому текст анотований та виправлений вручну був готовий аж у середині 2005 р.

Після анотації перекладу новели Джорджа Орвелла «1984», яка містить 102 000 токенів, ми вибрали для морфологічної ручної анотації тексти із щоденної газети SME, інтернетного журналу InZine та словацької версії інтернетної енциклопедії Wikipedia. Таке рішення мотивовано тим, щоби вручну утворений анотований корпус складався із текстів різноманітних стилів, оскільки його призначення використовуватися для тренування маркування великого корпусу. У червні, липні та серпні 2005 р. було здійснено морфологічну анотацію текстів обсягом приблизно 115 000 токенів, а саме тексти SME разом містили 21 268 токенів поділених до 51 файлів, InZine містив разом 42 703 токенів в 51 файлах та Wikipedia 50 077 токенів в 48 файлах.

На початковому етапі із згадуваних джерел були вибрані ті тексти, які були найоптимальнішими для анотування, тобто такі, які були написані як одне ціле, словацькою мовою, з оптимальним числом слів і ці тексти були поділені на окремі файли. Файл містив найменше 100 токенів, максимальна кількість токенів у файлі становила 6 500. Найкращою з погляду зосередженості на анотації та організації процесу анотування виявилася довжина файла приблизно 1 500 токенів.

Відібрані тексти були перетворені на так звані анотаційних файли, які відповідали XML-стандарту. У нашому відділенні була створена спеціальна GUI-програма написана за допомогою програмної мови Python, котра використовує для розбору XML-файли ElementTree бібліотеки. Програма зображає реєстр усіх слів у файлі і для кожного обраного слова список відповідних лем та міток. Анотатор має можливість обрати до токена відповідну пару лема-мітка, або самостійно доповнить відповідну пару. Цей метод попереднього позначення текстів великою мірою допомагає анотаторам і пришвидшує їх роботу (≈300 токен година$^{-1}$).

Це анотування здійснювало десять студентів Філософського факультету Університету імені Коменського в Братиславі. Переважна більшість файлів опрацьовувалася двічі, тобто той самий файл був окремо анотований двома студентами, щоб можна було пізніше порівнювати між собою дві анотації. Як правило анотатори робили помилки в приписувані міток, інколи лема і мітка, що відповідали одному токенові, були помилковими. Помилки спричиняло також недостатність знань про внутрішні закономірності граматичної системи словацької мови, а також внутрішньомовні неоднозначності, наприклад не цілком ясна межа при встановленні незмінних частин мови. Студенти також повинні були достосувати свої знання до обраного нами методу, який в деяких пунктах відрізняється від класичної морфології, а деякі правила доповнюють її, тому анотування виявилося для студентів не надто простою справою. Важливим фактором появи помилок були і інші обставини, наприклад втома анотаторів – анотація є працемісткою роботою, яка вимагає постійного зосередження уваги.

Часто помилковим було встановлення відмінка іменника у випадках, коли йшлося про омоформи, інколи при дієсловах був неправильно визначений знак афірмації або негації, багато разів помилково визначався вид дієслів, часто сплутувався числівник із займенником, частка з прислівником, помилково встановлювався рід іменників та тип парадигми змінюваних частин мови. У мітках інколи бракувало обов'язкової частини, наприклад встановлення ступеня порівняння у прикметниках та дієприкметниках тощо.

Таблиця 1. A_1 – перший анотатор, A_2 – другий анотатор, Б – остаточно виправлений текст

Доку-менти	**К-сть токенів**	**К-сть словес-них еле-ментів**	**порівняння**	**К-сть помилок [%]**		
				лема	**мітка**	**обидва**
SME	21 268	77, 35 %	A_1 – Б	0, 87	7, 87	0, 33
			A_2 – Б	1, 08	7, 21	0, 40
			A_1 – A_2	1, 73	11, 45	0, 63
WIKI	50 077	82, 86 %	А – Б	1, 01	6, 74	0, 50
INZINE	42 703	82, 01 %	A_1 – Б	1, 23	9, 40	0, 51
			A_2 – Б	1, 68	11, 07	0, 96
			A_1 – A_2	2, 52	14, 56	1, 36

Зважаючи на велику кількість помилок ми вважали необхідним додатково виправляти оці помилки. Перевірка текстів, здійснювана протягом серпня – жовтня 2005 р.,

складалася з трьох окремих фаз. За допомогою наших власних напівавтоматизованих знарядь було можливе швидке та надійне виявлення й виправлення більшості помилок.

У першій фазі були знайдені такі помилки: мітка не мала правильної довжину (не була складена з усіх обов'язкових знаків – деякого знаку бракувала, а деякий був зайвим), у лемі фіксувався пробіл між буквами або велика букву (лема повинна складатися із малих букв), у мітці була неправильна комбінація декотрих знаків (не правильне встановлення роду, числа тощо). Результати пошуку помилок були зображені в наочних таблицях.

Таблиця 2. Приклад результатів із програми, яка виконуює порівняння

номер рядка	**токен**	**лема**	**мітка**	**опис помилки**
126	nejakej	nejaký	PAfs6x	Bad length
267	"	"	Z	Spaces in lemma/orth
503	bola	byť	VLesc+	Bad length
651	mi	ja	PPms3	Bad gender
1333	tučné	tučný	AAfp4	Bad length
1456	sú	byť	VKefp+	Bad number
2500	zažili	zažiť	VLdpbm+	Bad gender
2932	—	—	Z	Spaces in lemma/orth
2990	niekoľko	niekoľko	PU	Bad length
3057	deviatich	deväť	NUip2w	Bad length
3332	ich	on	PPmp4	Bad gender
5856	pohrá	pohrať	VKmsc+	Bad aspect
6057	slúžia	slúžiť	VKepci+	Bad length
7818	II	II	C}---------------	Bad POS
11637	,	,	None	Not string

Друга фаза виправлення полягала в порівнянні відмінностей між текстами, анотація яких була виконана незалежно один від одного двома різними анотаторами. Таким чином виявилися всі випадки, у яких студенти не узгоджувалися у встановленні леми або мітки. Також були знайдені важливі помилки, які випливали із недостатнього порозуміння правил теґсету. Виявлено велику кількість неправильних інтерпретацій, але не всі, тому що в декотрих випадках помилялися обидва анотатори.

У третій фазі знов йшлося про порівняння двох текстів. У цей час ми уже мали і могли використовувати остаточно виправлений текст Дж. Орвелла «1984». На підставі порівняння триґрам токен – лема – мітка до таблиці були виписані всі відмінності між текстами, а не тільки неправильні, але й такі, яких у тексті новели не виявилося. Словниковий запас слів новели є само собою обмежений – лише 40 % триґрамів знаходилося в наступних текстах, тому ми ще остаточно розподіляли мітки за частинами мови, щоб читання інформацій в таблицях прискорилося.

Після закінчення цих фаз перевіряння всіх текстів було зреалізовано і особлива увага приділялася так званим проблемним словам, яким виявилися частки, сполучники, дієприкметники, номери, частини текстів іншомовного походження, у яких ще інколи знаходилися помилки і, відповідно, які були остаточно виправлені.

На початок листопада 2005 р. був відкритий доступ до нової версії СКН prim-2.0 (первісний корпус), який є більшим за кількістю текстів (понад 250 мільйонів токенів), а також першу версію автоматизованої морфологічної анотації словацькими мітками.

З погляду комп'ютерного опрацювання (йдеться про меншу гібридність лематизації та розпізнавання омонімів) було використано первісне чеське програмне знаряддя для маркування мітками TNT [1], натреноване на вище згадуваному тренувальному корпусі анотованому вручну обсягом ≈130 000 токенів. За інформаціями автора цього знаряддя, для статистичної анотації на базі наявності триґрамів у наперед анотованому зразку, зреалізованому на англійських та німецьких текстах, це знаряддя при 100 тисячному тренувальному корпусі досягає успішність 95 %, при 200 тисячному корпусі успішність 96 % і т.д. Правда, словацька мова типологічно відрізняється від німецької та англійської мови і успішність знижується передусім через високу частоту омонімічних форм, тому успішність першої версії автоматизованої анотації була приблизно 90 %.

Перші фільтри показали, що в автоматизованих анотаціях повторюються переважно помилки подібні до тих, які з'являються у анотованих вручну текстах, наприклад, не правильно позначається відмінок у омонімічних формах, помилковим є встановлення мітки для істот та неістот в іменниках чоловічого роду, не правильно встановлюється вид дієслів, не розрізняються формальні дієприкметники та прикметники. Окремі проблеми випливають із неоднозначності займенників, а також встановлення афірмації та негації в дієсловах та ступінь порівняння прикметників та прислівників, проблеми випливають із частої омонімії незмінних частин мови – ці слова краще анотувати вручну, за допомогою списку слів укладених за окремими ознаками, наприклад, займенники із іменниковою парадигмою, окремою займенниковою, змішаною, не повною парадигмами та прислівниковою парадигмою, і також списків виключних – неомонімічних часток, прийменників тощо.

Таблиця 3. Приклад результатів порівняння анотацій між двома анотаторками

1	Zakázaný	zakázaný	AAis1x
	Zakázaný	zakázaný	Gtis1x
7	BRATISLAVA	bratislava	SSfs1:r
	BRATISLAVA	bratislava	SSfs1
23	osobných	osobný	AAip2x
	osobných	osobný	Gtip2x
40	ho	on	PFis4
	ho	jeho	PUis4
54	autora	autor	SSms2
	autora	autora	SSms2
57	B	b	Q
	B	B	Wms1
59	Manna	mann	SSms2:r
	Manna	Mann	SSms2:r
64	je	nebyť	VKesc-
	je	byť	VKesc+
74	SAV	sav	W?
	SAV	sav	Wfs2
95	aj	aj	O
	aj	aj	T

Під час автоматизованої анотації ще додалося технічних помилок, які були спричинені:

а) різницею обраних підходів: за словацькою лематизацією для дієслова в заперечній формі була встановлена лема в заперечній формі; власні імена, хоча починаююся з великої букви, лематизуються з малою початковою буквою; дієприкметники лематизуються як прикметники, а не як дієслова. У корпусі анотованому вручну ці правила повинні бути дотримані, в корпусі, який був лематизований та анотований автоматично, леми встановлюються за чеським знаряддям, а тому в наведених випадках знаходяться відмінності, що інколи спричиняє також помилку в мітці;

б) не достатнім виявився розмір тренувального корпусу, в якому не виявилося, наприклад, частки *azda*, і тому була знаряддями визначена її належність до будь-якої частини мови, так само вигуки позначалися довільно. Наприклад до власного іменника-прізвища *Štúr* (Л. Штур був основоположником словацької літературної мови) була додана лема дієслова *štúrať* тощо; менш зрозумілими помилками були не системно призначувані мінки чи інші граматичні категорії в триґрамах, наприклад, в іменному складі слів *na maďarský národný štát* прийменнику був приділений 6. відмінок, першому прикметнику 1. відмінок, другому прикметнику та іменнику 4. відмінок.

Третій етап морфологічної анотації текстів СНК, що власне зараз відбувається, передбачає поповнення анотованого вручну корпусу виправленнями вибраних текстів автоматично анотованих таким чином, щоб він досяг розміру найменше 1 мільйон токенів. З середини лютого 2006 р. вже існує відкритий в Інтернеті підкорпус анотований вручну з обсягом 322 600 слів, що представляє 56 465 окремих словесних форм, 26 498 лем та 1 428 різних міток, на якому знов тренувалася автоматизована анотація, результатом чого була нова доступна версія автоматизовано лематизованого корпусу prim-2.1.

Загалом, крім використання нашого корпусу для наступного опрацювання словацької мови, якісна морфологічна анотація може використовуватися, а в цьому часі вже й використовується, для граматичних досліджень, зокрема для створення нової морфології словацької мови. Тексти анотовані вручну далі передаємо на синтаксичну анотацію, яка цілком реалізується за принципами Празького корпусу.

Література

1. Brants, T.: TNT – Statistical Part-of-Speech Tagging
2. URL: http://www.coli.uni-saarland.de/~thorsten/tnt
3. Forróová, M., Horák, A.: Morfologická anotácia korpusu. In: Slovenčina na začiatku 21. storočia. Ed. Mária Imrichová. Prešov : Prešovská univerzita, Fakulta humanitných a prírodných vied, 2004, s. 174 – 186.
4. Garabík, R., Gianitsová, L., Horák, A., Šimková, M.: Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu. h
5. Garabík, R., Gianitsová-Ološtiaková, L.: Manual Morphological Annotation of Slovak Translation of Orwell's Novel 1984 – Methods and Findings. In: Proceedings of International Conference Slovko 2005 – Computer Treatment of Slavic and East European Languages. Ed. R. Garabík. Bratislava : Veda, 2005.
6. Hajič, J.: Popis morfologických značek – poziční systém. h
7. Hajič, J., Krbec, P., Květoň, P., Oliva, K., Petkevič, V.: Serial Combination of Rules and Statistics: A Case Study in Czech Tagging. Prague Dependency Treebank. CD ROM. V. 1.0. Praha : Ústav formální a aplikované lingvistiky MFF UK, 2001.
8. Hlaváčová, J.: Orwell's 1984 – Playing with Czech and Slovak Versions. In: Proceedings of International Conference Slovko 2005 – Computer Treatment of Slavic and East European Languages. Ed. R. Garabík. Bratislava: Veda, 2005.

И. Лучкув, к. филол. н.*
Вроцлавский университет (Вроцлав, Польша)
УДК 81'33

ПОЛЬСКАЯ ЛИНГВИСТИЧЕСКАЯ ЛИТЕРАТУРА XIX – НАЧ. XX ВВ. В ЦИФРОВОМ ФОРМАТЕ. НА ПУТИ К СОЗДАНИЮ БИБЛИОГРАФИИ ПОЛЬСКОЙ ЛИНГВИСТИКИ В ФОНДАХ ИНТЕРНЕТ-БИБЛИОТЕК

Настоящая статья имеет сугубо информационный характер. Основная ее цель – обзор работ польских лингвистов (в основном XIX – нач. XX вв.), которые в цифровом формате

* © *И. Лучкув, 2006*

уже имеются в фондах польских интернет-библиотек, и в общих чертах определить возможность постоянного пополнения этого списка.

За последние годы общепризнанным каналом распространения разнообразной, в том числе научной, информации стал Интернет. Рядом с традиционными печатными изданиями книг и журналов все чаще появляются их электронные варианты, которые становятся столь же обычными, какими еще не так давно были «бумажные». Что же касается «классики», которая появилась в доинтернетовские времена, то в последние годы начался процесс восполнения этого пробела, т. е. перевод книг и журналов в цифровой формат. Дигитализация документов – это не только современный способ сохранения имеющихся культурных ценностей на длительный срок. Благодаря этому исследовательские институты, библиотеки и государственные учреждения могут предоставить доступ к своим архивам, в том числе лингвистическим, а цифровые библиотеки служат своеобразным промежуточным звеном между прошлым, настоящим и будущим.

Прежде чем перейти к краткому обзору лингвистической литературы в польских цифровых библиотеках, необходимо в общих чертах охарактеризовать их современное состояние. В поле нашего зрения остаются только бесплатные библиотеки открытого доступа, представляющие единую базу данных. В Польше они развиваются в рамках двух основных проектов. Во-первых, это финансируемая в рамках правительственной программы Polska Biblioteka Internetowa (www.pbi.edu.pl – [PBI]), а во-вторых, проект отдельных региональных цифровых библиотек, разрабатываемый на основе библиотечных фондов в крупных вузовских центрах. Все эти библиотеки работают с учетом авторского права, т. е. в них публикуются книги, изданные до 1945 г., или же с авторами подписаны договора. Основная цель обоих проектов – предоставить бесплатный доступ к культурным ценностям самому широкому кругу пользователей – ученым, студентам, ученикам и всем, кто захочет познакомиться с материалами по истории, географии, культуре, с текстами художественной литературы. Это отсканированные полнотекстовые версии книг, журналов, архивных документов, географических карт, портретов с факсимильными подписями и многое другое.

Самая крупная цифровая библиотека общего характера – Polska Biblioteka Internetowa [PBI] – w. (в ней представлены, между прочим, тексты художественной литературы -- в том числе переводы иностранной литературы на польский язык, -- публицистика, энциклопедии, словари, научная литература, архивные документы, журналы и газеты (март 2006 – 29243 файлов). Отдельные документы можно читать при прямом подключении к Интернету. Пока нет возможности скачивать книги на жесткий диск. В будущем предусмотрена возможность платного скачивания.

В отличие от «государственной», PBI региональные цифровые библиотеки [РЦБ] появились как инициатива снизу общественности библиотекарей. РЦБ развиваются на основе библиотечных фондов крупных вузовских центров. В октябре 2002 г. первой открылась Wielkopolska Biblioteka Cyfrowa [WBC] (www.wbc.poznan.pl) в Познани, в декабре 2005 г. Kujawsko-Pomorska Biblioteka Cyfrowa [KPBC] (http://kpbc.umk.pl) в Торуни и Быдгощи. Проект постоянно расширяется, и к нему подключились университетские центры во Вроцлаве [отдельно университет (http://www.bibliotekacyfrowa.pl/dlibra) и политехнический институт, к которому будут подключены фонды библиотеки Оссолинеум (http://dlib.bg.pwr.wroc.pl/dlibra)], Кракове (http://mbc.malopolska.pl), Лодзи (http://ebipol.p.lodz.pl), Зеленой Гуре (http://zbc.uz.zgora.pl).

В РЦБ нет практически текстов художественной литературы, зато особое внимание уделяется широко понимаемой «региональной» тематике. Это и тексты про регион, и изданные в данном регионе, и находящиеся в традиционных библиотеках. Кроме того, РЦБ постепенно пополняют фонды учебных пособий, которые составят будущий фонд литературы для дистантного обучения (e-learning). Все библиотеки работают в одной системе – dLibra (Digital Library Framenetwork). Документы записываются в формате DjVu, и , в отличие от PBI, все тексты можно скачать на жесткий диск.

Переходя к описанию лингвистических фондов польских интернет-библиотек, необходимо отметить, что найти в них интересующие меня работы – довольно сложный процесс. Это обусловлено тем, что поиск ведется практически только по фамилиям авторов или по заглавиям работ. Пока нет тематических рубрикаторов, по которым можно искать лингвистические тексты. А там, где они есть, оказывается, что лингвистика попадает под рубрику «литературоведение» (literaturoznawstwo), которую на самом деле следовало бы назвать «филологией» [KPBC]. К тому же, нет единой программы, по которой строится

фонд польской цифровой лингвистики. Поэтому и трудно собрать то, что уже имеется. Прилагаемый к докладу список литературы – это ок. 40 % того, что мне удалось уточнить до первых дней марта 2006 г. на трех основных сайтах – PBI, WBC и КРВС. Дальнейшее пополнение списка лингвистических текстов, появляющихся в цифровых библиотеках, требует постоянного пристального внимания. На главной странице сайта каждой РЦБ дается сноска на «новые поступления» (последних 25) и «ближайшие планы». Зарегистрированным пользователям на адрес e-mail пересылают информацию о новых поступлениях в библиотеку. Конечно, в такой информации есть все новые поступления (в том числе, напр., отдельные номера газет), но беглый просмотр такой информации позволяет отобрать то, что нас интересует. К сожалению, в PBI не предусмотрены такие возможности. После пополнения фонда поиск надо начинать опять по фамилиям и заглавиям.

Приложение представляет собой своего рода библиографию польской «интернет-лингвистики». Краткая характеристика Приложения:

1) в польских интернет-библиотеках в настоящее время представлены главным образом лингвистические работы XIX – нач. XX вв. Кроме первых учебников по грамматике польского языка О. Копчинского [41–46], в Приложении не учитываются школьные учебники, работы по стилистике и культуре речи, по методике преподавания польского языка, а также двуязычные словари, за исключением словаря Г. Кнапского (1717) [40];

2) словари и работы по лексикографии и лексикологии: словарь польского языка С. Линде [60], первый польский словарь синонимов А. Красиньского [56], кашубский словарь Л. Бискупского [17], диалектологические словари Я. Карловича [37], З. Глогера [23], словари заимствований в польском языке Я. Карловича [38] и Ш. Моравского [75]. Кроме того есть замечания к проекту словаря польского языка (т. наз. «варшавского») Я. Карловича [36], статьи посвященные анализу древнейших словарей польского языка Г. Лопацинского [62] и химической терминологии Я. Бодуэна де Куртенэ [1];

3) ономастика: особо надо выделить ряд работ по ономастике Я. Карловича [28; 30; 31; 33; 34] и С. Козеровского [47–55]. В ближайшее время в КРВС должны появиться работы Я. Розвадовского;

4) историческая грамматика: этот раздел языкознания довольно хорошо представлен. Тут и работы общего характера [14; 19; 20; 26; 27; 65; 66; 67; 69; 74], и анализ памятников [11; 63; 71];

5) диалектология: пока диалектология представлена довольно скромно [18; 23; 29];

6) польский язык на фоне других языков: сравнительно-исторический метод [16; 72; 76];

7) грамматика польского языка: Общая структура языка представлена в учебниках О. Копчинского [41–46] и А. Бема [15], а также в работе Я. Бодуэна де Куртенэ [10]. Ряд работ посвящен синтаксису [21; 24; 57] и фонетике [2; 8], а только одна морфологии [22]. В планах КРВС есть дигитализация ряда работ по фонетике Т. Бенни;

8) орфография: работы по орфографии представлены довольно хорошо [2; 32; 39; 58; 64; 68; 70; 73; 77]. На их основе можно проследить становление нормы польской орфографии, начиная с предложений Й. Кохановского и Л. Гурницкого. Оба эти предложения сопоставляются в работе Й. Янушовского *Nowy karakter polski* (1594) [25]. Конечно, вопросы орфографии появляются также во многих школьных учебниках польского языка;

9) работы по общему языкознанию Я. Бодуэна де Куртенэ [4; 5; 6; 7; 9; 11; 13];

10) представленные в статье лингвистические материалы – это только начало того длинного пути, который ждет интернет-библиотеки и лингвистов. Однако, с другой стороны, будем надеяться, что филологи не ограничатся интернет-библиотеками. Ведь непосредственный контакт с книгой совершенно другое дело, чем с бездушным дисплеем. Лучшим решением было бы мирное сосуществование обеих форм – традиционной “бумажной” и дигитализированной форм лингвистической литературы. А Интернет и новая информационная среда будут только средствами, которые дадут широкий и богатый доступ к источникам лингвистической литературы.

Приложение

1. Baudouin de Courtenay Jan, Głos członka Akademii J. Baudouina de Courtenay w sprawie słownictwa chemii, Kraków 1900. [PBI]
2. Baudouin de Courtenay J., Jedna z kwestji spornych pisowni polskiej, Warszawa 1890. [PBI]

3. Baudouin de Courtenay J., Mikołaj Kruszewski, jego życie i prace naukowe, Warszawa 1889. [PBI]
4. Baudouin de Courtenay J., O języku pomocniczym międzynarodowym. Odczyt publiczny wygłoszony w Warszawie 5 maja 1908 r., Kraków 1908. [PBI]
5. Baudouin de Courtenay J., O ogólnych przyczynach zmian językowych. Odczyt drugi, miany w Dorpacie 27 marca / 8 kwietnia 1889 r., na korzyść Kasy im. Mianowskiego, Warszawa 1890. [PBI]
6. Baudouin de Courtenay Jan, O zadaniach językoznawstwa. Odczyt pierwszy, miany w Dorpacie 25 marca / 6 kwietnia 1888 r. na korzyść Kasy imienia Mianowskiego, Warszawa 1888. [PBI]
7. Baudouin de Courtenay J., O zjeździe slawistów i O panslawizmie „platonicznym". Dwa odczyty wygłoszone w Krakowie 16 i 18 czerwca 1903 r. na korzyść Towarzystwa Pomocy Naukowej dla Polek im. Kraszewskiego, Kraków 1903. [PBI]
8. Baudouin de Courtenay J., Próba teorji alternacyj fonetycznych . Cz. 1, Ogólna, Kraków 1894. [PBI]
9. Baudouin de Courtenay Jan, Próba uzasadnienia samoistności zjawisk psychicznych na podstawie faktów językowych, Kraków 1903. [PBI]
10. Baudouin de Courtenay J., Rozbiór gramatyki polskiej księdza Malinowskiego, Warszawa 1875. [PBI]
11. Baudouin de Courtenay J., Strona językowa oryginału polskiego listu "Dymitra Samozwańca" do papieża Klemensa VIII. z dnia 24. kwietnia roku 1604, «Rozprawy Wydziału Filologicznego Akademii Umiejętności w Krakowie», t. 29, 1899, s. 183–213. [PBI]
12. Baudouin de Courtenay J., Szkice językoznawcze, t. 1, Warszawa 1904. [PBI]
13. Baudouin de Courtenay J., Z patologii i embryjologii języka, Warszawa 1885. [PBI]
14. Baudouin de Courtenay J., Zarys historji języka polskiego, Warszawa 1922. [PBI]
15. Bem A., Jak mówić po polsku czyli Gramatyka polska w zarysie popularnym, Warszawa 1889. [PBI]
16. Bem A., Zarys wykładu mowy polskiej według wskazówek językoznawstwa porównawczego, t. 1, Warszawa 1883. [PBI]
17. Biskupski L., Słownik kaszubski porównawczy, Warszawa 1891. [ps. Aleksander Berka] [PBI]
18. Bystroń J., O mowie polskiej w dorzeczu Stonawki i Łucyny w Księstwie Cieszyńskim, Kraków 1885. [PBI]
19. Bystroń J., O użyciu genetivu w języku polskim. Przyczynek do historycznej składni polskiej, Kraków 1893. [PBI]
20. Bystroń J., Przyczynek do historyi języka polskiego z początku XV wieku. Na podstawie zapisków sądowych w Księdze Ziemi Czerskiej, Kraków 1887. [PBI]
21. Bystroń J., Przyczynki do składni polskiej. 2, Kraków 1899. [PBI]
22. Cegielski H., O słowie polskiem i konjugacyach jego, wraz z wstępem krytycznym, Poznań 1852. [PBI]
23. Gloger Z., Słownik gwary ludowej w okręgu tykocińskim, Warszawa 1894. [PBI]
24. Gruszczyński S., Nauka o zdaniu zawierająca rozbiór zdania pojedynczego, złożonego i okresu oraz prawideł szyku ułożona na klasy średnie gimnazyów i szkół realnych, tudzież dla seminaryów nauczycielskich i wyższych szkół żeńskich, Poznań 1861. [KPBC]
25. Januszowski J., Nowy karakter Polski. Z drukarnie Lazarzowey y ortographia polska Iana Kochanowskiego, Ie M. P. Lukasza Gornickiego etc. etc., Kraków 1594. [PBI]
26. Kalina A., Historya języka polskiego. T. 1: Formy gramatyczne języka polskiego do końca XVIII wieku, Lwów 1883. [PBI]
27. Kalina A., O liczebnikach w języku staropolskiem jako przyczynek do historii języka polskiego, Petersburg 1876. [PBI]
28. Karłowicz J., Czterysta kilkadziesiąt nazwisk miejscowości litewskich, Odbitka z Pamiętnika Fizyjograficznego, t. III, Warszawa 1883. [PBI]
29. Karłowicz J., Gwara kaszubska, Warszawa 1898. [PBI]
30. Karłowicz J., Imiona niektórych plemion i ziem dawnej Polski, Warszawa 1882. [PBI]
31. Karłowicz J., Imiona własne polskich miejsc i ludzi od zatrudnień, Warszawa 1897. [PBI]
32. Karłowicz J., Niezałatwiona kwestia, kwestya, kwestyja, kwestja ortograficzna, Warszawa 1890. [PBI]
33. Karłowicz J., O imieniu Polaków i Polski, Warszawa 1881. [PBI]
34. Karłowicz J., O imionach własnych polskich miejsc i ludzi, Warszawa 1885. [PBI]
35. Karłowicz J., O języku litewskim // Z Rozpraw i sprawozdań z posiedzeń Wydziału Filologicznego Akademii Umiejętności, t. II, Kraków 1875: 135–376. [PBI]
36. Karłowicz J., Przyczynki do projektu wielkiego słownika polskiego, Kraków 1876. [PBI]
37. Karłowicz J., Słownik gwar polskich, t.1–6. Tom 4–6 do druku przygotował J. Łoś, Kraków 1900–1911. [PBI]
38. Karłowicz J., Słownik wyrazów obcego a mniej jasnego pochodzenia używanych w języku polskim, Kraków 1894–1905. [PBI]
39. Karłowicz J., W sprawie pisowni polskiej i wiążących się z nią zagadnień gramatycznych, Kraków 1883. [PBI]
40. Knapski G., Słownik polsko-łaciński ze skarbu Księdza Knapiusza... wybrany, niektóremi słowy... Benedykta Woronowskiego pomnożony..., t. 1–3, Kalisz 1717. [WBC]

41. Kopczyński O., Grammatyka języka polskiego przez X. Onufrego Kopczyńskiego, Piiara. dzieło pozgonne, Warszawa 1817. [PBI]
42. Kopczyński O., Gramatyka dla szkół narodowych na klassę III, Warszawa 1804. [PBI]
43. Kopczyński O., O duchu języka polskiego [przez...]. Wstęp. Na posiedzeniu publicznem dnia 16 listopada 1804, Warszawa 1804. [PBI]
44. Kopczyński O., Poprawa błędów w ustney i pisaney mowie polskiey, Warszawa 1808. [PBI]
45. Kopczyński O., Treść grammatyki polskiey napisaney przez X. Onufrego Kopczyńskiego, służąca za wstęp do nauk filozoficznych, Wilno 1806. [PBI]
46. Kopczyński O., Układ grammatyki dlá szkół narodowych z dziéła iuż skończonégo wyciągniony, Warszawa 1785. [PBI]
47. Kozierowski S., Badania nazw topograficznych dzisiejszej archidyecezyi gnieźnieńskiej, Poznań 1914. [WBC]
48. Kozierowski S., Badania nazw topograficznych dzisiejszej archidyecezyi poznańskiej. T. I: A–O; T. II: P–Ż. Dodatek A–Ż. Poznań 1916. [WBC]
49. Kozierowski S., Badania nazw topograficznych na obszarze dawnej zachodniej i środkowej Wielkopolski. T. 1: A–Ł. T.2: M–Z „Roczniki Poznańskiego Towarzystwa Przyjaciół Nauk", XLVII (1920); XLVIII (1921). Poznań 1921, 1922. [WBC]
50. Kozierowski S., Nazwiska, przezwiska, przydomki, imiona polskie niektórych typów słowotówrczych. Księga rodzaju ludu polskiego. Poznań 1938. [WBC]
51. Kozierowski S., Nazwy geograficzne i osadźcy ziemi średzkiej z mapą, Poznań 1931. [WBC]
52. Kozierowski S., Nazwy rzeczne w Lechji Przybałtyckiej i w przyległych częściach Słowiańszczyzny Północno-Zachodniej. II. Nazwy na -awa i -ew/-wa, Odbitka z „Slavia Occidentalis", t. X, 1931, s. 160–243. Poznań 1931. [WBC].
53. Kozierowski S., Pierwotne osiedlenie ziemi gnieźnieńskiej wraz z Pałukami w świetle nazw geograficznych i charakterystycznych imion rycerskich, Odbitka z „Slavia Occidentialis", t. III–IV, 1923–1924, s. 1–129. Poznań 1924 [WBC]
54. Kozierowski S., Pierwotne osiedlenie pogranicza wielkopolsko-śląskiego między Obrą i Odrą a Wartą i Bobrem w świetle nazw geograficznych, Odbitka z „Slavia Occidentialis", t. VII, 1928, Poznań 1928, s. 172–329. [WBC]
55. Kozierowski S., Pierwotne osiedlenie dorzecza Warty od Koła do Ujścia w świetle nazw geograficznych, Odbitka z „Slavia Occidentialis", t. VI, 1926, Poznań 1926, s. 112–246. [WBC]
56. Krasiński A. S., Słownik synonimów polskich, t. 1–2, Kraków 1895. [PBI]
57. Krasnowolski A., Systematyczna składnia języka polskiego, Warszawa 1909. [KPBC]
58. Kucharski A., Księdza Stanisława Zaborowskiego Ortografija polska. Z przydaniem uwag tłómocza, tudzież Ortografii Seklucyana i spisu bibliograficznego grammatyk i słowników polskich, Warszawa 1825. [PBI]
59. Linde S. B., Prawidła Etymologii, przystosowane do ięzyka polskiego // Roczniki Towarzystwa Warszawskiego, T. IV (1804), s. 73–155. [PBI]
60. Linde S. B., Słownik języka polskiego, t.II–VI cz.1 (G–W), Lwów 1855–1860. [PBI]
61. Łebiński W., Materiały do słownika łacińsko-polskiego średniowiecznej łaciny i starożytności polskich, Poznań 1885.
62. Łopaciński H., Najdawniejsze słowniki polskie drukowane, Warszawa 1897. [PBI]
63. Łoś J., Bogurodzica. Pierwszy polski hymn narodowy, Lublin 1922. [PBI]
64. Łoś J., Główne zasady pisowni polskiej według punktów ustalonych przez Akademję Umiejętności, Kraków 1919. [PBI]
65. Łoś J., Gramatyka polska. Cz. 1, Głosownia historyczna, Lwów 1922. [PBI]
66. Łoś J., Gramatyka polska. Cz. 2, Słowotwórstwo, Lwów 1925. [PBI]
67. Łoś J., Gramatyka polska. Cz. 3, Odmiennia (fleksja) historyczna, Lwów 1927. [PBI]
68. Łoś J., Jakóba syna Parkoszowego Traktat o ortografii polskiej, Kraków 1907. [PBI]
69. Łoś J., Krótka gramatyka historyczna języka polskiego, Lwów 1927. [PBI]
70. Łoś J., Pisownia polska w przeszłości i obecnie. Zagadnienia i wnioski, Kraków 1917. [PBI]
71. Łoś J., Przegląd językowych zabytków staropolskich do roku 1543, Kraków 1915. [PBI]
72. Malinowski F. K., Krytyczno-porównawcza gramatyka języka polskiego z dzisiejszego stanowiska lingwistyki porównawczéj, Poznań 1869. [PBI]
73. Malinowski F. K., Zasady i prawidła pisowni polskiej przez Deputacją od Król. Towarzystwa Warszawskiego wyznaczoną podane w dziele pod napisem: "Rozprawy i wnioski o ortografiji polskiej" w Warszawie 1830. Wraz z kriticznym rozbiorem i zmianą niektórych prawideł Deputacyji z wytknięciem niektórych błędów od czasu ks. Kopczyńskiego i później powstałych, aż dotąd język Polski każących, Poznań 1859. [PBI]
74. Malinowski L., Materyjały do historyi języka polskiego z końca XIV i początku XV wieku na podstawie ksiąg sądowych łęczyckich, orłowskich i brzezińskich (Kraków 1884) // Teki A. Pawińskiego, t. V: Księgi sądowe łęczyckie, cz. III, Warszawa 1898. [WBC]
75. Morawski Sz., Wyrazy fenickie w mowie polskiej, Lwów 1885. [PBI]
76. Papłoński J., Lekcyja wstępna filologii porównawczej sławiańskiej miana dnia 2 października, 9 listopada 1864 r. w Szkole Głównej Warszawskiej, Warszawa 1864. [PBI]
77. Stankiewicz Z., Pisownia polska według prof. Ad. Ant. Kryńskiego w zestawieniu z pisownią przyjętą przez Akademię Umiejętności w Krakowie i Zjazd Rejowski ze słowniczkiem wyrazów o pisowni wątpliwej. Warszawa 1908. [PBI]

ЛІНГВІСТИЧНІ ПРОЦЕСОРИ

І. Карамишева, Р. Карамишева*
Національний університет “Львівська політехніка” (Львів)
УДК 81’322

АНАЛІЗ МОРФЕМНОГО СКЛАДУ НЕОЛОГІЗМІВ ЗАСОБАМИ СУЧАСНИХ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ (НА ОСНОВІ АНГЛІЙСЬКОЇ МОВИ)

The presented article focuses on the morphemic analysis of English neologisms, namely those created by means of blend compounding. It has been found out that the base of such neologisms possesses vague semantic motivation, which hinders its correct interpretation. The IT means are not ready to facilitate their proper morphological analysis. The solution to this problem may lie in the creation of the new word stems’ database.

Неологізми – це слова або мовні звороти, створені для позначення нового предмета чи поняття. З’являючись у мові, кожне слово спочатку є неологізмом, згодом воно набуває широкого вжитку, входить до активного словникового складу [2: 105]. Такі одиниці можуть становити певні труднощі для аналізу через свої відмінності від основної частини вже існуючих слів. Саме тому аналіз неологізмів є надзвичайно актуальним. При дослідженні неологізмів труднощі виникають не лише при їхньому розумінні та перекладі, але й при морфологічному аналізі. Відтак мета нашої роботи –розглянути проблеми морфологічного аналізу неологізмів.

Для аналізу було зроблено вибірку з 300 одиниць неологізмів переважно суспільно-політичної тематики [6–10]. Особливий інтерес становлять 13% вибірки – слова, утворені не дуже продуктивним до останнього часу способом творення англійських слів – блендом або блендінгом.

Бленд – це особливий тип словотвору за допомогою злиття частин слів у одне слово, наприклад, *politainer (politician+entertainer).* Такий тип словотвору зумовлений прагненням до мовної економії, він стає особливо популярним у сучасній англійській мові [5: 98]. Процес словотвору, під час якого основи слів ніби входять одна в іншу, називають телескопією. При цьому збігаються дві однакових голосних або приголосних, наприклад, *slanguage (slang+language)* [4: 47].

Найчастіше таким способом утворюються іменники. Виокремлюють такі групи

Word to find

snoopervision

Results

Lang	Status	Type	Normal form	Information	ID / Source
Eng	Predicted		SNOOPERVISION	NOUN narr,sg,;	VISION
Eng	Predicted		SNOOPERVISION	VERB inf,;	VISION

блендів:

1) творення слова з початкових елементів одного слова та закінчення іншого: *banalysis (banal+analysis);*
2) поєднання одного слова із закінченням іншого слова: *eatertainment (eater+entertainment);*
3) поєднання початку одного слова із повним словом: *democrazy (democracy+crazy)* [5: 99].

Проведено морфологічний аналіз таких неологізмів за допомогою програми Dialing Morphology. Оскільки в базі даних програми ці неологізми відсутні, то їм надавався статус predicted (прогнозований) і їхній аналіз проводився за аналогом подібних слів:

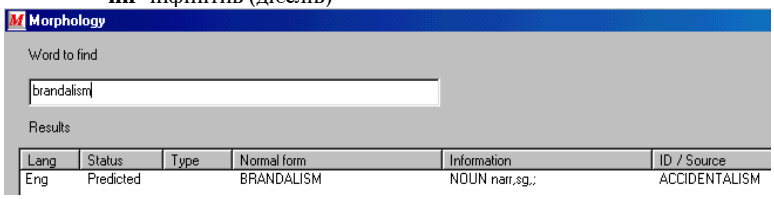
Word to find

businesscrat

Results

Lang	Status	Type	Normal form	Information	ID / Source
Eng	Predicted		BUSINESSCRAT	NOUN narr,sg,;	ANTIDEMOCRAT

NOUN – ІМЕННИК
VERB – ДІЄСЛОВО
narr – загальна назва
sg – однина
inf- інфінітив (дієслів)



Результати комп'ютерного морфологічного аналізу є цілком задовільними. Проте в комп'ютерній лінгвістиці під морфологічним аналізом розуміють аналіз окремих словоформ поза контекстом. У результаті такого аналізу послідовність словоформ вхідного тексту замінюються послідовністю інформації про ці словоформи. Комп'ютер для кожного слова в тексті визначає його граматичний клас, або частиномовну належність, та в межах граматичних класів – граматичний підклас, або граматичні підкласи, тобто розряди слів зі спільними змістовими, формальними та функціональними властивостями. Нові для себе слова програма аналізує за зразком уже відомих, обираючи ті, які мають однакові формотворчі суфікси. Все решта поза суфіксом буде для програми основою слова, яку вона далі не членує.

Для нас морфологічний аналіз неологізмів, що утворились блендом, є важливим насамперед тому, щоб зрозуміти їхню основу. Неологізми, які утворюються афіксацією чи складанням основ, містять вже відомі основи, що піддаються аналізу. Те, що ми отримуємо шляхом бленду, є новим типом основи, яка не має прозорої семантичної мотивації. Н. М. Раєвська пише, що результатом бленду є монема, тобто просте слово, яке не можна проаналізувати, а не вмотивована синтагма [5: 99]. Справді, адже суфікс *–ism*, який виділив комп'ютерний морфологічний аналіз у слові *brandalism* не був доданим до існуючої основи *brandal*, а є частиною слова *vandalism*, яке інтегрувалось зі словом *brand* утворивши просте слово. Це просте слово може згодом стати основою для словотворення. Такі основи створюють проблему для аналізу, подібно до тих, які М. А. Кронгауз називає семантично спустошеними або порожніми, оскільки ми не можемо чітко визначити їх семантику [3: 210].

Отже, доходимо висновку, що в цьому разі при морфологічному аналізі за допомогою комп'ютерних програм ми не отримаємо бажаного результату, адже людина, аналізуючи неологізми, перш за все, буде намагатись розібрати основу, зрозуміти, з яких компонентів вона складається, і лише потім зверне увагу на формотворчі афікси. Проведене дослідження показує, що морфологічний аналіз неологізмів засобами інформаційних технологій вдвічі ускладнюється, оскільки основа неологізмів часто представлена на поверхневому рівні непрозоро. При автоматичному морфологічному аналізі використовується поняття машинної основи та машинної флексії [1: 78], які можуть відрізнятися від традиційних. Відтак, можливим розв'язанням цієї проблеми вбачаємо створення баз даних нових основ.

Література

1. Беляева Л. Н. Лингвистические автоматы в современных информационных технологиях: Монография. – СПб.: Изд-во РГПУ им. А. И. Герцена, 2001. – 130 с.
2. Єрмоленко С. Я., Бибик С. П., Тодор О. Г. Українська мова. Короткий тлумачний словник лінгвістичних термінів / За ред. С. Я. Єрмоленко. – К.: Либідь, 2001. – 224 с.
3. Кронгауз М. А. Семантика. Учебник для вузов. – М.: Рос. гос. гуманит. ун-т, 2001. – 399 с.
4. Мостовий М. І. Лексикологія англійської мови. – Харків: Основа, 1993. – 256 с.
5. Раєвська Н. М.. Лексикологія англійської мови. – К.: Вища школа, 1979. – 304 с.
6. www.andromeda.rutgos.edu
7. www.ditl.info
8. www.lisatalk.su
9. www.wikipedia.org
10. www.wordspy.com

Т. Юзепольська*
Київський національний лінгвістичний університет (Київ)
УДК 81'322

ПАРАДИГМАТИЧНІ КЛАСИ ІМЕННИКА АНГЛІЙСЬКОЇ МОВИ (КАТЕГОРІЇ ЧИСЛА: АНАЛІЗ ТА СИНТЕЗ)

The article deals with the problem of classification of nouns: setting different types on the base of their plural forms. To differentiate the nouns morphological analysis has been performed. The data-base has been worked out that includes a great variety of noun characteristics. Applying this data-base, the application "Noun" has been created. It could help to find plural forms of the nouns and to transform the given Singular nouns into corresponding Plural forms.

Опрацювання текстової інформації незмінно залишається провідним завданням комп'ютерної лінгвістики. Наші знання про дійсність втілюються в певній вербальній формі. Навчити комп'ютер „розуміти" текст і означає наділити його здатністю видобувати з нього потрібну для виконання того чи іншого завдання інформацію. Таке „розуміння" тексту полягає у вмінні аналізувати його на різних рівнях представлення інформації: морфологічному, синтаксичному, логіко-семантичному і узагальнювати одержані внаслідок подібного аналізу результати у певній визначеній формі [2].

Об'єктом цього дослідження є морфологічні ознаки іменника англійської мови. Морфологічний аналіз слугує основою семантичної обробки інформації, що представляється природною мовою, і тому є необхідним компонентом процесу автоматичного індексування. В обчислювальній лінгвістиці під морфологічним аналізом розуміють аналіз окремих словоформ поза контекстом. У результаті морфологічного аналізу послідовність словоформ вихідного тексту замінюється на послідовність інформації про ці словоформи.

Поняття „морфологічний аналіз" в обчислювальній лінгвістиці відрізняється від традиційного лінгвістичного уявлення про морфологічний аналіз. Відповідно до останнього морфологічний аналіз веде до визначення частини мови і морфологічної будови слова. Ціллю аналізу є опис форми слова. В обчислювальній лінгвістиці морфологічний аналіз може слугувати засобом отримання різних даних, і навіть тих, що не стосуються морфології [1].

Предметом дослідження було обрано категорія числа іменників англійської мови, що в майбутньому може стати основою для автоматичного перекладу й розв'язання багатьох інших завдань комп'ютерної лінгвістики. Також результати морфологічного аналізу можуть бути вихідними даними етапів автоматичного аналізу науково-реферативного тексту — синтаксичного, лексичного, семантичного, аналізу текстових надфразних відношень, а також слугувати базою для виявлення особливостей морфологічної структури наукових рефератів.

Матеріалом дослідження є група іменників англійської мови, відібраних зі словника Macmillan Essential Dictionary (for learners of English). Саме на основі їх формального аналізу були визначені парадигматичні класи іменників і створена програма аналізу та синтезу форми множини.

Метою дослідження є створення програми, яка могла б утворювати форму множини заданого їй іменника, розпізнавати цю форму в тексті та визначати її початкову форму, тобто форму однини.

Визначаючи парадигматичні класи іменників, можна ще раз пересвідчитись у тому, що форма слова відіграє суттєву роль у визначенні певних граматичних категорій. Завдяки їй можна видобути з тексту багато граматичної інформації, яку потім можна використати у створенні певних систем автоматичного аналізу тексту. Однак, на жаль, не в усіх мовах форма говорить так ясно про свою граматичну приналежність. Це залежить від типу природної мови (аналітична чи флективна). Наприклад, у мови з багатою морфологією (українська чи латинська) аналіз окремо взятої словоформи дає можливість отримати велику кількість різноманітної інформації про різні граматичні категорії. А в такій мові як англійська (аналітична) форма слова не дозволяє дістати всю ту інформацію, яку б нам хотілось. Більше того, виникає проблема лексико-граматичної омонімії

* © *Т. Юзепольська, 2006*

(наприклад, *answers* як форма множини іменника і як форма третьої особи однини дієслова; *(the) rich* як іменник, що позначає групу людей з певними особливими ознаками і як прикметник). Тому в таких випадках треба звертати увагу ще й на оточення того слова, що ми аналізуємо.

Створення бази даних «mac_noun»

Щоб зробити класифікацію іменників за ознакою утворення множини, потрібно було опрацювати близько 11000 іменників, відібраних зі словника Macmillan. Для цього за допомогою Microsoft Access було створено базу даних "*mac_noun*". Таблиця "noun" (рис. 1) складається з наступних полів: id, wrd, c/u, rule_pl, plural, rule_pl2, plural2, signs, par_kl, chastotn, the, sphere, B/A, definition.

В першу чергу було зафіксовано інформацію щодо сфери вживання (formal (anomaly, refuse), informal (bike, flop), medical (diabetic, plasma), linguistics (gerund, participle)) та частоти вживаності (частотності) деяких слів. Наступним етапом дослідження було визначення типу іменника (злічуваний (balcony, ladder, yacht), незлічуваний (bodywork, marble, turbulence), що вживається лише в однині (Singularia Tantum) (atmosphere, monotone, subconscious), що вживається лише в множині (Pluralia Tantum) (bedclothes, remains, acrobatics), злічуваний або незлічуваний (cake, exposure, size) та ін.). Відомо, що не всі іменники англійської мови утворюють форму множини. Протягом цього етапу перевірка типу іменників відбувалася за допомогою словника Macmillan і граматики англійської мови. Таким чином було отримано ту групу іменників (7542 слів), над якою надалі буде проводитися аналіз. Потім були формально описані правила утворення форми множини простих та складених іменників (<e* (basis, congressman, saleswoman), >s (anagram, house, formula), <s-in-la >w (brother-in-law, sister-in-law), <v >es (elf, werewolf, sheaf) і т. д.). На завершальному етапі було визначено 98 парадигматичних класів на основі типів іменників, правил утворення форми множини та кількості слів в одному іменникові.

Створення програми «Noun»

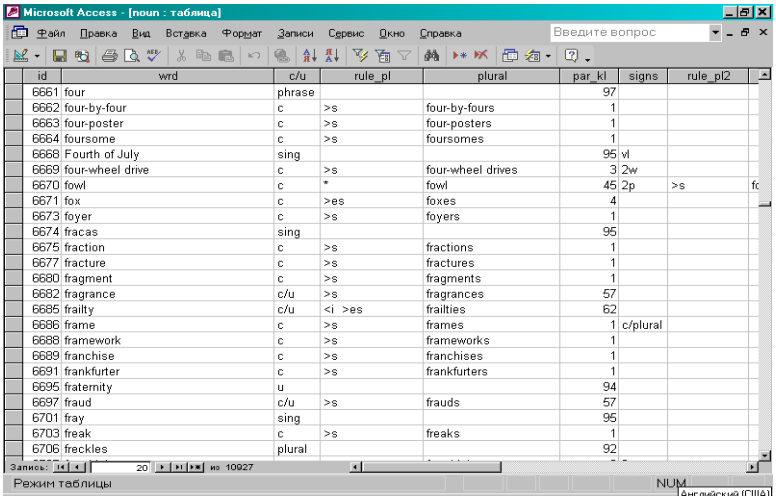
Існує два підходи до створення програм такого роду: на основі певних правил та на основі словника. Програма, основана на першому підході виконує синтез та аналіз множини на основі заданих їй формальних правил. Зазвичай така програма займає зовсім небагато місця, але її недоліком є досить значна кількість помилок, адже неможливо задати абсолютно всі правила і врахувати усі тонкості мови. Програма, створена на основі словника, звичайно займає набагато більше місця, але в ній менше помилок. А так, як зараз комп'ютерні об'єми пам'яті майже необмежені, то, на мою думку, доцільніше використовувати саме такий підхід. Що й було зроблено в даному дослідженні.

Сам робочий словник, а саме форму множини усіх слів (адже ми мали лише список слів у формі однини), було отримано за допомогою допоміжної програми, написаної мовою C#. Програма "Noun" має два блоки (рис. 2): аналіз та синтез множини іменника.

У результаті дослідження парадигматичних класів іменників англійської мови було зроблено наступне:

1) Створена база даних з 11000 іменників, у якій міститься детальна характеристика кожного слова. 2) Сформульовані формальні правила утворення форми множини. 3) Визначений тип іменника: countable or uncountable, countable, uncountable, plural, singular, singular or uncountable та ін. 4) Призначений номер парадигматичного класу. 5) Визначена сфера вживання.

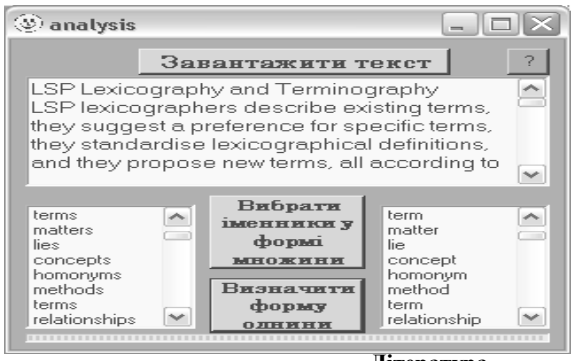
Рисунок 1. База даних "*mac_noun*". Таблиця "noun"



id	wrd	c/u	rule_pl	plural	par_kl	signs	rule_pl2
6661	four	phrase			97		
6662	four-by-four	c	>s	four-by-fours	1		
6663	four-poster	c	>s	four-posters	1		
6664	foursome	c	>s	foursomes	1		
6668	Fourth of July	sing			95	vl	
6669	four-wheel drive	c	>s	four-wheel drives	3	2w	
6670	fowl	c	*	fowl	45	2p	>s
6671	fox	c	>es	foxes	4		
6673	foyer	c	>s	foyers	1		
6674	fracas	sing			95		
6675	fraction	c	>s	fractions	1		
6677	fracture	c	>s	fractures	1		
6680	fragment	c	>s	fragments	1		
6682	fragrance	c/u	>s	fragrances	57		
6685	frailty	c/u	<i >es	frailties	62		
6686	frame	c	>s	frames	1	c/plural	
6688	framework	c	>s	frameworks	1		
6689	franchise	c	>s	franchises	1		
6691	frankfurter	c	>s	frankfurters	1		
6695	fraternity	u			94		
6697	fraud	c/u	>s	frauds	57		
6701	fray	sing			95		
6703	freak	c	>s	freaks	1		
6706	freckles	plural			92		

Цей етап дослідження вказує на те, що морфологічний аналіз слова дозволяє автоматизувати певні види роботи з мовними одиницями, що значно полегшує працю (роботу) лінгвістів. На основі цієї бази даних було створено програму «Noun», яка виконує аналіз та синтез форми множини англійського іменника. Вона утворює множину іменника від заданого, а також розпізнає у тексті усі іменники, що стоять у формі множини і потім визначає їх форму однини. Результати даного дослідження доводять, що проблема зняття омонімії залишається актуальною так само, як і розрізнення типу іменника, який одночасно може бути і злічуваним, і незлічуваним.

Рисунок 2. Програма “Noun”



Література

1. Василевский А. А., Марчук Ю. Н. Вычислительная лингвистика. – М., 1970.
2. Волошин В. М. Вступ до комп’ютерної лінгвістики (Донецький національний університет). — Донецьк: ТОВ „Юго-Восток, Лтд”, 2003. — 184 с.
3. Использование ЭВМ в лингвистических исследованиях. – К.: Наук. думка, 1999.
4. Каушанская В. Л. Грамматика английского языка. — М., 2000.
5. Качалова К. Н., Израилевич Е. Е. Практическая грамматика английского языка с упражнениями и ключами. — СПб.: БАЗИС, КАРО, 2004. — 608с.
6. Морфологический анализ научного текста на ЭВМ. – Киев: Наук. думка, 1989.
7. Digby Beaumont, Colin Granger. The Heinemann ELT English Grammar.
8. Macmillan English Dictionary (for learners of English).

В. Дембіцький*
Київський національний лінгвістичний університет (Київ)
УДК 81’33

СТУПЕНІ ПОРІВНЯННЯ ПРИКМЕТНИКІВ І ПРИСЛІВНИКІВ. СИНТЕЗ І РОЗПІЗНАВАННЯ В ТЕКСТІ ПАРАДИГМАТИЧНИХ КЛАСІВ АНГЛІЙСЬКИХ ПРИКМЕТНИКІВ ТА ПРИСЛІВНИКІВ

Стаття присвячена проблемі комп’ютерного оброблення письмових текстів, зокрема морфологічного аналізуі. Автор пропонує програму, яка, керуючись правилами утворення ступенів порівняння, автоматично утворює їх для відібраних з англо-українського навчального словника прикметників та прислівників. Наступний модуль програми дозволяє виділити прикметники та прислівники у тексті, а також пояснює ступінь порівняння знайденого слова.

Здатність комп’ютера опрацьовувати мовний матеріал обмежена набором операцій і засобів, котрі доступні обчислювальній техніці на сучасному етапі її розвитку. Ми можемо використовувати комп’ютер як помічника при умові, що сформулюємо йому завдання так, щоб його можна було б викласти однією з мов програмування. Обов’язковою складовою частиною лінгвістичного забезпечення будь-якої системи

* © *В.Дембіцький, 2006*

автоматичної обробки письмових текстів є морфологічний аналіз, який, власне, і є одним з основних етапів роботи системи автоматичного аналізу тексту.

На основі опрацьованих (досліджених) прикметників та прислівників слід розробити алгоритми утворення парадигми англійських прикметників та прислівників (де це можливо), а також принципи розпізнавання їх (у будь-якому ступені порівняння) у тексті та утворення від вищого та найвищого ступенів початкової форми прикметника або прислівника, а також створити програму, яка б практично застосовувала отримані принципи (правила).

Для проведення дослідження з навчального англо-українського словника другого рівня відібрано усі реєстрові прикметники та прислівники і збережено у таблицях бази даних:

прикметників 721;
прислівників 260;

Згідно з граматикою англійської мови розроблено правила утворення і розпізнавання ступенів порівняння вищезгаданих частин мови.

На мові програмування C# складена програма, яка, керуючись правилами утворення ступенів порівняння, автоматично їх утворює для відібраних з англо-українського навчального словника прикметників та прислівників. Другий модуль програми дозволяє виділити прикметники та прислівники у тексті, а також пояснює ступінь порівняння знайденого слова.

Характеристика досліджуваних частин мови.

Залежно від значення прикметники поділяють на якісні (qualitative) та відносні (relative). Більшість якісних прикметників мають ступені порівняння, тоді як відносні взагалі не мають ступенів порівняння. Однак навіть деякі якісні прикметники не мають ступенів порівняння (greenish, darkish, incurable, unsuitable, chief, principal…). Більшість якісних прикметників можуть утворювати прислівники (adverbs) за допомогою додавання суфікса -ly.

Ступені порівняння прикметників.

Більшість якісних прикметників мають ступені порівняння (degree of comparison):

- **початкова форма (positive degree):**

початкова форма вказує на якість предмета і відповідає словниковій формі, тобто прикметники у початковій формі не мають ніяких закінчень: difficult, green. Однак часто коли говорять про однакову ступінь якості різних предметів використовують сполучник «as … as» – «такий же…, як» або його заперечний варіант «not so … as» – «не такий …, як».

The line AB is as long as the line CD.

Конструкція «as (adjective) as» може бути доповнене у передпозиції прислівниками not, almost, twice, three times, half, one-third та ін.

Mabel is as clever an administrator as Robin.

New York is as distant from San Francisco as Boston is from London.

- **вищий (comparative degree) – позначає вищий рівень якості:**

наприклад, якщо потрібно вказати, що один предмет володіє більш вираженою ознакою у порівнянні з іншим предметом, то використовують прикметник у вищому ступені, який утворюється шляхом додавання суфікса «-er» до основи прикметника (який складається з одного або двох складів) наприклад:

short – shorter, dark – darker

- **найвищий (superlative degree)- позначає найвищий рівень якості,**

вказує на те, особа, предмет чи група володіє найбільш вираженою ознакою у порівнянні з усіма іншими (наприклад, biggest/(the) most beautiful)

Розбіжності у способах творення ступенів порівняння (одночасно і синтетичний, і аналітичний способи) спричинені історичним розвитком англійської мови.

Прикметники (як і прислівники) утворюють свої форми порівняння наступними способами:

1. За допомогою закінчення «–er»; «-est» (синтетичний спосіб).

Прикметники, які утворюють ступені порівняння шляхом додавання закінчень (синтетичним способом), включають односкладові прикметники та двоскладові з закінченням «-y». Наприклад: brave (1)[1], happy (2).

Ще кілька двоскладових прикметників інколи вживаються з закінченнями «-er» / «-est», наприклад: clever, quiet.

Варто відзначити, що односкладові дієслова у формі дієприслівників минулого часу (past participle), які у тексті вжиті як прикметники, зазвичай вживаються без закінчень.

2. За допомогою додавання прислівників more/most перед прикметником (аналітичний спосіб).

Як видно з наведеної вище інформації, односкладові прикметники зазвичай утворюють вищий і найвищий способи порівняння у перший спосіб (синтетичний), а багатоскладові – у другий спосіб (аналітичний).

Правила утворення ступенів порівняння прикметників:

Регулярні прикметники

Односкладові прикметники утворюють вищий ступінь порівняння шляхом додавання до форми прикметника у початковій формі суфікса «-er». Найвищий ступінь порівняння утворюється способом додавання суфікса «-est» (синтетичний спосіб): hard – harder – (the) hardest.

Односкладові прикметники, які закінчуються на «-e» утворюють ступені порівняння додаванням «-r», «-st»: Nice – nicer – (the) nicest. А якщо основа прикметника закінчується на букву «-y» з попереднім приголосним, то при додаванні суфікса «-er/-est» буква «-y» переходить в «-i»: dry – drier – (the) driest; easy – easier – (the) easiest.

В односкладових прикметників з комбінацією літер типу cvc (consonant-vowel-consonant/приголосний-голосний-приголосний), де кінцевим є будь-який єдиний приголосний, крім “w”, “x” або “y”, перед «-er»/ «-est» подвоюється кінцева приголосна у вищому і найвищому ступені порівняння, щоб зберегти закритий склад: big – bigger – (the) biggest.

Коли прикметник закінчується на “w”, “x” or “y”, перед яким знаходиться єдиний наголошений голосний, останній приголосний не подвоюється перед додаванням «-er»/ «-est», наприклад: slow – slower, gray – greyer.

Варто відзначити, що коли прикметник закінчується на єдиний приголосний, який стоїть після двох голосних, останній приголосний не подвоюється перед додаванням «-er»/ «-est», наприклад: loud – louder, neat – neater, soon – sooner

Однак існують деякі багатоскладові прикметники, які утворюють вищий і найвищий ступені порівняння за допомогою закінчень (синтетичним способом). Деякі двоскладові прикметники також можуть утворювати ступені порівняння за допомогою «-er»/»-est» (як правило прикметники, які закінчуються на –y; -ow; -er; -le.), а також слова (polite, severe, sincere, common):

Прикметники, що мають 2 склади, але з наголосом на останньому складі:

complete – completer – (the) completest

concise – conciser – (the) concisest

Двоскладові прикметники, які закінчуються на «–y», змінюють «-y» на «–i»:

happy – happier – (the) happiest

nasty – nastier – (the) nastiest

Двоскладові прикметники, які закінчуються на ненаголошену голосну:

hollow – hollower – (the) hollowest (having a cavity, sunken)

Двоскладові прикметники, які закінчуються на [l], тобто «-el», «-le»

simple – simpler – (the) simplest

Двоскладові прикметники, які закінчуються на [r]

clever – cleverer – (the) cleverest

Багато двоскладових прикметників і всі прикметники, які мають більше двох складів утворюють ступені порівняння за допомогою відповідно more та most, less та least. Ці слова ставляться перед прикметниками у початковій формі (зазвичай цим способом утворюють ступені порівняння похідні та складні якісні прикметники, про які йшлося вище):

active – more active – (the) most active

---

[1] У дужках вказано кількість складів.

difficult – more difficult – (the) most difficult

Нерегулярні прикметники:

Деякі прикметники мають нерегулярні форми у вищому та найвищому ступенях:

good – better – (the) best
bad – worse – (the) worst

Деякі прикметники мають дві форми порівняння: друга має особливе значення, яке, власне, є зовсім іншим словом:

many – more – most (with countable nouns)
much – more – most (with uncountable nouns)
far – farther – farthest (of distance)
far – further – furthest (of distance or time)

Further має ще одне значення, наприклад воно може мати значення «додатковий» (additional, extra).

Do you need any further information?
Near – nearer – nearest
near – nearer – next
Наприклад: «What is the name of the next village?»
(old – older – oldest) & (old – elder – eldest)
(late – later – latest) & (late – latter – last)

Прислівник – це частина мови, яка виражає деякі умови, що стосуються дії або стану, або відзначає певні характерні риси дії або якості. Прислівники вказують на те, коли, де, чому, за яких обставин чи як (у який спосіб) щось зроблено (у той час як прикметники показують ознаку предмета чи особи), або щось відбувалось чи відбувається. Основна синтаксична функція прислівника – обставинна. Прислівники часто закінчуються на **«-ly»**, однак багато з них не мають такого закінчення, і до того ж воно не завжди означає, що слово з цим закінченням – прислівник. Наприклад, існує кілька прислівників на «**-ly»**, які не мають відповідних собі прикметників: friendly, likely.

Деякі прислівники **мають дві форми**: в одній слово закінчується на **«-ly»**, а у другій ні. У деяких випадках, ці дві форми мають різні значення:

He arrived late.
Lately, he couldn't seem to be on time for anything.

Прикметник	прислівник(значення)	прислівник(значення)
good	well	
difficult	with difficulty	
public	publicly	
deep	deep (place)	deeply (feeling)
direct	direct	directly (=soon)

Треба відзначити, що *well* крім прислівника, може бути ще й прикметником, що означає *healthy* (здоровий). Прикметник *well* частіше вживається як предикативний ад'єктив.

***Well** (предикативний прикметник, означуючий ім. **you**): I hope you are **well**.*
***Well** (прислівник способу дії): He did **well** on the examination.*

Подані нижче прикметники можуть також вживатись як прислівники (без змін): fast, hard, little, loud, hourly, enough, long.

Прислівники з і без «ly»

У наступній таблиці подано приклади пар прислівників, які тісно пов'язані, але мають різні значення.

Прислівник	Значення	Прислівник	Значення
hard	with effort	**hardly**	scarcely
high	opposite of **low**	**highly**	very; very well
late	opposite of **early**	**lately**	recently
near	opposite of **far**	**nearly**	almost
wide	opposite of **narrow**	**widely**	commonly

Ступені порівняння прислівників:

Більшість прислівників, головним чином прислівники способу дії (серед яких прості (simple) та похідні (derivative)), можуть мати ступені порівняння, що утворюються так само, як і ступені порівняння прикметників.

Прислівники, які мають ту ж початкову форму, що і відповідні прикметники, мають ту ж саму форму у вищому та найвищому ступенях порівняння. Наприклад, якщо слово має лише один склад, вищий ступінь утворюється за допомогою додавання до простої форми **«–er»** для утворення вищого ступеня, а щоб утворити найвищий **«-est»**: fast – faster – tastest, near – nearer – nearest.

Ми часто використовуємо *more* та *most, less* та *least,* цим самим виражаючи ступінь порівняння прислівника. Тут перша пара слів вживається для нарощення ознаки, а друга — для відображення спадання ознаки. Зазвичай прислівники, утворені від прикметників шляхом додавання суфікса «-ly», утворюють вищий ступінь саме таким способом: *clearly – more clearly – most clearly, softly – more softly – most softly.*

Варто відзначити, що прислівники, утворені додаванням **«ly»** до односкладового прикметника, інколи вживаються з закінченням «**er»** або «**est»**. *We walked slower and slower. О*днак у сучасній англійській мові буде правильніше сказати: *We walked more and more slowly*.

Ступені порівняння прислівників *often* (часто), *quickly* (швидко), *slowly* (повільно) утворюються обома способами: often – more often/oftener – most often/oftenest; quickly – more quickly/quicklier – most quickly/quickliest; slowly – more slowly/slowlier – most slowly/slowliest.

Деякі прислівники, як і відповідні їм прикметники, утворюють ступені порівняння не за правилами (нерегулярні форми). Вони мають ті ж форми порівняння, що і відповідні нерегулярні прикметники: well – better – best; far – farther/further – farthest/furthest.

Важливо, що такі прислівники, як sometimes, never, here, there, now, then, first, again, yesterday і daily, не мають ступенів порівняння.

Робота основного алгоритму визначення прикметників і прислівників, вжитих у вищому та найвищому ступенях (головним чином утворених синтетичним способом), а також і у початковій формі, побудована на аналізі кінцевих ланцюжків літер (квазіфлексій) слововживань, які відбивають особливості формотворчих елементів, властивих даному класу слів. Фактично, аналіз починається зі знаходженням основи слова у словнику і виділенням можливих змін у морфемному складі. Враховуючи зв'язок автоматичного морфологічного аналізу зі словником потрібно згадати про морфологічний аналіз зі словником основ.

Цей принцип використовується досить часто. Процес пошуку починається з того, що кожну словоформу, взяту з тексту, порівнюють на (повне) збігання з лексичними одиницями словника. Якщо словоформа не знайдена, то від слова відсікається потенційна формотворча морфема і знову відбувається пошук на збігання, але тепер з основою. Якщо вибрана і словникова форми основи збіглися, то першій надається інформація, яка міститься у словнику при початковій формі. Якщо шуканої словоформи в словнику немає, то вона вважається такою, що не належить до цієї частини мови або просто не внесена до бази даних.

У кожній мові існують особливості суфіксального словотворення і словозміни, які ускладнюють алгоритм розпізнавання основ. Так, при знаходженні англійської словоформи "bigger" при пошуку відкидається спочатку "er". Залишок "bigg" шукається у словнику.

З вищесказаного випливає, що алгоритм власне морфологічного аналізу залежить від того, що розуміється під основою слова. Словник основ може містити одиниці двох видів: а) канонічні основи, тобто основи, які містяться у звичайному словнику (іменник у називному відмінку однини, дієслово в інфінітиві, прикметник і прислівник у початковій формі); б) основи, представлені графічно незмінною частиною словоформ (тобто основою).

Класифікація словоформ при використанні канонічних основ відбувається для прикметників та прислівників наступним чином: існують лише три форми (початкова, вищий та найвищий ступені порівняння).

Розробка принципів автоматичного аналізу і синтезу прикметників і прислівників англійської мови (початкова форма, вищий та найвищий ступінь).

Згідно з граматикою англійської мови прикметники та прислівники можуть утворювати ступені порівняння лише двома способами — аналітичним та синтетичним (або не утворювати їх взагалі).

Для того, щоб розробити систему автоматичного синтезу ступенів порівняння англійських прикметників, як і для утворення парадигми будь-якої іншої самостійної частини мови, перш за все потрібно укласти систему правил, які було б легко інтерпретувати під комп'ютерну програму. Тут обов'язково потрібно врахувати усі можливі нерегулярні випадки утворення парадигматичних форм, оскільки у іншому випадку програма просто буде невірно утворювати парадигму.

Так, якщо подивитись на прикметники, варто відзначити, що не всі з них можуть утворювати ступені порівняння, а деякі можуть утворювати тільки найвищий ступінь (utter – …-utmost/uttermost). Суплетивні форми типу good – better – best потрібно просто задати окремо як винятки.

Основна проблема полягає у тому, чи має слово ступінь порівняння взагалі, навіть якщо воно і підпадає під певне правило словозміни, оскільки серед прикметників ступені порівняння мають лише більшість якісних, та й серед прислівників процент тих, які можуть утворювати ступені порівняння — досить невисокий.

Згідно з граматикою для прислівників та прикметників була розроблена система правил словозміни при утворенні ступенів порівняння. Таблиця з правилами подана нижче:

Таблиця . Уніфікована система правил утворення ступенів порівняння прикметників та прислівників англійської мови

правило	вищий	найвищий	приклад	пояснення
PR0	—	—	—	ступенів пор. нема
PR1	>er	>est	fast-faster-fastest	додавання частки до основи
PR2	< *>*er	< *>*er	big-bigger-biggest	подвоєння останньої приголосної плюс додавання частки до основи
PR3	>r	>st	wide-wider-widest	додавання частки коли основа закінчується на -е
PR4	< i>er	< i>est	sly-slier-sliest	додавання частки коли слово закінчується на –у
PR5	more < >	most < >	interesting – more interesting – most interesting	аналітичний спосіб
PR6	<***>	<***>	good-better-best	суплетивні форми

Розроблена система правил повністю задовольняє усі можливі випадки утворення парадигми англійських прикметників і прислівників, включаючи винятки, випадки, коли слово взагалі не має ступенів порівняння, а також коли утворює їх аналітичним або синтетичним способами.

Стосовно синтезу ніяких проблем не виникає через те, що кожне слово згаданих частин мови (за кількома винятками) має усталену форму утворення парадигми і повністю підпадає під відповідне правило словозміни. З іншого боку, потрібно відзначити, що при аналізі виникає проблема з омонімією (прикметника і прислівника, прислівника і прийменника, прикметника та іменника і т. д.). У зв'язку з цим без розроблення окремого алгоритму розпізнавання омонімії не можна стовідсотково автоматично опрацьовувати текст і розмежовувати ці частини мови, а потрібне втручання користувача у процес розпізнавання.

Суттєвою перевагою створеної програми є те, що вона може доповнюватись лексичними одиницями (оскільки зараз база даних містить лише прислівники та прикметники, які входять до навчального англо-українського словника II рівня).

Література

1. Использование ЭВМ в лингвистических исследованиях / Т. А. Грязнухина, Н. П. Дарчук, Н. Ф. Клименко и др.; Отв. ред. В. И. Перебейнос; АН УССР. Ин-т языковедения им. А. А. Потебни.— Киев: Наук. думка, 1990. — 288 с.

2. Крылова И. П., Гордон Е. М. Грамматика современного английского языка: Учебник для ин-тов и фак. иностр. яз.— 5-е изд.—М.: Книжный дом «Университет», 2000.— 488 с.— На англ. яз.
3. Рыбаков Ф. И. и др. Автоматическое индексирование на естественном языке/ Рыбаков Ф. И., Руднев Е. А., Петухов В. А. — М.: Энергия, 1980. — 160 с.
4. Карпіловська Є. А. Вступ до комп'ютерної лінгвістики / Донецький національний університет. — Донецьк: ТОВ «Юго-Восток, Лтд», 203. — 184 с.
5. Качалова К. Н., Израилевич Е. Е. Практическая граматика английского языка: В 2-х т. К.: Методика, 2000 г. — 368 с.
6. Положення про навчально-наукові та кваліфікаційні роботи студентів Київського державного лінгвістичного університету / Укл.: Болдирєв Р. В., Соловей М. І., Сажко Л. А., Спіцин Є. С. – К.: Вид. центр КДЛУ, 1999. – 46с.
7. M. A. Ganshina, N. M. Vasilevskaya English Grammar: Ninth edition revised – M.: Higher school publishing house, 1964. – 548p.
8. Освітні лінгвістичні Інтернет-портали: www.webster.commnet.edu; www.ego4u.com; www.paulnoll.com; www.vionet.hit.bg; www.wordpower.ws; www.algebra.com; www.fortunecity.com; www.ielanguages.com

О. Бугаков*
Український мовно-інформаційний фонд НАН України (Київ)
УДК 81'322.5

ВСТАНОВЛЕННЯ ЗОН ПРИЙМЕННИКОВИХ ЗВ'ЯЗКІВ В УКРАЇНСЬКОМУ ТЕКСТІ (МОДУЛЬ АВТОМАТИЧНОГО СИНТАКСИЧНОГО АНАЛІЗУ)

The principles of construcnting the algorythm for identification of prepositional connections zones in the Ukrainian text (module of parsing) are given. The algorythm has been constructed on the material of linguistic database as a subcorpus of the Ukrainian National Linguistic Corpus according to the analysis of prepositional connections zones in syntactic structure of the Ukrainian sentence.

Прийняте в Українському мовно-інформаційному фонді (УМІФ) поетапне здійснення синтаксичної розмітки лінгвістичного корпусу текстів базується на принципі часткової розмітки з використанням методу фракційного синтаксичного аналізу. Суть останнього полягає у побудові синтаксичної структури речення з окремих фракцій – синтаксичних конструкцій певного типу, а саме: ланцюжків слів, об'єднаних предикативним зв'язком, іменних безприйменникових конструкцій, прикметниково-іменникових конструкцій, прийменникових конструкцій (зон прийменникових зв'язків), на аналіз яких, власне, і спрямоване наше дослідження. Шляхом об'єднання синтаксичних конструкцій одержуємо повне представлення синтаксичної структури речення. Звідси й назва – фракційний синтаксичний аналіз.

Зона прийменникових зв'язків (ЗПЗ) містить прийменник, головне слово (тобто слово, яке керує прийменниково-іменниковою синтаксемою) і залежне слово (тобто слово, яке підпорядковується головному за допомогою прийменника).

Метою дослідження є побудова алгоритму встановлення зон прийменникових зв'язків в українському тексті (як окремого модуля автоматичного синтаксичного аналізу (АСА)) з урахуванням якісних та кількісних характеристик організації прийменникових зв'язків в українському тексті.

Виведення алгоритмічних правил визначення ЗПЗ передбачає проведення аналізу на репрезентативному матеріалі з метою забезпечення належного рівня достовірності результатів. Таким матеріалом слугував Український національний лінгвістичний корпус (УНЛК) обсягом 36 млн. слововживань, створений в УМІФі.

Програмне забезпечення УНЛК, побудованого як інформаційна система, дозволяє створювати спеціалізовані субкорпуси, орієнтовані на розв'язання поставлених завдань. За допомогою спеціально розробленої в УМІФі програми зазначені субкорпуси переводяться у формати баз даних з певною структурою, орієнтованою на проведення конкретних досліджень синтаксичної структури текста. Лінгвістичні бази даних (ЛБД), які виконують функцію інструмента і матеріалу дослідження мовного явища, структуровані за таким принципом: текстовим сегментам (контекстам), що містять конкретну мовну одиницю (прийменник), ставляться у відповідність визначені наперед диференційні ознаки, за якими здійснюється аналіз. Структурування ЛБД за полями, що

* © *О.Бугаков, 2006*

відповідають множині параметрів аналізу діагностуючих контекстів, та організація доступу до цих полів дозволяють автоматично класифікувати матеріал за кожним із параметрів та будь-якою їх комбінацією.

Інформаційною базою для визначення граматичних, позиційних та семантичних ознак компонентів ЗПЗ слугувала створена лінгвістична база зон прийменникових зв'язків (ЛБЗПЗ), сформована на субкорпусі обсягом 6 млн. слововживань. Обсяг бази – 20.768 речень. База структурована за полями, що відповідають множині параметрів, попередньо визначених прогнозуючими текстовими ознаками для алгоритмічного встановлення ЗПЗ при АСА. При формуванні полів цієї ЛБД виходили з того, що для визначення алгоритмічних правил АСА суттєвими і формально встановлюваними з тексту типами лінгвістичної інформації є: 1) морфологічна інформація текстових словоформ; 2) інформація про сполучувальні властивості граматичних класів слів у межах певних синтаксичних конструкцій; 3) дані про пунктуаційні засоби, які використовуються при структуруванні речення; 4) відомості про позиційні умови реалізації певних типів синтаксичних зв'язків; 5) лексико-статистична та 6) лексико-семантична інформація [4; 5]. Зазначеним типам інформації відповідають поля ЛБЗПЗ: 1) "Контекст", 2) "Довжина ЗПЗ", 3) "Перша позиція прийменника", 4) "Постпозиція ГС", 5) "Контактність ГС", 6) "ГС", 7) "Граматичний код ГС", 8) „Семантичний клас ГС", 9) "Контактність ЗС", 10) "ЗС", 11) "Граматичний код ЗС", 12) „Семантичний клас ЗС", 13) „Відношення", 14) „Ремарки" [3].

Виведені на основі ЛБЗПЗ якісні та кількісні характеристики організації прийменникових зв'язків в українському тексті [2] лягли в основу алгоритму встановлення зон прийменникових зв'язків в українському тексті. Їх також передбачається використати при ідентифікації прийменникових конструкцій на етапі синтаксичної розмітки УНЛК.

Розробка лінгвістичних алгоритмічних процедур в АСА будується з урахуванням принципів [1]: 1) доцільності (наприклад, для прийменників з перевагою постпозитивного розташування ГС слід починати пошук ГС саме з цієї позиції); 2) найбільшого виграшу (спочатку розглядаються зв'язки з найбільшою діагностуючою силою: наприклад, перевага контактного ГС-дієслова зумовлює початок перевірки саме з цієї ситуації); 3) частоти вживання (в алгоритмі частота прийменників враховується при визначенні ієрархії правил); 4) економічності (поступове ускладнення процедури перевірок).

Через те, що встановлене залежне слово може допомогти у пошуку головного, алгоритм встановлення ЗПЗ у тексті повинен починатися з визначення ЗС. По-перше, для залежного слова характерні обмежені синтагматичні контакти з прийменником, оскільки воно чіткіше визначене з погляду своєї позиції (переважно постпозиція стосовно прийменника) та граматичних характеристик (число класів, які можуть виконувати роль ЗС, обмежене), а отже, немає потреби залучати семантичну інформацію. По-друге, і це головне, результати цього етапу можуть використовуватися на наступних, зокрема, при використанні семантичних характеристик ЗС для пошуку ГС. Крім того, граматичні ознаки залежного слова можуть передбачати напрямок пошуку ГС і встановлювати порядок перевірки лівого і правого оточення прийменника.

При визначенні ЗС враховувалися результати попереднього аналізу функціонування прийменників у тексті: між прийменником і ЗС ніколи не може знаходитися жоден із компонентів предикативної пари; прийменник і ЗС не можуть знаходитися в різних предикативних частинах (ПЧ) складного речення.

Визначення ЗС починається з перевірки правого контексту прийменника на наявність у ньому іменника, займенника-іменника або числівника, які виконують функцію ЗС у 95, 9% усіх зафіксованих випадків у ЛБД: ***Навколо<PB> них<OT>*** *уже<Z0> сьогодні<T0> формується<YO> культ<MA> .<e>*

У випадку наявності одного з цих класів передбачається перевірка на узгодженість його відмінка з інформацією про відмінок, яким може керувати конкретний прийменник.

При негативному результаті перевіряється правий контекст на наявність у ньому цифри, абревіатури або скорочення, які переважно розташовуються в контактній позиції до прийменника: *Під час<PB> виступу<MB> в<PF> альма-матер<FF> Клінтон<JA> наголосив<VR> на<PF> необхідності<FF> продовження<NB> роботи<FB>* ***над<PE> ПРО<2A>*** *.<e>*

Для прийменників, у яких попередньо визначена можливість препозитивного розташування ЗС (*вслід/услід, навздогін/наздогін, навперейми, назустріч/навстріч,*

наперекір, напереріз, на зміну, на противагу, ради, заради), після перевірки правого контексту додатково перевіряється лівий на наявність у ньому контактно розташованих іменника та займенника-іменника у родовому (для *ради, заради*) або давальному (для інших прийменників) відмінках:

Правда<FA> йде<UO> ***Правді<FC> вслід<PC>*** *.<e>*

Після завершення роботи алгоритму пошуку залежного слова в алгоритм пошуку головного слова ЗПЗ передається інформація про місце розташування прийменника і ЗС в реченні і про граматичні характеристики останнього.

Межами інтервалу пошуку ГС при перевірці препозитивного контексту є: слово в початковій позиції речення або ПЧ, початок звороту (дієприкметникового або дієприслівникового), член предикативної пари, сполучення цифри з дужкою, яка вживається при переліку пунктів, підрядний або сурядний сполучник з попередньою комою. При пошуку ГС межі включаються в число об'єктів, які перевіряються на роль ГС. Межею постпозитивного пошуку є: присудок (оскільки для аналізу подається текст з установленим предикативним центром), останнє слово речення або ПЧ, для вставних конструкцій у дужках – дужка. При визначенні меж передбачено пропуск вставних конструкцій у дужках при лівому і правому пошуку і відокремлених зворотів – при лівому.

Пошук ГС починається з перевірки позиції прийменника. Якщо він знаходиться на початку речення або предикативної частини (ПЧ), головним словом визнається присудок (переважно виражений дієсловом, присудковим словом чи прикметником):

Незважаючи_на<PD> *дві<HV> поразки<FK>, ми<OS> не<Z0>* ***втратили<VU>*** *шансів<MI> знову<T0> вийти<VA> у<PD> чвертьфінал<MD> .<e>* – ГС у ЗПЗ прийменника *незважаючи на* виражене дієсловом *втратили.*

У випадку розташування прийменника не в першій позиції при виборі напрямку пошуку ГС повинні враховуватися дані про співвідношення ліво- та правобічних зв'язків конкретних прийменників. Як показав попередній аналіз, обидва способи розташування властиві майже всім прийменникам (крім *у справах, вглиб, на засадах, осторонь, у напрямі, усередину*, для яких було зафіксовано виключно лівобічне розташування ГС), тому правила аналізу повинні бути розраховані як на лівий, так і на правий контексти. Для прийменників з перевагою лівобічних зв'язків правила пошуку ГС у правому контексті повинні працювати в останню чергу, а для перерахованих вище прийменників в алгоритмі повинно передбачатися звернення до правого контексту як резервного правила, оскільки принципова можливість постановки ГС праворуч від прийменника не виключається. Для прийменників з перевагою правобічних зв'язків (*згідно, незважаючи на, пріч, згідно з, на відміну від, попри*) пошук ГС в правому контексті передує роботі правил, які аналізують лівий контекст. У цьому випадку відбувається пошук присудка праворуч від прийменника, роль якого переважно виконують дієслово, присудкове слово, прикметник чи дієприкметник:

Примітно<T0>, що<SS> ***згідно_з<PE>*** *латвійським<AE> законом<ME> про<PD> приватизацію<FD>, у<PF> країні<FF> не<Z0>* ***існує<UO>*** *заборонених<AT> до<PB> приватизації<FB> об'єктів<MI> ...<e>*

Для інших прийменників пошук ГС починається з аналізу лівого контексту. Як показав аналіз всіх ЗПЗ, при розташуванні дієслова, присудкового слова, прикметника чи дієприкметника в контактній позиції до прийменника в 99, 9% випадків вони виконують роль ГС.

При лівобічному розташуванні ГС важливою є інформація про перевагу у прийменників дієслівного чи іменного керування, яка була отримана внаслідок попередньо проведеного аналізу. Для прийменників, які віддають перевагу дієслівному керуванню (переважна більшість прийменників), за алгоритмом передбачається, насамперед, пошук дієслова або присудкового слова у дистантній позиції стосовно прийменника, але в межах інтервалу пошуку, оскільки, як показав матеріал, прийменники, які мають частку дієслівного керування більшу, ніж іменного, в тексті надають перевагу віддаленим зв'язкам з дієсловом (або його формами), ніж близьким зв'язкам з віддієслівними іменниками на відміну від прийменників, у яких в тексті переважає іменне керування: *Канівські<AS> судді<MH> дійшли<VU> висновку<MB>, що<SS> Володимир<JA> скоїв<VR> убивство<ND> у<PF> стані<MF> афекту<MB>, і<SC> звільнили<VU> його<OB> з-під<PB> варти<FB>, де<T0> він<OA>* ***перебував<UR>*** *під_час<PB> слідства<NB>* ***протягом<PB>*** *10<1A> місяців<MI> .<e>*

– головним словом у ЗПЗ прийменника *протягом* є не контактно розташований іменник з процесуальним значенням *слідство*, а дієслово *перебувати*, оскільки цей прийменник віддає перевагу дієслівному керуванню (84, 2% вживання в ЛБЗПЗ).

Для прийменників, у яких за попередніми даними частка іменної залежності більше, ніж дієслівної (*у справах*, *у галузі*, *з боку*, *на зразок*, *щодо*, *між*, *у справі* тощо), спочатку відбувається перевірка на пошук іменника з процесуальним значенням (розташованого як контактно, так і дистантно стосовно прийменника) у межах інтервалу пошуку:

Щоправда<T0>, в<PF> Головному<AR> столичному<AR> ***управлінні<NF> у_справах<PB>*** *захисту<MB> прав<NI> споживачів<MI> переконані<CS> : не<Z0> можна<X0> отоварюватися<YA> на<PF> стихійних<AX> ринках<MM> .<e>*

Виходячи з властивості проективності речень української мови (тобто неможливості перетину синтаксичних стрілок у структурі речення), виключається можливість іменного керування прийменником при вклинюванні між ними присудка, вираженого дієсловом, присудковим словом, прикметником чи дієприкметником. Відповідно, якщо ГС прийменника – іменник, то він може знаходитись тільки в інтервалі між присудком та прийменником. Якщо в ланцюжку слів ліворуч від прийменника не знайдено іменника, який задовольняє ознаки ГС-імені, а в інтервалі знайдено один з трьох класів, то останній і буде визначений як ГС прийменника.

Описаний вище принцип розташування правил є спільним для всіх прийменників. До спільних правил відноситься також використання семантико-граматичної інформації, зокрема, перевірка словоформ, розташованих ліворуч від прийменника, на входження до класів *істота* і *дія*. Перший формально може бути виведений за номером парадигматичного класу, що визначається для кожної текстової словоформи на етапі попереднього АМА тексту. У результаті проведеного аналізу було визначено прийменники, у яких функцію ГС може виконувати іменник-істота (*без*, *в/у*, *від*, *для*, *до*, *з*, *за*, *на*, *під*, *по*, *біля*, *з боку*, *над*, *поза*, *понад*, *протягом*, *у галузі*, *внаслідок*, *з погляду*, *з-за*, *з-під*, *з-поміж*, *поміж*, *у ролі*, *у справах*, *на зразок*, *у випадку*, *у порівнянні з*, *посеред*). Властивість використовуватиметься при визначенні пріоритетів при наявності кількох іменників-претендентів на роль ГС.

При визначенні класу *дія* семантичні ознаки виводилися з граматичних ознак слова, заданих його граматичною семантикою. Клас формувався за списком квазіфлексій віддієслівних іменників, що містить кінцеві буквосполучення слів та їх словоформ, за якими в межах граматичних класів іменників можуть визначатися слова на позначення *дії* (*-ання*, *-яння*, *-ення*, *-єння*, *-іння*, *-ація*, *-яція*, *-иття*, *-яття*, *-уття*, *-ництво*, *-ство*, *-ність*, *-илля*, *-овка*, *-обка; -анню*, *-янню* і т. д.)

Але оскільки ці квазіфлексії можуть мати слова з іншою семантикою (*населення*, *інформація*, *поняття*), передбачається перевірка претендентів на роль ГС за електронним семантичним словником прийменникових конструкцій.

Використання блоку правил для конкретних прийменників орієнтоване на неоднозначні ситуації, які репрезентують переважно прийменникове керування при дистантному розташуванні ГС. Сюди ж відносяться правила, які враховують лексико-статистичні характеристики прийменників, а також лексико-семантичні, які передбачають перевірку іменників на входження до семантичних класів (СК) – груп слів, які мають спільну сему в їх лексичних значеннях. Так, для прийменника *в/у* ГС-іменники, які не ввійшли до класів *істота* і *дія* (350 одиниць), було розподілено за 28 семантичними класами. Найбільшими з них є класи *процес* (*переговори*, *побудова*, *рибальство*), *подія* (*арешт*, *вибух*, *виступ*), *рух* (*виїзд*, *експорт*, *перехід*), *захід* (*виставка*, *нарада*, *турнір*), *ставлення/оцінка* (*невіра*, *недооцінка*). Семантичні класи задаються списками квазіоснов. При ідентифікації претендента на роль ГС із квазіосновою зі списку, він отримуватиме мітку головного слова в ЗПЗ.

Синтаксичну розмітку українського речення продемонструємо на прикладі. На вхід синтаксичного аналізатора поступає морфологічно розмічений текст: *У<PF> референдумі<MF> планується<YO> участь<FA> і<SC> палестинців<MI>, які<OS>* ***мешкають<UP> за<PE> межами<FL>*** *автономії<FB> .<e>*. Для зони зв'язків прийменника *за* синтаксична розмітка матиме такий вигляд: *мешкають<UP><10PR> за<PE><11PR> межами<FL><12PR>*, де поруч з морфологічною інформацією приписуватиметься номер залежного слова у реченні і тип синтаксичного зв'язку – *PR* (керування).

У ході встановлення ЗПЗ передбачається розв'язання проблем синтаксичної неоднозначності, а також усунення граматичної омонімії, яка залишилась після АМА.

Література

1. Аполлонская Т. А., Марашлец Е. Ф., Попескул А. Н. Устранение омонимии при АПТ // Инженерная лингвистика и оптимизация преподавания иностранных языков в вузе. – Л.: ЛГПИ, 1983. – С. 92–102.
2. Бугаков О. В. Зони прийменникових зв'язків у синтаксичній структурі українського речення // Мовознавство. – 2005. – № 5. – С. 75–87.
3. Бугаков О. В., Грязнухіна Т. О., Рабулець О. Г. Формування прийменникових текстозорієнтованих реляційних баз даних на корпусі українських текстів // MegaLing'2005 Прикладная лингвистика в поиске новых путей. Материалы международной конференции. 27 июня – 2 июля 2005 Крым, Меганом. – ТНУ им. В. И. Вернадского, Симферополь, 2005. – С. 19–20.
4. Грязнухина Т. А. Анализ предложных связей в научном тексте. – К.: Наукова думка, 1985. – 148 с.
5. Синтаксический анализ научного текста на ЭВМ / Под ред. Т. А. Грязнухиной. – К.: Наукова думка, 1999. – С. 31–32.

Н. Ерреро Бікус*
Київський національний лінгвістичний університет (Київ)
УДК81'322.5–6

АНАЛІЗ І СИНТЕЗ ПЕРФЕКТНИХ ФОРМ АНГЛІЙСЬКОГО ДІЄСЛОВА АКТИВНОГО СТАНУ

The paper focuses on the creation rules of perfect forms for determining paradigmatic classes of English verbs. The material is based on the English-Ukrainian dictionary of the second level provided by the Computational Linguistics laboratory. Had been used the paradigm of English verb, the created program characterizes with division word-change forms into paradigmatic classes of regular (12) and irregular verbs (56). It encapsulates the automatic analysis of perfect forms in the input text and the automatic synthesis of verbs had been taken into account main methods of verb expression.

Відомо, що англійське дієслово характеризується складною і розгалуженою системою словозмінних форм, яка становить значні труднощі для тих, хто вивчає англійську мову. Ця розгалуженість спричинена не так кількістю граматичних категорій, що в англійській мові не більша, ніж в інших індоєвропейських мовах, як специфічним співвідношенням граматичних категорій і засобів їх вираження.

Об'єктом дослідження в роботі є система дієслівних словозмінних форм англійської мови. П**редметом дослідження** – перфектний (доконаний) час англійської мови, правила утворення перфектних форм активного стану. Вибірка дієслів здійснена на основі англо-українського словника II рівня, розробленого в лабораторії комп'ютерної лінгвістики Київського національного лінгвістичного університету. М**ета дослідження** полягає у створенні програми, яка розпізнаватиме в тексті та синтезуватиме перфектні форми дієслів англійської мови.

Перфектні (доконані) часи позначають дії, що закінчилися раніше певного моменту в теперішньому (the Non-Past Perfect) чи минулому (the Past Perfect) і, як правило, пов'язані з цим моментом. Ці часи вважаються вторинними, оскільки виражають дію не саму по собі, а у співвіднесеності з якимось наступним моментом або дією. Доконані часи мають форму активного й пасивного стану, але в нашому дослідженні розглянуто лише дієслова активного стану.

Перфектні часи утворюються з відповідних неозначених часів допоміжного дієслова to have (в Non-Past Perfect (теперішній час) — have, has, це дієслово, яке використовується для утворення аналітичних форм дієслова; в Past Perfect (минулий час) — had) та дієприкметника минулого часу (Past Participle) основного дієслова. В англійській мові Non-Past Perfect ще називають Present Perfect на позначення теперішнього часу, проте майбутнього часу немає, є тільки форми вираження майбутнього часу. Дієслівна парадигма (ДП) — система словозмінних форм дієслова, що виражають граматичні категорії фінітності, способу, стану, часу, числа і особи [1: 13]. ДП в англійській мові досить складна і розгалужена, але точна кількість форм, що її утворюють, залишається

* © *Н. Ерреро Бікус, 2006*

невизначеною: вона залежить від трактування як граматичних категорій, так і поняття словозмінної форми. ДП притаманні такі риси:

1. Словозмінна форма повинна мати властивий тільки їй спосіб вираження, тобто свій набір елементів (включаючи питальний і окличний знаки).
2. ДП включає синтетичні й аналітичні форми.
3. Кожна форма може мати інваріант, регулярні й нерегулярні варіанти.
4. Непрямі способи словозмінними формами не вважаються, отже, в ДП не включаються.
5. Англійське дієслово не має словозмінних форм майбутнього часу й майбутнього в минулому.
6. Герундій об'єднується з дієприкметником у групу інгових форм.
7. Неособові форми можуть бути заперечними.
8. ДП має дві периферійні підсистеми: одна включає застарілі форми 2-ої особи однини і псевдопасивні форми дієслів руху, друга — get-пасив, що активно утверджується в мові.
9. Емфатичні форми з неповною інверсією є словозмінними і включаються у ДП.

Числення варіантів форм Perfect Active Tense. Нерегулярними варіантами є зміни втілення форми, яка залежить від індивідуальної вимови мовця. Регулярні ж варіанти форм визначаються за диференційними ознаками (ДО) п'яти видів:

ДО-1. Варіанти частки ***not,*** їх п'ять: ***not, n't, nt, n, na*** — і всі вони регулярні. Другий, четвертий варіанти ***not*** пишуться разом з допоміжним дієсловом.

У ряді випадків допоміжне дієслово, зливається з ***not*** і перетворюється в ***ain't***, ***hain't*** або ***ent***, що збільшує кількість варіантів заперечної форми, проте такі варіанти маловживані. В роботі використано основні варіанти заперечної форми, які є найчастотними***: have not, haven't, has not, hasn't, had not, hadn't***.

ДО-2. Варіанти допоміжного дієслова. В неминулому часі дієслово ***have*** має варіанти ***'ave***, ***av***, ***a***, ***'ve***, ***hev***, ***haf'***; форма ***has*** має варіанти ***'s***, ***'as***, ***hath***; у минулому часі ***had*** має варіанти ***'d***, ***'ad. В*** алгоритм включені такі частотні форми, як ', ***'ve***, ***'d***, і відповідно заперечні форми ***'not***, ***'ve not***, ***'d not.***

ДО-3. Варіанти основного дієслова. Ряд дієслів утворюють Participle II двома способами: ***born*** і ***borne***, ***proved*** і ***proven***, ***dreamt*** і ***dreamed*** і т. ін.

ДО-4. Перестановка елементів форми. Найчастіше трапляється перестановка ***not*** в питально-заперечній формі, як наприклад, в формі 148.

ДО-5. Зміна кількості елементів форми. Воно досягається опусканням допоміжного дієслова або частки.ДО-6. В деяких випадках перфектні часи утворюються за допомогою ***have got***, а не ***have,*** що призводить до збільшення варіантів цих форм.

Як і при будь-якому численні, при численні регулярних варіантів спостерігається надлишковість: обчислюються варіанти, теоретично можливі, але не ті, що реалізуються. Тому кількість варіантів значна і не всяка форма супроводжується формами [2: 13–28]. Урахувавши це, для початку були введені основні ДО — ДО-1, ДО-2, ДО-3, щоб уникнути надлишковості форм.

Правила утворення Past Participle регулярних і нерегулярних дієслів. Загальна кількість досліджуваних дієслів у англо-українському словнику II рівня — 795 (з варіантами форм становить 808 дієслів), з них 634 правильних і 161 неправильних дієслів.

Як відомо, в англійській мові розрізняють регулярні і нерегулярні дієслова. Регулярні дієслова — це правильні дієслова, які утворюються за правилами. Нерегулярні дієслова (їх ще називають неправильними) утворюються не за правилами регулярних.

Парадигматичні класи регулярних і нерегулярних дієслів за правилами утворення Past Participle. На основі вже створених правил утворення Past Participle були виділені парадигматичні класи для регулярних та нерегулярних дієслів. Таким чином, нараховується 56 парадигматичних класів нерегулярних дієслів і 12 парадигматичних класів регулярних дієслів.

Алгоритми автоматичного аналізу та синтезу мовної одиниці (дієслова). Алгоритм синтезу дієслова враховує всі основні форми вираження перфектного часу: стверджувальну, заперечну (з повною і скороченою формами допоміжного дієслова ***have not, haven't, has not, hasn't, had not, hadn't***), питальну, питально-заперечну (з повною і скороченою формами допоміжного дієслова) та емфатичну форми. Алгоритм при синтезі не видає варіанти форм ***'s, 've, 'd, 's not, 've not, 'd not*** з метою показати загальні правила утворення перфектного часу англійського дієслова активного стану.

Удосконалення програми можливо через урахування всіх ДО: 1) варіантів форм допоміжного дієслова *('ave, av, a, 've, hev, haf; 'as, hath)*; 2) варіантів частки *not (ain't, hain't ent)*; 3) способів утворення перфектного часу за допомогою *have (has) got.*

Таким чином, програма створювалася на основі статистичних розрахунків про вживання того чи іншого варіанта форми і може стати у пригоді тому, хто має на меті вивчення англійської мови.

Література

1. Дворжецкая М. П., Перебейнос В. И.. Методические рекомендации по изучению системных и функциональних характеристик словоизменительных форм английского глагола (Perfect Tenses Active). Выпуск 4. — Львов, 1995. — С. 13–28.
2. Мілих Н. Г., Перебийніс В. С., Рукіна Е. П. Морфологія англійського дієслова. — К.: Либідь, 1995. — С. 13.

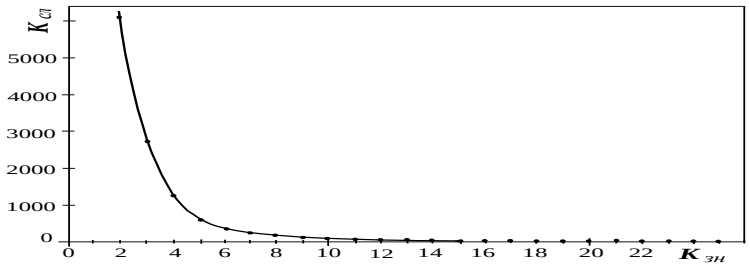
Я. Чаварга*
(Ужгород)
УДК 81'322.344

СЕМАНТИЧНИЙ ПОТЕНЦІАЛ ПРИКМЕТНИКІВ АНГЛІЙСЬКОЇ МОВИ

The article deals with the exponential dependence of the quantity of polysemantic adjectives in Modern English on the quantity of their meanings. The concrete form of this exponent is found and its linguistic interpretation is given. Three semantic potentials of English polysemantic adjectives are distinguished on the basis of our investigations: the usual, the increased and the high potential of creating new meanings.

Загальновідомо, що семантичний обсяг слова та кількість слів із певним числом значень, зафіксованих у словнику статистично пов'язані: найбільшу частку в ньому складають однозначні слова, далі йдуть (у порядку спадання частот) слова з двома, трьома і т. ін. значеннями. Такий розподіл є універсальною квантитативно-системною характеристикою полісемії природних мов. Вищезгадану статистичну залежність підтверджують і результати нашого дослідження, графічне зображення яких показано на рис. 1:

Рисунок 1. Залежність кількості прикметників від кількості значень.



За зовнішнім виглядом крива на рис. 1 нагадує відомий у математиці тип кривих, які називаються експонентами й описуються формулою:

$$y = a \cdot e^{-bx} \qquad (1)$$

де для нашого випадку y – кількість слів ($K_{сл}$ на графіку),
x – ступінь полісемічності ($K_{зн}$ на графіку),
a – константа, що визначає масштаб по осі y, точку перетину кривої з віссю y,
b – константа, що задає крутизну експоненти,
e – основа натурального логарифму ($e \approx 2,718$).

* © *Я. Чаварга, 2006*

Мета нашого дослідження – з'ясувати, наскільки строго залежність кількості слів від кількості значень узгоджується з експоненціальною залежністю, і знайти конкретний вигляд цієї експоненти та дати її лінгвістичну інтерпретацію.

Відзначимо, що експоненціальна залежність інтенсивності реакції (відчуття) від величини сили подразнення, в тому числі й у психофізичних експериментах, добре відома в науці, що вивчає поведінку живих організмів, – у біофізиці. Це так званий закон Вебера – Фехнера, який стверджує, що інтенсивність відчуття пропорційна натуральному логарифму сили подразнення [1:439]. У нашому випадку роль подразника відіграє поява нового предмета, явища чи властивості і, відповідно, потреба в номінації певного предмета чи явища засобами мови, а роль реакції – саме найменування (ім'я). Подібність нашої кривої до експоненти, а також низький рівень розкиду експериментальних точок вказує на те, що процес когнітивної діяльності людини в галузі творення нових значень слів на основі вже наявних не є хаотичним, безсистемним, а підпорядковується строгим закономірностям, характерним для живих організмів узагалі, і діяльності людини в напрямку творення полісемії зокрема.

З'ясуванню питання аналітичного виду залежності кількості слів від кількості значень (встановленню математичної формули) в літературі, в тому числі і в лінгвістичній, приділяють значну увагу, оскільки правильно визначена математична формула може сприяти з'ясуванню механізму творення полісемії, а відтак і впливу на неї когнітивної діяльності, що її визначає, оскільки при цій діяльності вона, власне, і твориться. Ф. Папп, 1967, [2] у своїх дослідженнях полісемії в угорській мові на базі тлумачного словника спробував представити результати досліджень як

$$y = V\,2^{-x} \tag{2}$$

де y – частка у словнику слів, що мають x значень,

V – обсяг словника.

Ю. А. Тулдава, досліджуючи полісемію в естонській мові [4], запропонував формулу, яка краще, ніж формула Ф. Паппа, узгоджується з експериментальними даними:

$$p(m) = a \cdot e^{-b \cdot m^{c}} \tag{3}$$

де $p(m)$ – частка слів із даною кількістю значень,

m – кількість значень слова,

a, b, c – параметри, що залежать від того, яка мова, яка частина мови та який словник чи текстовий масив досліджується.

У дослідженнях Ю. А. Тулдави $c \approx 0,5$, тобто формула має вигляд

$$p(m) = a \cdot e^{-b\sqrt{m}} \tag{4}$$

У всіх згаданих випадках робилася спроба представити результати експерименту через одну експоненту. Однак, якщо експериментальна або теоретична крива є результатом складання декількох простих експонент, то встановити це за зовнішнім виглядом графіка типу того, що зображений на рис. 1, важко, а в більшості випадків просто неможливо, оскільки при накладанні декількох експонент одна ділянка кривої плавно переходить у наступну. При цьому створюється враження наявності однієї кривої. Отже, при аналізі графіка на рис. 1 нам треба відповісти на запитання: чи є насправді дана експериментальна крива експонентою, чи тільки подібна на неї за формою (наприклад, гіпербола), або чи не є ця крива сумою декількох простих експонент типу

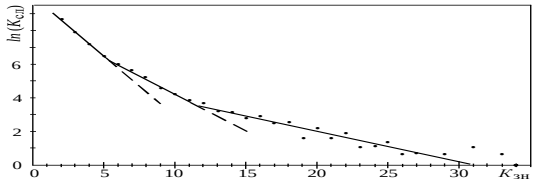
$$y = a \cdot e^{-b \cdot m}.$$

Для того, щоб відповісти на це запитання, криву треба прологарифмувати. У разі, якщо отриманий графік буде мати вигляд прямої лінії, це означатиме, по-перше, що вихідна (експериментальна) крива насправді є експонентою, і, по-друге, що процес творення багатозначності ймовірно є однаковим для всього масиву прикметників.

Якщо ж крива є комбінацією двох або більше експонент типу $y = a \cdot e^{-b \cdot x}$, то це означатиме, що творення полісемії відбувається за одним і тим самим законом, але матеріал (масив прикметників) по-різному робить свій вплив на дію закону. З курсу математики відомо, що в цьому випадку при логарифмуванні замість однієї прямої лінії отримаємо ламану лінію, при цьому прямолінійні ділянки ламаної відповідають тим

значенням аргументу, при яких дана експонента робить домінуючий внесок у загальну експоненту. Отже, щоб визначити, скільки різних експонент репрезентує крива на рисунку 1, представимо її у логарифмічній шкалі. На рис. 2 показано графік залежності кількості слів від кількості значень у логарифмічній шкалі.

Рисунок 2. Залежність кількості прикметників від кількості значень у логарифмічній шкалі.



Як бачимо, логарифмічний графік досить впевнено можна інтерпретувати як комбінацію трьох прямих ліній, а це свідчить про те, що на кожній із ділянок домінує одна з трьох експонент, тобто в кожному випадку утворення нових значень слова відбувається за законом Вебера – Фехнера. Як видно з графіка, вплив першої експоненти є домінуючим до кількості значень 5. Від кількості значень 5 і до 12 переважає друга експонента, і, нарешті, від 12 й вище домінує третя експонента.

Графік на рисунку 2 не тільки вказує нам на наявність трьох різних причин, що впливають на форму кривої, але й уможливлює встановлення загального аналітичного виду залежності кількості слів від кількості значень. Оскільки на логарифмічному графіку наявні три лінійні ділянки, це дає нам підстави зробити припущення про те, що вираз, яким можна аналітично представити експериментальну криву, повинен мати вигляд:

$$y = y_0 + y_1 e^{-b_1 m} + a_2 e^{-b_2 m} + a_3 e^{-b_3 m} \qquad (5)$$

За допомогою програми ORIGIN 7 ми отримали формулу для інтерпретації залежності кількості прикметників від кількості значень в англійській мові:

$$y = 35499 \cdot e^{-1,007 m} + 3049 \cdot e^{-0,395 m} + 36 \cdot e^{-0,099 m} \qquad (6)$$

Результати розрахунків за цією формулою разом з експериментальними даними представлені в таблиці:

*Таблиця. Р*озрахунки внеску кожної експоненти формули (6) в загальну криву і порівняння з експериментальними даними

К-сть значень	exp1	exp2	exp3	exp1+ exp2+ exp3=	екс-пери-мент	теорія/ експ.
2	4737	1384	30	6151	6124	1, 004
3	1731	932	27	2690	2740	0, 982
4	632	628	24	1285	1275	1, 008
5	231	423	22	677	638	1, 061
6	84	285	20	390	389	1, 003
7	31	192	18	242	261	0, 927
8	11	129	16	158	174	0, 908
9	4	87	15	107	103	1, 039
10	2	59	13	75	68	1, 103
11	1	40	12	53	47	1, 127
12	-	27	11	39	42	0, 929
13		18	10	29	24	1, 208
14		12	9	22	23	0, 957
15		8	8	17	12	0, 895
16		5	7	14	19	0, 737

К-сть значень	exp1	exp2	exp3	exp1+ exp2+ exp3=	екс-пери-мент	теорія/ експ.
17		4	7	11	12	0, 917
18		2	6	10	13	0, 769
19		2	6	7	5	1, 400
20		1	5	6	9	0, 667
21		1	5	5	5	1, 000
22		-	4	5	7	0, 667
23			4	5	3	1, 667
24			3	4	3	1, 333
25			3	4	4	1, 000
26			3	4	2	2, 000
27			3	3	2	1, 500
28			2	3	0	------
29			2	3	2	1, 500
30			2	2	0	------
31			2	2	3	0, 667
32			2	2	0	------
33			1	2	2	1, 000
34			1	2	1	2, 000
35			1	2	1	2, 000
36			1	1	1	1, 000

Незважаючи на те, що формула (6) має вигляд складного виразу, насправді це простіша залежність порівняно з формулою Ю. А. Тулдави (4), оскільки являє собою суму простих експонент, у той час як у формулі Ю. А. Тулдави наявна нелінійна залежність показника степеня від кількості значень, тобто $\sqrt{m}$, лінгвістичну інтерпретацію якій знайти важко, якщо взагалі можливо.

У випадку, коли математична залежність між двома величинами має вигляд декількох доданків, як у (6), використання математики дозволяє встановити внесок кожної складової в загальну криву – для формули (6) можна встановити величину внеску кожної з трьох експонент у сумарну експоненту. На рисунку 3 графічно представлено результати таких обчислень. Так, графік 1 побудований як:

$$y_1 = \frac{a_1\, e^{-b_1 m}}{a_1\, e^{-b_1 m} + a_2\, e^{-b_2 m} + a_3\, e^{-b_3 m}} \tag{7}$$

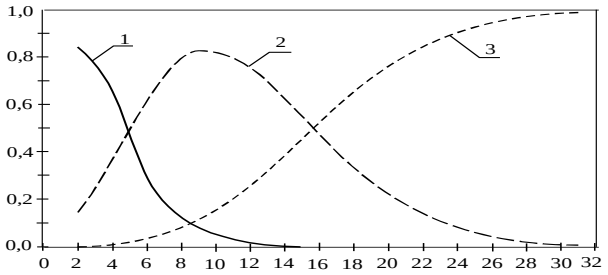
графік 2 як:

$$y_2 = \frac{a_2\, e^{-b_2 m}}{a_1\, e^{-b_1 m} + a_2\, e^{-b_2 m} + a_3\, e^{-b_3 m}} \tag{8}$$

а графік 3 як:

$$y_3 = \frac{a_3\, e^{-b_3 m}}{a_1\, e^{-b_1 m} + a_2\, e^{-b_2 m} + a_3\, e^{-b_3 m}} \tag{9}$$

Рисунок 3. Графіки внеску окремих експонент формули (5) у загальну криву, програма ORIGIN 7: 1 – внесок першої експоненти; 2 – внесок другої експоненти; 3 – внесок третьої експоненти.



Той факт, що наша експериментальна крива залежності кількості прикметників від ступеня полісемії виявилася сумою трьох експонент, свідчить на користь гіпотези про те, що на утворення нових значень, крім інших факторів, впливає семантичний потенціал слова. Як правило, слова при утворенні є однозначними, але в ході подальшого вживання вони можуть послідовно реалізувати свій асоціативно-семантичний потенціал у вигляді нарощування числа своїх значень. На матеріалі російської мови при аналізі даних словника „Новые слова и значения" [3] висуваються припущення про вичерпання потенціалу семантичного розвитку слів, сповільнення темпу утворення нових значень слів при просуванні до вищих семантичних зон аж до повної зупинки на певному етапі. Ми припускаємо, що наявність трьох експонент у згаданому законі пов'язана з тим, що слова не просто мають внутрішній потенціал творення багатозначності, але ще й діляться на декілька типів за величиною цього потенціалу, за різними закладеними в них потенційними можливостями творення полісемії.

Аналізуючи графіки на рисунку 3, бачимо, що перша крива описує внесок у полісемію прикметників англійської мови з порівняно невисоким (назвемо його звичайним) потенціалом творення багатозначності. Друга крива показує прикметники з дещо вищим потенціалом творення полісемії (назвемо його підвищеним потенціалом), які складають порівняно менш частотну групу. Третю групу прикметників представляють слова з високим семантичним потенціалом, (див. третю криву).

Ми скористаємося отриманими графічними результатами для того, щоб дати відповідь на запитання, що насправді являють собою ці потенціали, і в чому полягає їхня лінгвістична суть. Згідно з графіком 1 на рисунку 3, для з'ясування лінгвістичної суті першого потенціалу потрібно проаналізувати слова з полісемією 2–5. Переважна більшість цих слів повинні мати спільну ознаку, яка і складе характеристику для першого семантичного потенціалу прикметників. Це основна кількість досліджуваних нами слів (10777 із 12019, або 89, 67 %). Визначення когнітивних моделей цих слів через вивчення та зіставлення словникових дефініцій показали, що більшість із них, які входять до діапазону 2–5 значень, тлумачаться за таким словниковим шаблоном: 1) of or pertaining to …; 2) related (relating) to …; 3) caused by …; 4) having (composed of, made of) …; 5) being in … тощо. Водночас наші спостереження показали, що основний масив прикметників цього діапазону складають багатоморфемні (похідні та складні) слова – 10116 із 10777, ~94%. Простих прикметників у цьому діапазоні відносно небагато – усього 661, ~6%, що за графіком 3 на рис. 3 майже відповідає внеску слів із високим потенціалом. Отже, кількість багатоморфемних слів приблизно у 15, 3 раз більша, ніж кількість одноморфемних. Цей факт дає підставу для висновку про те, що афіксально марковані прикметники мають звичайний семантичний потенціал. У словах із першим (звичайним) потенціалом чітко проявляється вплив словотвірного фактора на полісемію.

Другий семантичний потенціал досягає максимуму (~82% внеску) при кількості значень 9–12. Отже, відповідно до графіка 2 на рис. 3, суть лінгвістичного змісту другого (підвищеного) потенціалу полісемії треба шукати, аналізуючи слова з кількістю значень 9–12. Наші дослідження показали, що переважна більшість цих ад'єктивних одиниць є абстрактними, тобто такими, що позначають абстрактні ознаки. Наприклад: *Christian*, *classical*, *drawn*, *exclusive*, *imperfect*, *legal*, *personal*, *primitive*, *privy*, *progressive*, *sovereign*, *spiritual*. *В*ідношення між простими й афіксально маркованими прикметниками змінюється: перепад між їхньою кількістю суттєво переміщується у бік зростання кількості простих слів. Кількість багатоморфемних одиниць перевищує кількість одноморфемних у 4, 65 разів проти 15, 3 разів у першій групі, що добре узгоджується з ходом графіків на рис. 3.

Принагідно зауважимо, що в англійській мові прикметників із підвищеним потенціалом творення багатозначності приблизно в десять разів менше, ніж прикметників зі звичайним потенціалом. Це дає вагому підставу для висновку про те, що однією з головних причин появи у прикметників більш високого (підвищеного) потенціалу творення полісемії є узагальнення поняття до абстрактного рівня.

Графік 3 на рисунку 3 вказує на те, що лінгвістичну природу високого семантичного потенціалу потрібно шукати у списку слів у діапазоні від 16 значень і більше. Ця група прикметників є порівняно нечисленною. У досліджуваному нами матеріалі нараховується 99 слів із багатозначністю 16 і вище, що складає всього 0, 82 % від загальної кількості прикметників. Ці 99 слів дають 2154 значення, або 5, 4 % від сумарної кількості значень. Ми проаналізували слова з кількістю значень 20 і більше як такі, що є

найхарактернішими представниками ад'єктивної лексики з високим семантичним потенціалом. Ці прикметники можна об'єднати в декілька груп. Більшість прикметників із високим потенціалом полісемії є конкретними, основні значення яких характеризують ті чи інші сторони базових понять природознавства, що у процесі когнітивної діяльності були поширені й на інші предмети. Найпомітнішу частину прикметників із високим семантичним потенціалом складають ті з них, що позначають просторово-часові відношення, наприклад: *high, short, quick, running,* ознаки з точки зору сили, міцності, гостроти: *strong, weak, tight, loose, set,* слова на позначення відчуттів, наприклад: *cold, dry, light, soft, hard, heavy* та інші. Досить помітну частину високополісемічних прикметників складають оцінні слова, наприклад: *nice, fine, fair, good, foul* та інші.

Очевидним є той безперечний факт, що основна маса прикметників із високим потенціалом – прості (90 %). І тільки незначна частина слів цього діапазону є афіксально маркованими, точніше, похідними, що цілком узгоджується з рисунком 3. Складних ад'єктивних одиниць із високим семантичним потенціалом не виявлено. Наші дослідження також показали, що переважна більшість цих слів – англосаксонського походження.

На основі проведених досліджень ми робимо висновок про те, що всі полісемічні прикметники англійської мови можна умовно поділити на три групи залежно від величини притаманного їм потенціалу творення нових значень. Першу групу складають прикметники зі звичайним потенціалом полісемії, другу групу утворюють слова з підвищеним потенціалом, третю групу становлять прикметники з високим потенціалом полісемії, що зумовлено словотвірними та семантичними чинниками. Як бачимо, нескладний математичний аналіз статистичних даних дозволив нам зробити цілком конкретні лінгвістичні висновки. Впадає у вічі той факт, що семантичний потенціал наростає не плавно, а стрибкоподібно, тобто величина потенціалу квантується (змінюється дискретно) – звичайний → підвищений → високий, що дає змогу стверджувати про наявність своєрідного „квантового ефекту в лінгвістиці".

У перспективі доцільно провести аналогічні дослідження семантичного потенціалу різних лексико-граматичних класів слів у різних мовах, а також здійснити порівняльний аналіз семантичного потенціалу певних груп слів.

Література

1. Биофизика / Костюк П. Г., Гродзинский Д. М., Зима В. Л., Магура И. С., Сидорик Е. П., Шуба М. Ф. – К.: Выща школа, 1988. – 504 с.
2. Папп Ф. О некоторых количественных характеристиках словарного состава языка // Slavica VII. – Debrecen, 1967. – С. 51–58.
3. Поликарпов А. А., Каримова Г. О. Словарь «Новые слова и значения» (НСЗ–1984) как источник данных об эволюции языка // Прикладная лингвистика и автоматический анализ текста. Тез. докл. научн. конф. – Тарту, 1988.
4. Тулдава Ю. А. О некоторых квантитативно-системных характеристиках полисемии // Учен. зап. Тартуского ун-та. – Вып. 502. Linguistica XI. – Тарту, 1979. – С. 107–141.

С. Курстак*
Київський національній лінгвістичний університет (Київ)
УДК 81'322

АЛГОРИТМ ЗНЯТТЯ ОМОНІМІЇ ТИПУ «ІМЕННИК-ДІЄСЛОВО» В АНГЛІЙСЬКІЙ МОВІ ТА ІМПЛЕМЕНТАЦІЯ ЙОГО В ПРОГРАМНОМУ ПРОДУКТІ

Актуальність статтьи зумовлена тим, що жодне точне дослідження словникового складу мови не може обійтися без глибоко розгляду питань омонімії. Не менш важливою є й розробка проблем омонімії для вирішення задач інженерної, структурної, комп'ютерної лінгвістики в цілому.

Омонімія (від гр. homos – однаковий і onyma – ім'я) – це прояв особливої якості мовного знаку: різним означуваним можуть відповідати тотожні означення. Через це омонімія створює труднощі в процесі комунікації. Часто слухач опиняється в складній ситуації: яке з декількох значень, які виражаються певною мовною формою, треба обрати для правильного розуміння повідомлення. Труднощі можуть виникати не лише в слухача (читача), а й в мовця (письменника), який задля оптимізації процесу спілкування

* © *С.Курстак, 2006*

намагається будувати повідомлення так, щоб його сприйняли однозначно. Окрім іншого, омонімія додає складності у вивченні іноземної мови, коли студент стикається з такою проблемою: одна й та сама мовна форма може мати абсолютно різні значення – факт, на який в рідній мові він зазвичай не звертає уваги. Аналіз таких форм суттєво погіршує сприйняття іншомовного тексту або повідомлення.

Значні труднощі додає омонімія при створенні систем автоматичного аналізу тексту, а також в лексикографічній роботі: при виборі одиниць для частотного словника, при визначенні меж між окремими словами в загальному словнику мови, при розробці способів представлення в ньому омонімічних слів і т.п.

Повне вирішення проблем омонімії наразі є недосяжною мрією, але воно розпочне революцію в найрізноманітніших галузях науки і повсякденного життя: від машинного перекладу і штучного інтелекту до побутових роботів і автоматичних систем забезпечення життя та надання послуг. Всі ці системи оперують з мовним матеріалом і нерідко – з звуковим. Такі системи зазвичай не мають права на помилку, тому повідомлення які вони сприймають/відсилають повинні бути стовідсотково однозначними. І хоча вивчення омонімії ведеться вже давно, нажаль в цій області досягнення невеликі. А час вимагає поглибленого дослідження проблем омонімії, вияву її внутрішніх закономірностей, визначення місця, яке вона займає в загальній системі мови.

Основним структурним елементом омоніміки в мові є **омонімічний ряд. О**монімічним рядом називається сукупність слів (лексем) або словоформ однієї й тієї ж мови в один і той самий період його існування, взаємно тотожних в плані вираження, але суттєво різних в плані змісту.

Омонімічний ряд – це свого роду макросистема, цілісність якої базується на загальності форми її елементів. Вивчення омонімічного ряду може йти як по лінії дослідження його системних характеристик, так і по лінії аналізу його структури. В першому випадку омонімічний ряд розглядається як складна єдність, і тоді нас цікавлять його якості як цілого, в другому ж він розглядається як розчленована єдність, і тоді нас цікавить його склад і внутрішня організація. В курсовій роботі переважно використовується другий підхід.

Омоніміка мови представляє собою сукупність всіх омонімічних рядів мови, одні з яких знаходяться між собою в визначених системних відносинах (наприклад, ряди вихідних і похідних омонімів або ряди основних і похідних омоформ), інші – без жодного зв'язку між собою. Таким чином, омоніміка в цілому не утворює єдиної системи. Інтенсивний розвиток омонімії в англійській мові очевидно спричинений більш ніж одним фактором. Насправді, існує кілька непов'язаних між собою причин, таких як односкладовий характер англійської мови і її аналітична структура.

Велика кількість омонімів також тісно пов'язана з такою характерною рисою англійської мови як фонетичною ідентичністю слова і основи, чи, іншими словами, предомінантністю вільних форм серед найчастотніших коренів. Факт, що частотність слів знаходиться в оберненій залежності від їх довжини, вказує на те, що односкладні слова будуть найчастотнішими. Крім того, так як найчастотніші слова полісемантичні, природнім є те, що вони можуть набути значень, які суттєво відрізнятимуться від вихідних. Коли посередницькі відношення руйнуються, деякі з цих нових значень можуть втратити будь-які зв'язки з рештою структури і розпочати незалежне існування. Цей феномен відомий під назвою *дезінтеграція* або *розділення полісемії*.

Причини виникнення омонімії можна розділити на дві групи:

– омонімія, що викликана конвергентним розвитком звуку, коли два чи три слова різного походження випадково співпадають в звуковому складі;

– омонімія як розвинута полісемія, через дивергентний розвиток значення (включаючи конверсію).

Обидві причини можуть поєднуватися з втратами закінчень і іншими морфологічними процесами. Як приклад можна навести два латинських дієслова: cadere 'падати' і capere 'тримати(ся)', в першому з яких, з плином часу 'd' змінилося на 'р', частково під впливом аналогічної форми з старо-французької мови. Такий тип омонімії неважко впізнати. Існування другого типу – питання до дискусії. В. І. Абаєв вважає омонімічними лише ті словоформи, які є різними з точки зору етимології. Всі інші – просто полісемічні. Багато інших вчених не згодні з В. І. Абаєвим і наполягають на необхідності семантичних і структурних критеріїв для розрізнення омонімії і полісемії.

На відміну від ‘capere’ члени омонімічного ряду ‘box’, які утворилися внаслідок дезінтеграції, є етимологічно поєднаними. Ідентичність форми – невипадкова; вона базується на генетичних зв’язках. Всі слова цього ряду (5 за даними Оксфордського словника) походять одне від одного і простежуються до латинського ‘buxus’.

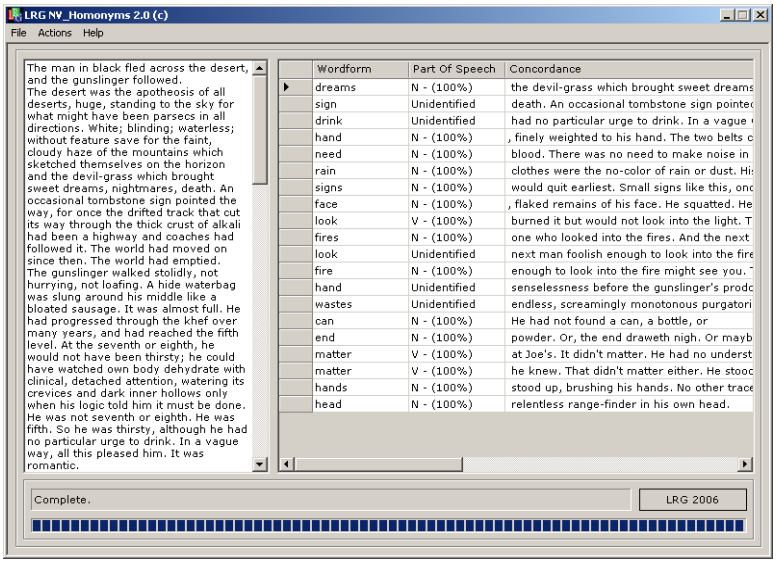
Такі омоніми можуть частково або повністю походити один від одного, але спільна точка їх зародження знаходиться за межами англійської мови. В цих словах, з появою нового значення, яке суттєво відрізняється від інших, семантична структура вихідного слова розділяється. Нове значення починає незалежне існування і отримує власну семантичну структуру. Звідси й термін *дезінтеграція*.

Підбиваючи підсумок, омонімія типу «іменник-дієслово», яка викликана з одного боку процесом втрати закінчень і т.п. і з іншого – конверсією (див. Таб. 1 з аналізом походження та конверсії омонімічних форм), є однією з найвизначніших характеристик сучасної англійської мови.

У роботі проводилося дослідження омонімічного складу англо-українського навчального словника другого рівня. Омоніми досліджувалися як групи, виділені за типом омонімії. Предмет дослідження: 266 омонімічних пар типу «іменник-дієслово» англійської мови. На основі проведеного дослідження запропонувати спосіб вирішення омонімії типу «іменник-дієслово» на прикладі 266 омонімічних пар.

Для проведення дослідження з англо-українського навчального словника другого рівня вибрано всі омонімічні пари, трійки, четвірки та п’ятірки. Омоніми розділено на групи за типом і подано їх частоту. На основі найчастотнішої групи «іменник-дієслово» укладено механічну вибірку з 266-ти омонімічних пар. Засобами власноруч створеної програми на мові програмування C# запропоновано спосіб автоматичного зняття омонімії в її частинному варіанті «іменник-дієслово» для 266-и омонімічних пар. Омонімічні словоформи закладено в масив. Окрім 266 вихідних форм, в масиві присутні їх омонімічних похідні (наприклад: “advance”, “advances” і т.п.);

Отримавши текст, програма розділяє його на словоформи і, з ціллю полегшення доступу до нього, записує в базу даних. В базі даних кожне слово отримує своє унікальне місце і номер. З текстом в такому вигляді, наступною операцією є порівняння словоформ введеного тексту з омонімічними формами в масиві. При відповідному збігові починає роботу модуль “Determine”, в якому зібрані диференційні, діагностичні ознаки та властивості відповідних частин мови, за якими однозначно (або з високою ймовірністю правильної відповіді) можна зняти омонімію. За допомогою низки простих правил в алгоритм “Determine” включена також система визначення вірогідності правильності результату. На виході користувач отримує впорядковану в вигляді списку інформацію про омонімічні форми, які зустрілись в введеному тексті і визначену частину мови для кожної з них з відповідною вірогідністю.



Модуль "Determine" розділений на 2 частини:

набір правил для іменника (NDefining)

- Якщо слово подане з великої літери і стоїть не на початку речення;
- Якщо слово закінчується на "'s", "s'".
- Якщо попереднє слово – неомонімічний прикметник (вибірка з 760 найчастотніших прикметників з англо-українського навчального словника 2-ого рівня);
- Якщо попереднє слово – артикль;
- Якщо попереднє слово – присвійний займенник;
- Якщо попереднє слово – числівник (кількісний або порядковий);
- Якщо попереднє та наступне слова в різних варіаціях – займенники, прислівники, прийменники...

набір правил для дієслова (VDefining)

- Якщо попереднє слово – особовий займенник;
- Якщо попереднє слово – допоміжне дієслово;
- Якщо попереднє слово – прислівник, а перед ним – особовий займенник.
- Різноманітні комбінації часток, прийменників і т. д.

В кожному наборі правила також поділяються на *графічні (однозначні)*, які визначають частину мови безпосередньо за графічними показниками або іншими, які можуть вказати на частину мови без контекстного аналізу, та *синтаксичні (контекстні)*, які визначають частину мови за контекстом, сполучуваністю і т. д.

Згадані вище ознаки майже стовідсотково визначають частину мови, в той час як інші надають лише часткову вірогідність і спрацьовують лише в взаємному поєднанні. Кількість комбінацій суттєво велика. Ця робота обмежена розміром вибірки і масштабами дослідження. І все ж запропонований способів зняття омонімії типу «іменник-дієслово» спрацьовує для даного масиву омонімічних форм з високою вірогідністю правильного результату.

А. Прыстаўка*
Беларускі дзяржаўны універсітэт (Мінск)
УДК 81'322

ТЫПАЛОГІЯ СЕМАНТЫЧНЫХ ПЕРАЎТВАРЭННЯЎ ТЭКСТАЎ

Зварот да вызначэння тыпаў семантычных пераўтварэнняў тэкстаў абумоўлены важнасцю гэтага пытання для патрэб камп'ютэрнай лінгвістыкі. Даследаванні асаблівасцей будовы чалавечага мыслення можа растлумачыць далейшыя кірункі працы па ўдасканаленні перакладных праграм, праграм рэферыравання тэкстаў і г. д.

Прац, прысвечаных даследаванню семантыкі, дастаткова шмат і ў лінгвістыцы, і ў логіцы, але, на нашу думку, яны носяць больш філасофскі, чым практычны характар.

Мы разглядаем пераўтварэнні семантыкі, грунтуючыся на законах мовы як сістэмы, даказаных Агульнай тэорыяй сістэм (АТС) Ю. А. Урманцава і пацверджаных на розных узроўнях у межах сістэмнай лінгвістыкі У. А. Карпавым [гл. 1; 2].

Разгледзім з'яву тэкставых рэмінісцэнцый (ТР) у газетных загалоўках, пабудаваных на падставе прэцэдэнтных тэкстаў (ПТ). Прэцэдэнтны тэкст – гэта інварыянт сітуацыі, які ў выглядзе закончанага выказвання захоўваецца ў свядомасці носьбіта мовы ў гатовым выглядзе. Самымі распаўсюджанымі прыкладамі прэцэдэнтных тэкстаў з'яўляюцца прыказкі, прымаўкі, выслоўі, крылатыя выразы, афарызмы і г. д. Загалоўкі публіцыстычных артыкулаў, створаныя на падставе прэцэдэнтных тэкстаў, прыцягваюць увагу, выклікаюць адпаведныя асацыяцыі, правакуюць цікавасць да артыкула, а таму даволі папулярны ў сучасным перыядычным друку. Разгледзім газетныя загалоўкі сучаснай беларускамоўнай прэсы, пабудаваныя на аснове прэцэдэнтных тэкстаў літаратурнага ці фразеалагічнага характару ("*Ці загляне сонца ў наша аконца", "А хто там стаіць?", "Зямля пад надзейнымі крыламі"* і г. д.).

Змены ці пераўтварэнні, якія адбыліся з прэцэдэнтнымі тэкстамі ў загалоўках, падпарадкаваны сістэмным законам. Згодна АТС "аб'ект-сістэма можа або заставацца ў

**© А. Прыстаўка, 2006*

спакоі, або змяняцца. Пры фіксаваным законе кампазіцыі ў аб'екце-сістэме па вызначэнню нельга змяніць нічога іншага, акрамя колькасці, якасці, адносінаў адзінства "першасных" элементаў (усіх або часткі). Камбінаторыка з 3 асноўных пераўтварэнняў дае 8 іх варыянтаў. На аснове гэтага сцверджання фармулюецца асноўны закон пераўтварэння аб'екта-сістэмы: "Аб'ект-сістэма ў межах сістэмы аб'ектаў таго ж кшталту дзякуючы свайму існаванню пераходзіць або ў сябе праз тоеснае пераўтварэнне, або ў іншыя аб'екты-сістэмы – праз адно з сямі і толькі сямі разнастайных пераўтварэнняў, а менавіта: 1. колькасці, 2. якасці, 3. адносінаў, 4. колькасці і якасці, 5. колькасці і адносінаў, 6. якасці і адносінаў, 7. колькасці, якасці, адносінаў усіх або часткі яго "першасных" элементаў. Уся сукупнасць сістэмных пераўтварэнняў складаецца, такім чынам, з аднаго тоеснага і сямі нятоесных" [1: 92]

Пад колькаснымі зменамі (К) мы будзем лічыць змены даўжыні кампазіцыі (павелічэнне ці скарачэнне элементаў) за кошт паўназначных слоў (бо службовыя служаць для ўсталявання адносінаў ці злучэння, а не сэнсу) адносна зыходнай.

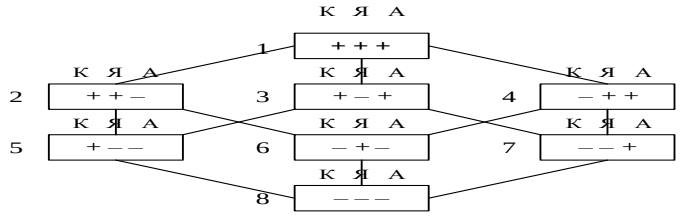
Пад якаснымі зменамі (Я), больш складаны від, – змяненне значэння слова, замена лексічнай адзінкі, некалькіх слоў ці выразаў новымі. Дадаванне новых паўназначных слоў не будзе лічыцца якаснымі зменамі, таму што адбываецца колькаснае змяненне кампазіцыі.

Пад зменамі адносінаў (А) мы будзем разумець змены паслядоўнасці элементаў кампазіцыі, змены граматычнай функцыі словаформы ў межах кампазіцыі, змены сінтаксічнай структуры выказвання, змены ці знікненне рыфмоўкі, змены рытмічнай арганізаванасці кампазіцыі. Гэты аспект мае значэнне, бо рытмічнае гучанне, рыфма нясуць часта функцыю запамінання і распазнавання кампазіцыі, а іх змяненне істотна ўплывае на прыцягненне ўвагі чытача (адна з галоўных мэт выкарыстання рэмінісцэнцый у загалоўках).

На схеме, пабудаванай на аснове мадыфікаванага кода Грэя, гэта будзе выглядаць наступным чынам (знак '+' паказвае на наяўнасць, знак '–' – на адсутнасць характарыстыкі):

Падсістэма 1, тоесныя змены. Змены, якія не закранулі ніякі з відаў. Гэта штуршок да з'яўлення магчымых светаў. Сюды трапілі ТР як яны ёсць па форме, але якія нясуць новую семантыку. Напрыклад, у артыкуле пад назвай "*Герой нашага часу"* апавядаецца пра сучаснае пакаленне і яго ідэалы; радкі вядомай у савецкія часы песні *"З чаго пачынаецца Радзіма?"* сталі назвай конкурсу юных мастакоў, і трапілі ў загаловак артыкула пра гэтую падзею; вядомы эўфемізм *"Рамантыкі з вялікай дарогі"* даў назву артыкулу пра бандытаў. Такім чынам, хоць у прыведзеных прыкладах і назіраецца пэўная адпаведнасць сэнсу прэцэдэнтнага тэксту і ТР, але гэта ўсе ж зусім іншая сітуацыя, а таму семантычныя пераўтварэнні маюць месца.

Схема 1.



Падсістэма 2, адносныя змены. Тут размясціліся кампазіцыі, дзе змяніўся парадак размяшчэння элементаў, ці іх граматычная форма, рытміка і рыфмоўка кампазіцый, змянілася сінтаксічная будова, ці адносіны паміж адзінкамі кампазіцыі. Змяненне граматычнага ліку назіраецца ў прыкладзе *"Белыя вароны";* у загалоўку *"Нам застанецца спадчына"* ў параўнанні з зыходным тэкстам ("Нам засталася спадчына") змяніўся час дзеяслова; прыклад *"Свіння гусям не таварыш"* ілюструе і змяненне парадку размяшчэння элементаў, і змяненне ліку ("Гусь свінні не таварыш"). Змененая назва вядомага выдавецтва пачатку 20 ст. "Загляне сонца і ў наша аконца" трапіла ў гэтую

падсістэму, таму што ў ёй змянілася сінтаксічнае афармленне: з апавядальнага па структуры сказа яно стала пытальным: "*Ці загляне сонца ў наша аконца?"*. Прыкладаў, дзе б назіралася толькі змяненне рытмічнага афармлення ў чыстым выглядзе нам не трапілася, але такія прыклады будуць у іншых падсістэмах, дзе мае месца змена адносінаў (падсістэмы 5, 6, 8).

Падсістэма 3, якасныя змены. Тут знаходзяцца кампазіцыі, якія змяшчаюць змены значэння асобных слоў ці лексічныя адзінкі: *"Да нас едзе Кавэрдэйл"* у арыгінале "К нам едет ревизор"*; "І толькі надпіс – "Мілавіца"* радок з песні гурта "Песняры" "І толькі надпіс "Вераніка"*; "Навука на барыкадах"* назва вядомай карціны "Свобода на баррикадах". ТР *"А ці быў вірус?"*, які ўзыходзіць да рэфранічнага выразу рамана М. Горкага "Жизнь Клима Самгина" "А был ли мальчик?", на першы погляд, мае і змяненне парадку размяшчэння элементаў, але мы не ўлічвалі асаблівасці перакладу выразаў на/з беларускай мовы, а сінтаксічная канструкцыя з часціцай "ли" мае ў беларускай мове адпаведную канструкцыю з часціцай "ці" і іншым парадкам слоў.

Падсістэма 4, колькасныя змены. Сюды трапілі кампазіцыі адрозныя па даўжыні ад зыходнай за кошт толькі паўназначных адзінак мовы (не службовых слоў). У нас гэта падсістэма атрымалася другой па малаколькаснасці, у ёй толькі 5 прыкладаў. Патлумачыць гэта можна неактуальнасцю для мовы ўвогуле такога параметра як колькасць у чыстым выглядзе, бо часцей мае значэнне колькасны параметр у сукупнасці з іншымі, напрыклад, якасцю. У ТР *"Зямля пад белымі крыламі стала другой Радзімай"* да назвы артыкула У. Караткевіча ("Зямля пад белымі крыламі") далучылася частка, якая павінна была ўдакладніць тэму артыкула. Выраз аднаго з герояў фільма "Ирония судьбы" з удакладненнем стаў назвай артыкула пра асаблівасці камунальнага гаспадарання на Беларусі: *"Цёпленькая" пайшла. Ці надоўга?";* артыкул пра аднаўленне помнікаў архітэктуры цытуе Біблейскія радкі з невялікім дадаткам: *"Час збіраць і аднаўляць камяні"*.

Падсістэма 5, якасна-адносныя змены. У гэтай падсістэме размясціліся кампазіцыі, якія змяшчаюць і змяненне сэнсу і змяненне парадку ці граматычных параметраў новай кампазіцыі адносна зыходнай. Напрыклад, "*Залатая пятля Расеі"* у назве экскурсійнага маршрута "Золотое кольцо России" змянілася адно слова, якое стала прычынай змянення роду прыметніка*;* ТР *"Кансэрватары ўсіх краін, не яднайцеся!"* узыходзіць да палітычнага лозунга Маніфеста камуністычнай партыі і разам са змененым словам мае і зменене сінтаксічнае аблічча за кошт адмоўнай часціцы "не". У ТР *"Што табе падарыць, мой выкладчык дарагі?"*, якая паходзіць з радка песні («Что тебе подарить, человек мой дорогой…») акрамя якасных зменаў слова назіраецца таксама змяненне парадку размяшчэння элементаў, што дае падставы змясціць гэту ТР у падсістэму якасна-адносных пераўтварэнняў.

Падсістэма 6, колькасна-адносныя змены. У гэту падсістэму трапілі выразы, адрозныя па колькасці элементаў і парадкам, але без якасных змяненняў элементаў. Хоць гэтая падсістэма ў нас атрымалася самай малаколькаснай (толькі 4 прыклады), але ТР тут досыць цікавыя па характары пераўтварэнняў. Устойлівы выраз часоў савецкай улады "битва за урожай" даў назву артыкулу, але з дадаваннем новай лексемы, што пацягнула склонавае змяненне назоўніка "бітва": *"Ахвяры бітвы за ўраджай"*. З назвы фільма «Свой среди чужих, чужой среди своих» знікла адна частка, гэтым тлумачацца колькасныя змены, а таксама род займенніка:*"Свая сярод чужых"*. Назва рамана Маркеса «Сто лет одиночества» у беларускамоўным артыкуле набыў выгляд *"Стагоддзе адзіноты"*. Колькаснае змяненне (тры лексемы ў прэцэдэнтным тэксце-дзве ў ТР) справакавала і адносныя змены: словазлучэнне-слова.

Падсістэма 7, колькасна-якасныя змены. У гэтых кампазіцыях змяніліся і самі элементы, і колькасць элементаў адносна зыходнай. Радкі верша Пляшчэва не змянілі ў ТР сваёй рытмікі і рыфмоўкі, але колькасць паўназначных элементаў зменшылася, некаторыя словаформы змяніліся іншымі, але адносных зменаў мы тут не назіраем, бо сінтаксічныя адносіны апошніх прэдыкатыўных частак нельга параўнаць:*"Кактус зелянее, сонейка блішчыць, на маім акенцы кветачка стаіць"*. Прыказка "На бязрыб'і і рак – рыба" ў ТР скарацілася і змяніла сваё лексічнае аблічча:*"На бязрыб'і і баран…*. Тое самае назіраецца і ў прыказцы «Через тернии – к звёздам», дзе змяніліся словы і дадалося адно новае: *"Праз церні эфіру — да «Тэлевяршыні»"*.

Падсістэма 8, колькасна-якасна-адносныя змены. Сюды трапілі кампазіцыі, якія менш за ўсё падобны да зыходнай, бо ў іх іншы парадак размяшчэння элементаў ці

граматычныя формы слоў, іншыя словы ці выразы і іншая колькасць элементаў. Біблейны выраз «не бывает пророк без чести, разве только в отечестве своём», які пазней трансфармаваўся ў выраз “нет пророка в своем отечестве”, у загалоўку набыў выгляд: *“Дык што, няма мастака ў Айчыне сваёй?”*. Тут мы бачым і новую лексему (мастак), і іншую сінтаксічную канструкцыю (пытальны сказ), і змяненне парадку элементаў, а таксама і павелічэнне кампазіцыі. Пераўтвараная назва рамана Э. Хемингуэя «На Западном фронте без перемен», у загалоўку змянілася колькасна (са спалучэння назоўніка з прыназоўнікам атрымалася спалучэнне займенніка і прыметніка), якасна (западный-Маскоўскі, перемен-нічога новага) і адносныя змены ў рытме фразы, хоць яна і не рыфмаваная, але яскрава адчуваецца іншая сістэма лагічных націскаў: *“На Маскоўскім фронце нічога новага”*.

Запоўніўшы ўсе падсістэмы канкрэтным матэрыялам мы можам зрабіць наступныя вынікі. Самыя невялікія прыклады далі падсістэмы 2 (колькасныя), 4 (адносныя) і 6 (колькасна-адносныя) пераўтварэнні, адпаведна 6, 5 і 4 прыклады са 134 разгледжаных намі. Відавочна, што дзве з падсістэм перасякаюцца ў трэцяй, гэтым тлумачыцца невялікая колькасць прыкладаў падсістэм 2 і 4, якая ў падсістэме 6 яшчэ больш знізілася.

Наступнымі па колькасці прыкладаў з’яўляюцца падсістэмы 8, колькасна-якасна-адносныя змены дзе змяшчаецца 10 ТР. Мы тлумачым невялікую колькасць запаўнення тым, што даследаванне, якое тычыцца семантычных зменаў, спецыфічнае па сваёй сутнасці. Падсістэма 7, колькасна-якасныя змены, мае 12 ТР. Гэта, на наш погляд, тлумачыцца самой па сабе вялікай колькасцю якасных зменаў, якія і паўплывалі на дадзеныя гэтай падсістэмы, бо колькасныя змены, як згадана вышэй, маюць малы працэнт выкарыстання.

Па 22 прыклады маюць падсістэмы 1 і 5, тоесныя і якасна-адносныя змены. Першы тып зменаў канстатуе наяўнасць у нашай сістэме месца і для мадэляў, якія не маюць фармальнага выражэння зменаў, такім чынам пацвярджаецца універсальнасць сістэмнага падыходу. Таксама відавочна паступовае развіццё сістэмы ў бок ускладнення пераўтварэнняў, пачынаючы з тоесных. Даволі значную долю тоесных зменаў мы тлумачым на падставе сіметрычна-асіметрычнага характару мовы, які праяўляецца ў развіцці яе сістэмы ад вядомых, у нашым выпадку фармальна тоесных кампазіцый, да асіметрычных, тых, што імкнуцца пазбавіцца агульных параметраў з зыходнай. У працэнтным суаднясенні тоесныя:нятоесныя змены складаюць суаднесенасць 1:6, ці 16, 4% да 83, 6%, што, на наш погляд, з’яўляецца даволі заканамерным і пацвярджае дадзеныя іншых даследаванняў сіметрычна-асіметрычнага характару мовы. Якасна-адносныя змены (5 падсістэма) маюць таксама 22 прыклады. Якасныя змены, якія прыводзяць да замены лексічнага складу ТР адносна зыходнай, выклікаюць і адносныя змены, звязаныя са зменай граматычнай афарбоўкі іншых кампанентаў ТР ці яе сінтаксічнага аблічча.

Самая вялікая і ў колькасным і значная ў працэнтных суадносінах 3 падсістэма якасных зменаў мае 53 прыклады, што складае 39, 6% ад агульнай колькасці. Такі вынік можна растлумачыць імкненне аўтараў ТР захоўваючы максімальнаю блізасць, а тым самым і пазнавальнасць з прэцэдэнтным тэкстам, надаць сваё, аказіянальнае значэнне вядомаму выразу. У табліцы нашы дадзеныя размесцяцца наступным чынам

Табліца 1.

№ падсістэмы	колькасць	працэнт
1	22	16, 4%
2	6	4, 5%
3	53	39, 6%
4	5	3, 7%
5	22	16, 4%
6	4	3%
7	12	8, 9%
8	10	7, 5%

Прааналізаваўшы шэраг кампазіцый, пабудаваных на аснове прэцэдэнтных тэкстаў, мы прыйшлі да наступных вынікаў. Пераважная большасць кампазіцый ствараецца шляхам змянення лексічнага складу (якасныя змены). На другім месцы змяненне тоесныя, якія не закранаюць фармальны бок тэкстаў і змены якасна-адносныя. Кампазіцыі,

утвораныя шляхам дадавання ці выдалення элементаў, змянення парадку ці граматычнай афарбоўкі, а таксама спалучэнне гэтых відаў з'яўляюцца самымі малакольканымі, а значыць такі спосаб стварэння кампазіцый можна лічыць непрадуктыўным.

Такім чынам, сістэма семантычных пераўтварэнняў складаецца з 1 тоеснага і 7 нятоесных пераўтварэнняў, кожны з якіх мае запаўненне ў той ці іншай ступені моўным матэрыялам, а значыць прапанаваная сістэма можа лічыцца поўнай.

Літаратура

1. Карпов В. А. Язык как система. Мн.: Выш. шк., 1992. – 302 с.
2. Урманцев Ю А. «Девять плюс один этюд о системной философии». – М.: «Современные тетради», 2001. – 168 с.

Н. Коваль, М. Лангенбах*
Київський національний університет (Київ)
УДК 81'322

ПАРАДИГМАТИЧНІ І СИНТАГМАТИЧНІ ХАРАКТЕРИСТИКИ ПРИЙМЕННИКОВИХ ТА БЕЗПРИЙМЕННИКОВИХ ДІЄСЛІВНИХ СПОЛУЧЕНЬ

У статті пропонується опис сполучуваності у вигляді корпусу текстів із автоматизованою та автоматичною системою аналізу. Передбачено, що за допомогою комп'ютерних засобів буде організовано масив конструкцій з контекстним оточенням у гнучку систему, у якій буде доступним сортування за різними параметрами, подаватимуться для кожного окремо взятого елемента та моделі в цілому граматична, лексико-семантична та статистична інформація.

Сполучуваність – явище універсальне, притаманне усім мовним рівням, але на кожному рівні вона має свою специфіку, відповідно до особливостей тих елементів, з якими вона «має справу». Скажімо, *фонетична* сполучуваність – це поєднання фізичних та фізіологічних факторів із мовно-етнічними, *морфологічна* та *синтаксична* сполучуваність – явище, яке характеризується двобічністю, оскільки одиниці, що сполучаються на цих рівнях, мають два плани – план вираження та план значення, формальний та змістовий рівень. Семантична сполучуваність (а також і синтаксична, до певної міри), окрім загальних уявлень людини про навколишній світ, включає у себе також закони логіки.

Досліджуючи й описуючи сполучуваність, ми тим самим робимо спроби зафіксувати та відтворити принципи перетворення мови у мовлення, реалізації потенційних зв'язків мовних елементів у зв'язному висловлюванні. І якщо сполучуваність – це формальні закони реалізації валентних можливостей, реальне їх втілення, то валентність – це потенційні можливості мовної одиниці. Таким чином, можна сказати, що валентність – явище парадигматичне, а сполучуваність – синтагматичне, валентність ближча до мови, сполучуваність – до мовлення.

Якщо говорити про частиномовну сполучуваність як один із аспектів мовної сполучуваності загалом, то варто зазначити, що кожна частина мови має певні обмеження на сполучення з іншими. Проте лише частиномовної характеристики сполучуваності недостатньо. Крім приналежності до тих чи ніших лексико-граматичних класів, можливість чи неможливість сполучення слів характеризують їх граматичні форми. Синтаксис багатьох мов вимагає певних змін форми слова для узгодження його з іншими словами, певного порядку слів, допоміжних словоформ тощо. Як слушно зазначає В. Ш. Рубашкін, «деталізація опису синтаксичної сполучуваності зводиться до зазначення, – принаймні, для деяких із можливих конфігурацій, – додаткових умов синтаксичного зв'язку» [7].

Але і формально-граматичного критерію недостатньо. Якщо брати до уваги лише формальні показники, ми можемо отримати фразу граматично правильну, проте беззмістовну. З точки зору синтаксису та морфології можливими є набагато більше конструкцій, ніж насправді, з урахуванням значення кожного з елементів та всього висловлення (згадаймо Щербу з його «глокой куздрой» або породжуючі граматики Хомського). Тому необхідним є введення до правил сполучуваності ще й елементів семантики. Рубашкін підкреслює, що реалізація семантичної сполучуваності може бути

* *©Н.Коваль, М. Лангенбах, 2006*

визначена перш за все як необхідність посилення обмежень на синтаксичну сполучуваність.

Семантична сполучуваність може бути представлена як додаткові умови сполучуваності синтаксичної, певні коригуючі правила. Найочевидніше рішення у такому випадку: оформлювати семантичний опис у вигляді певних семантичних класів та формул їх комбінації. Присвоївши кожному слову певні семантичні мітки, можна було б поєднати правила синтаксичні з правилами семантичними.

Прийменниково-відмінкова граматична форма – аналітичний спосіб вираження граматичного значення, а отже, формальне вираження зв'язку. Проте, враховуючи, що граматичне значення до певної міри пов'язане з лексичним, оскільки обидва вони містять у собі елементи реального смислу, варто говорити не тільки про граматичне, а й про семантичне значення членів сполуки.

Опис сполучуваності можна представити у вигляді корпусу текстів із автоматизованою та автоматичною системою аналізу. За допомогою комп'ютерних засобів можна організувати масив конструкцій з контекстним оточенням у гнучку систему, в якій буде доступним сортування за різними параметрами, подаватимуться для кожного окремо взятого елемента та моделі в цілому граматична, лексико-семантична та статистична інформація.

Проект, який зараз перебуває на стадії розробки у лабораторії комп'ютерної лінгвістики, присвячений дієслівній сполучуваності, зокрема дієслівно-іменним та дієслівно-прийменниково-іменним сполукам. Ми зупинимося лише на тих випадках, коли дієслово виявляє однозначно домінуючу роль у словосполученні тобто поєднується із залежним словом зв'язком керування, хоча ми свідомі того, що поняття сполучуваності дієслова включає в себе усі його зв'язки з іншими словами, отже, варто було б досліджувати й ті іменники, які виконують у реченні синтаксичну функцію підмета.

Першим кроком було створення бази дієслівних сполук. За основу бралися тексти української публіцистики, оскільки публіцистичний стиль включає у себе елементи усіх інших, тобто є найбільш універсальним з-поміж усіх функціональних стилів мови. Масив вибраних конструкцій був організований у базу даних, яка включає у себе контекстуальні відомості, граматичні відомості (морфологічні коди та складені з них синтаксичні моделі), лексичні та семантичні відомості, частотність слів, які складають аналізовані конструкції, а також частотність виявлених морфолого-синтаксичних та семантико-синтаксичних моделей.

Відповідно до перелічених вище елементів бази даних аналіз дієслівних сполук розподіляється на кілька етапів:

1. Відбір необхідних конструкцій. З комплексу публіцистичних текстів за допомогою програмного забезпечення було відібрано дієслівно-іменникові та дієслівно-прийменниково-іменникові конструкції. Таблиця бази містить у собі поля дієслова, прийменника та іменника, леми дієслів, частотність дієслів, прийменників та іменників, а також контекст.

2. Морфолого-синтаксичний аналіз матеріалу. Усім наявним словам, що входять у конструкції, автоматично приписуються морфологічні коди, після чого складаються і обраховуються морфолого-синтаксичні моделі.

Морфолого-синтаксичний аналіз матеріалу проводиться за рядом параметрів:

- дистантність (між дієсловом й іменником);
- наявність чи відсутність у конструкції прийменника;
- позиція іменника відносно дієслова (пре- чи постпозиція);
- кількість залежних іменникових груп (складність конструкції, яка визначається валентністю дієслова);
- частотність моделей та слів, які входять до їх складу;

3. Лексико-семантичний аналіз. Складовим елементам конструкцій ставляться у відповідність певні лексичні значення та семантичні коди, внаслідок чого створюється синоптична схема дієслів (класифікація дії за напрямком, тривалістю, інтенсивністю, якістю тощо), іменників (світ, природа, предмет, річ, речовина, матеріал, середовище та ін. з подальшим структуруванням цих лексико-семантичних груп) та прийменників. Після цього складаються та проходять статистичну обробку семантико-синтаксичні моделі аналізованих сполук.

4. Створена програма, яка дозволяє користувачеві оперувати базою даних. Інтерфейс включає кілька полів:

- ✓ початкова форма дієслова;
- ✓ прийменник (факультативно заповнене поле, залежно від типу конструкції);
- ✓ початкова форма іменника;
- ✓ лексико-семантична група дієслова;
- ✓ лексико-семантична група прийменника;
- ✓ лексико-семантична група іменника;
- ✓ пошук за граматичним кодом (можна задати повну граматичну модель або лише один з її елементів);

Усі ці поля доступні у двох режимах – введення потрібного слова (параметру) вручну чи пошук через комбобокс.

- ✓ поле «Контекст», в якому міститься приклад обраної конструкції, поданий у контексті, достатньому для повного розкриття значення сполуки;
- ✓ кнопка «Пошук», яка здійснює пошук за обраними параметрами;
- ✓ кнопка «Додати модель», яка дозволяє доповнити базу сполук;
- ✓ кнопка «Додати приклад»;
- ✓ «Статистика за» (вибрати поле);
- ✓ Кнопка «тезаурус».

В кожному полі можна задати вихідні параметри, за якими в базі данних здійснюватиметься пошук (за морфологічним класом, за встановленою користувачем початковою формою дієслова чи іменника чи за прийменником, за відмінком іменника, за лексико-семантичною групою дієслова, іменника чи прийменника).

Кнопка «тезаурус» переводить програму у режим тезаурусного пошуку. Робота у режимі тезаурусного пошуку відбуватиметься наступним чином.

При переході у цей режим, користувачеві відкривається вікно із такими полями:

1. Загальна синоптична схема, яка має деревовидну структуру.
2. Оскільки кожне поняття представлене безвідносно до частини мови, якою представляється (наприклад, дієслово „читати” й іменник „читання” представляють одне поняття таких категорій, як *інтелект. діяльність (Ді)*, *процес (Ац)*, хоча належать до різних частин мови, і входять до морфолого-синтаксичних моделей як керуючий і залежний елементи – *читати книгу*, *припинити читання*), то для зручнішого користування є поле вибору, де можна обрати потрібну синоптичну схему, враховуючи й частиномовну характеристику, тобто – дієслівну, іменникову чи прийменникову схеми – шляхом натискання на відповідну кнопку («дієслова», «іменники», «прийменники»).
3. Поле синоптичної схеми, де показуються слова обраної частини мови, організовані у деревовидну структуру за семантичними групами та показано, яким чином вони поєднуються зі словами інших категорій. При натисканні на відповідній категорії з'являється додаткове вікно з прикладами та статистикою.

Ця програма має чітке прагматичне спрямування і покликана задовольняти потреби користувачів залежно від їхніх інтересів і рівня знань. Для користувачів, які погано володіють мовою і хотіли б її вивчити, програма стане у нагоді як навчальний посібник. Програма буде корисною і для науковців, які зможуть на основі проводити власні дослідження, моніторинг, змінювати чи розширювати конфігурацію параметрів програми за власними потребами.

Рисунок 1. Інтерфейс програми.

Література

1. Апресян Ю. Д. О толковом словаре управления и сочетаемости русского глагола// Словарь. Грамматика. Текст. – М., 1996
2. Апресян Ю. Д. Формальная модель языка и представление лексикографических знаний //Вопросы языкознания. – 1990. – №6
3. Вихованець І. Р. Прийменникова система української мови. – К., 1980
4. Грязнухина Т. А., Дарчук Н. П. Синтаксический анализ в системе автоматической обработки научных текстов// Актуал. пробл. компьютер. лингв./ Тез. докл. Всес. конф./ Тарту, 29–31 мая 1990 г. – Тарту – 1990.
5. Москвин В. П. Семантика и синтаксис русского глагола. Автореферат диссертации на соискание ученой степени доктора филологических наук. – Краснодар, 1993
6. Перебийніс В. С. Теоретичні та прикладні проблеми структурної лінгвістики // Мовознавство. – 1981. – № 4.
7. Рубашкин В. Ш. Представление и анализ смысла в интеллектуальных информационных системах. http://www.lingmar.narod.ru/library/rubashkin.html.

КОМП'ЮТЕРНЕ РОЗПІЗНАВАННЯ Й СИНТЕЗ УСНОЇ МОВИ

Н. Бардіна, д. філол. н., ***І. Каташинська****
Одеський національний університет ім. І. І. Мєчнікова (Одеса)
УДК 81'322.5

СИСТЕМНІСТЬ ЛІНГВІСТИЧНОГО АНАЛІЗУ ПРИ ПРОВЕДЕННІ ІДЕНТИФІКАЦІЙНИХ ЕКСПЕРТИЗ

The modern linguistic criminalistic examination refers to the classical multilevel language model. Methodology based on the discourse model has been suggested for efficiency increase.

Поняття лінгвістичної експертизи лише нещодавно було запроваджено до мовознавчої практики [див. 1; 2; 4]. Це новітній напрям прикладної лінгвістики, який знаходиться на межі досліджень лінгвістики, криміналістики, акустики, авторознавства, дискурсології, когнітології тощо.

Сучасні проблеми розбудови держави, з одного боку, та проблеми формування шляхів міждисциплінарних розроблянь, експансія лінгвістичного знання, з іншого, обумовлюють **актуальність** пошуків адекватних методів встановлення ідентифікації особистості на підставі відносно невеличкої кількості ознак.

Взагалі, коли говорять про експертизу, мають на увазі дослідження, у центрі якого знаходиться певна проблема, що для свого вирішення потребує особливих (наукових, професійних) знань. Її сутність полягає в аналізі матеріальних об'єктів, а також різноманітних документів з метою встановлення фактичних даних, що мають значення для вірного з'ясування справи. [16: 5]

Залежність експертизи від технічного прогресу робить постійними зміни її змісту, сприяє розширенню її можливостей, переліку завдань та проблем, які вона спроможна вирішити. В результаті і сама система експертної діяльності стає динамічною: з'являються нові експертні методики, а, отже, і нові типи самої експертизи.

Так, довідник "Судово-експертна діяльність", виданий Міністерством Юстиції України у 2001 році, подає наступну класифікацію криміналістичних експертиз з точки зору предмета експертизи:

1.почеркознавча, авторознавча;
2. технічна експертиза документів;
3. фототехнічна;
4. технічна експертиза матеріалів і засобів звукозапису;
5. портретна;
6. трасологічна;
7. балістична;
8. вибухових пристроїв, матеріалів і речовин.

Окремим різновидом, який не відноситься до класу „криміналістична", подається „психологічна" експертиза. [14: 91].

* © *Н. Бардіна, І. Каташинська, 2006*

Треба зазначити, що в недавньому минулому дослідження мовлення обмежувалось почеркознавчою, авторознавчою та фоноскопічною експертизами. Кожна з них торкається лише одного параметру вивчення відповідно письмового чи усного мовлення. Почеркознавча здатна встановити певні риси особистості, але її об'єктом може бути лише письмовий текст і тільки в рукописному варіанті. [13: 496]. Таким чином, в ситуації поширення використання електронних засобів зв'язку звернення до почеркознавчої експертизи втрачає свою актуальність.

Авторознавча експертиза також досліджує переважно письмове мовлення, хоч в останнє десятиріччя у криміналістиці з'являються спроби дослідження в її межах і усного мовлення також. Але, згідно з традицією проведення цього типу експертиз, у фокусі свідомості дослідника перебувають лише дискретні параметри мовлення.

Особливості побудування мовленнєвого потоку, що притаманні конкретній людині, залишаються поза увагою експертів, і таким чином, авторознавча експертиза може вирішити тільки питання авторства і встановити деякі ознаки певної особистості: походження, статеву приналежність, ступінь освіти; але не торкається, наприклад, проблем стану свідомості в момент створення тексту.

Проблемі фоноскопічної експертизи – експертизи, яка встановлює тотожність мовця на основі його усного мовлення, – присвячено достатню кількість досліджень (див., наприклад, роботи О. Ф. Дьяченко, О. І. Галяшиної, Л. В. Златоустової, М. В. Китайгородської, Н. М. Розанової, Р. К. Потапової, О. Р. Шляхова та ін.).

Фоноскопічна експертиза враховує, перш за все, акустичні характеристики мовного сигналу, фонетичне втілювання вислову. Традиційно фоноскопічні експертизи проводили фізики-акустики, інженери-радіотехніки: довгий час вважали, що ідентифікацію мовця можна провести, спираючись саме на акустичні ознаки – спектральні характеристики, частоту основного тону, швидкість вимови тощо. Але дуже швидко стало ясно, що ці ознаки мовленнєвого сигналу найчастіше незіставні: оригінальна суперечлива фонограма і зразки голосу підозрюваного отримували в різних умовах, з використанням різних засобів магнітофонного запису.

Але ж лінгвістами вже давно доведено, що емоційний стан, соціальне оточення, відносини між комунікантами, сам тип мовлення (спонтанне або підготовлене, читання готового тексту) значно впливають на тембральні, інтонаційні характеристики, на вимову окремих звуків [див., наприклад, 9].

Виникла нагальна потреба використати власне лінгвістичні знання, і експерти-фоноскопісти звернулись до підручників з мовознавства. Але при цьому зовсім не враховувалося те, що традиційні посібники і підручники було зорієнтовано на ієрархічну модель мови, мовлення переважно оцінювали як реалізацію мовної системи, з точки зору відповідності літературним нормам. [докладніше див. 4].

Таким чином, повз уваги залишались особливості розгортання саме дискурсу, різноманітні хезитації (мовленнєві коливання), які вже з 70 років XX століття описували як маркери кодування інформації, а не як "слова-паразити" з точки зору моделі літературної мови.

Експерти з фоноскопії і зараз не аналізують особливості синтаксичної, семантичної, когнітивної організації дискурсу як комплексної ситуації породження мовлення; не враховують новітньої антропоцентричної парадигми сучасної науки, що значне обмежує інформативність досліджуваного матеріалу.

Нерідко експерт не спроможний взагалі винести експертний висновок, посилаючись на "недостатність фактичного матеріалу".

Наприклад, матеріали одної з експертиз показали, що в суперечливих фонограмах всі мовці (мешканці одного міста Півдня України, всі з вищою освітою) в неофіційній обстановці вживали в російському мовленні [Г] фрикативне, [W] білабіальне, велику кількість інвектив, а при відборі зразків голосу певно дотримувалися літературної норми, тому, спираючись на традиційну методику аналізу, на ті параметри, які пропонували комп'ютерні програми Sound Forge або "Діалект", неможливо було визначити, кому власне належить голос, записаний на плівку з телефону.

Настала потреба ввести новий тип експертизи, яку б виконували фахівці з лінгвістики, обізнані в новітніх тенденціях моделювання мовних явищ. Про це йшла мова на великій кількості наукових конференцій [див., наприклад, 8].

Тому цілком закономірним стало офіційне введення в 2002 році нової спеціальності в експертній практиці – лінгвіст-експерт [11]. Ситуація, на наш погляд, при проведенні

індифікаційних та класифікаційних (пошукових) експертизах значно покращилась.

Однак і досі немає єдиної думки, закріпленого пояснення, який саме аспект лінгвістики доцільно впроваджувати до процесу криміналістичної практики і як новий тип лінгвістичної експертизи пов'язати зі вже відомими комп'ютерними ідентифікаційними програмами Cool Edit і Sound Forge, що традиційно були вживані під час проведення фоноскопічних експертиз.

Тому ми і звернулись до створення **нової, оригінальної** методики проведення *системних* лінгвістичних експертиз у криміналістиці, яка вже пройшла випробування на практиці протягом трьох років.

Уперше такого роду експертизи було проведено на матеріалі російської мови у полілінгвальному середовищі (Південь України) на базі Одеського науково-дослідного інституту судових експертиз, співробітником якого є один з авторів даної статті.

На нашу думку, сучасний стан наукового знання вимагає рішення будь-якої проблеми не в межах конкретної дисципліни, а у міждисциплінарному просторі, який охоплює цю проблему. Проведення ідентифікаційних і класифікаційних лінгвістичних експертиз має бути комплексним

Взагалі *комплексною* експертизою називається дослідження, при проведення якого поставлена проблема вирішується за допомогою спеціальних знань з декількох суміжних наук і стосується до компетенції двох чи більше видів експертизи. Саме таким типом може стати лінгвістична експертиза, практика якої об'єднує досягнення загального мовознавства, когнітивної і психолінгвістики, дискурсології, психології і ряду інших дисциплін, – експертиза, яка аналізує всі сторони мовленнєвої поведінки людини: виявлення особливостей не лише фонетичного, але й семантичного, когнітивного, синтаксичного ритму, фокусування свідомості на певних категоріях, що віддзеркалюється в змінах тону, інтенсивності, темпу, тембру, синтагматичному членуванні, наявності хезитацій, стилістичних і комунікативних особливостях акта спілкування.

На нашу думку, порівняльний лінгвістичний аналіз має спиратися на дві схеми: *антропоцентричну* і *текстоцентричну*.

Системним центром антропоцентричної схеми порівняння є мовна особистість. В цьому випадку ми повинні уявити мовця як біосоціокогнітивну систему, тобто систему, що породжує мовлення і має біологічні, соціальні і когнітивні параметри.

Біологічні параметри зазвичай обумовлюють тембр голосу, наявність або відсутність дефектів вимови (артикуляції і інтонування), темп мовлення, ритм мовлення, дистрибуцію фізіологічних пауз, нейролінгвістичну (право- або лівопівкулеву) акцентуацію, психіатричну акцентуацію дискурсу [див.: 5; 10], зміни мовлення під впливом неприродного положення тіла, емоційного напруження, захворювання тощо.

Якщо тембр голосу, наявність дефектів практично незмінні за умов змінення психофізіологічних параметрів комунікації, то темп, ритм варіюються в залежності від ситуації і може слугувати підставою для визначення психіатричної акцентуації, психічної лабільності.

Соціальні параметри визначають володіння орфоепічними, просодичними, лексичними, словотвірними, морфологічними та синтаксичними нормами, словниковий запас, вміння вибрати відповідний соціальний регістр та стиль спілкування.

Когнітивні параметри виявляють себе в зміні тем, чергуванні інформаційно навантажених і формальних (хезитаційних, контактних, потокотворчих) елементів, специфіці вибору дискурсних слів, парадигматичного вибору квазисинонимів, особливості фокусування свідомості на якій-небудь онтологічній категорії (речі, властивості, відношення, час тощо), у глибині фраз, уподобанні вибору активних або пасивних конструкцій, дрібності синтагматичного членування, у своєрідній зміні інтонації в передтакті та шкалі фрази тощо.

Текстоцентрична схема експертизи передбачає використання традиційного зразку аналізу, який спирається на рівневу модель мови та її реалізації в дискурсі. При цьому ми розглядаємо сегментні та суперсегментні фонетичні особливості, особливості лексичної організації, слововживання, фразової та надфразової граматики, діалогічну структуру (у випадку діалогічного мовлення).

Антропоцентричний аналіз вбачається більш перспективним з погляду ідентифікації та діагностики особистості, однак з метою підвищення достовірності висновків доцільно використовувати обидві схеми як такі, що доповнюють одна одну. Якщо більшість параметрів з аналізованих фонограм збігається можна твердити про приналежність їх однієї

особі [докладніше див. 1–3].

Практика проведення експертиз за означеною схемою підтвердила її ефективність. Проблема виникає в її трудомісткості, що обумовлює продовження встановленого законом терміну проведення експертиз. Тому головна подальша мета – створення комп'ютерної програми аналізу тексту на запропонованих теоретичних засадах.

Література

1. Бардина Н. В. Системный антропоцентрический подход к проведению лингвистических экспертиз // Теорія та практика судової експертизи і криміналістики. – Вип..2. Збірник матеріалів міжнародної науково-практичної конференції. Харків, 19–20 червня 2002 р. – Харків: Право, 2002. – С.217 – 221.
2. Бардина Н. В. Проблема системности дискурсивного анализа в лингвистических идентификационных экспертизах устной речи // «Роль и значение деятельности профессора р.С. Белкина в становлении и развитии современной криминалистики»: Материалы международной научной конференции. – М., 2002. – С.411 – 416.
3. Бардина Н. В. Этнокультурный фактор в фоноскопических лингвистических экспертизах // Мова. – Вип.8. – Одеса:Астропринт, 2003. – С.20 – 27.
4. Бардіна Н. В. Лінгвістична фоноскопія як новий напрямок прикладної лінгвістики // Українське мовознавство. Міжвідомчий науковий збірник. Вип. 32–33. – К., 2004. – С.22–26.
5. Белянин В. П. Основы психолингвистической диагностики. – М.: Тривола, 2000. – 248 с.
6. Галяшина Е. И. Основы судебного речеведения. – М.:Станси, 2003. – 232 с.
7. Дьяченко О. Ф. Системи ідентифікації особи за ознаками усного мовлення // Теорія та практика судової експертизи і криміналістики: Зб. науково-практ матеріалів. – Х., 2001.
8. Златоустова Л. В. Универсальное, групповое и индивидуальное в речи (лингвокриминалистический аспект) // Русский язык: исторические судьбы и современность: Международный конгресс исследователей русского языка: Труды и материалы. – М., 2001. – С.405 – 406.
9. Китайгородская М. В., Розанова Н. Н. Языковая личность в аспекте проблем судебной экспертизы устной речи // Язык и личность. – М.: Наука, 1989. – С.130 – 143
10. Леонгард К. Акцентуированные личности. – Р/на-Дону: Фенікс, 2000. – 544 с.
11. Наказ Міністерства юстиції України від 17.01.02 р. №4/5 "Про внесення змін до наказу Міністра юстиції України від 15.07.97 р. №285/7– А "Про експертно-кваліфікаційні комісії та атестацію судових експертів//Document HTML. – http://www.lawukraine.com.ua
12. Потапова р.К. Речь: коммуникация, информатика, кибернетика. – М.: Эдиториал УРСС, 2001. – 568 с.
13. Справочная книга криминалиста. – М.:Норма, 2000. – 712 с.
14. Судово-експертна діяльність. – К.:ІнЮре, 2002. – 343 с.
15. Теория и практика судебной экспертизы. – СПб.: Питер, 2003. – 704 с.
16. Шляхов А. Р. Судебная экспертиза, организация и проведение. – М.: Юридическая литература, 1979. – 165 с.

Н. Глізнуца*

Одеський національний університет ім. І. І. Мєчнікова(Одеса)

УДК 81'322.5

ВИМОВА ДОВГИХ ТА КОРОТКИХ ПРИГОЛОСНИХ ЯК ІДЕНТИФІКАЦІЙНА ОЗНАКА

У статті розглянуто питання вимови довгих та коротких приголосних сучасної української мови як ідентифікаційна ознака мовця.

Ключові слова: ідентифікація, довгі та короткі приголосні, сучасна українська мова, антропоцентризм, фоноскопія.

Сучасне мовознавство, орієнтоване на антропоцентричну парадигму, все більше уваги концентрує, по-перше, на зв'язок мовних і психологічних особливостей мовця і, по-друге, на прикладному аспекті характерологічних ознаках усного мовлення.

Таким чином, поєднання класичних теоретичних досліджень з новітніми експериментами є **актуальним.**

Проблема вимови довгих та коротких приголосних в українському мовленні багато років була проблемою суто фонологічною (М. А. Жовтобрюх, Ф. Т. Жилко, І. І. Ковалик, П. П. Коструба, Г. М. Міжевська, А. А. Москаленко, М. Ф. Наконечний, В. С. Перебийніс, Л. І. Прокопова, І. П. Сунцова, Г. Ф. Шило), але з початком XXI століття, вона набула

* © *Н. Глізнуца, 2006*

характеру міждисциплінарного – такого, що пов'язує лінгвістичні дослідження з прикладними.

В останні роки не припиняються пошуки фонетичних ознак, які б допомагали визначити особистість за критеріями, вимови окремих звуків. Дослідники зосередили увагу на комп'ютерних технологіях аналізу мовленнєвого сигналу, пошуках унікального співвіднесення частот у вимові, перш за все голосних [4]. Треба відзначити, що в цьому напрямку зроблено багато цікавого [1].

Але ж, треба і не забувати про вагомість приголосних, перш за все, це стосується більшості слов'янських мов. І. О. Бодуне де Куртене [5] сформулював закон, який сьогодні науковці називають „законом Бодуена де Куртене" про те, що спрощується система голосних та ускладнюється система приголосних. А спрощення системи голосних означає, що голосні відіграють все меншу роль для розпізнання слів. Наприклад: *х-р-ш-, ф-л-л-г* – дуже легко зрозуміти ці слова. Як видко, приголосні відіграють найважливішу роль у розумінні значення слова. Це диктується тим, що приголосні зустрічаються рідше, ніж голосні.

Тому розрізнювальна сила у голосних (тобто інформація, яку вони несуть) в середньому нижча ніж у приголосних. Приголосні завжди більше несуть інформації у мові. Голосні несуть частоту основного тону, яка сприймається правою півкулею. Частота основного тону, тобто інтонація змінюється залежно від ситуації, то приголосні навпаки не змінюються і несуть основну інформацію, яка обробляється лівою півкулею мозку і відповідає за інтелектуальне мовлення.

Сьогодні бракує експериментального аналізу вимови деяких звуків у сучасній українській мові. Ця **проблема,** перш за все стосується теоретичного обґрунтування фоноскопії.

„Предмет лінгвістичної фоноскопії можна визначити як дослідження віддзеркалення мовної особистості в усному мовленні й розробку методів ідентифікування, кваліфікування та діагностування людини на підставі її фонації (звукової об'єктивації)" [2].

На російському мовному матеріалі зроблено багато [1; 3; 4] Проте на українському ґрунті нема жодного дослідження, яке б теоретично перевіряло особливості сприйняття довгих українських приголосних як ідентифікаційну ознаку особистості. Тому дослідження можна вважати **новим**.

Метою роботи стало визначення **ваги** вимови довгих та коротких приголосних для ідентифікації мовця. Було висунуто гіпотезу, що на вимову довгих приголосних впливає тип мовлення: читання (де на фонацію впливає графічний образ) та переказ (коли мовець інколи запобігає збігу довгих приголосних, щоб підвищити ефект порозуміння).

Для експерименту було створено штучний текст, у словах якого були наявні довгі та короткі приголосні типу: *ссе – се, Ганна – гана, до Ганни – догана, ллють – лють, у судді – у суді, читання – чи Таня, у почутті – у почуті, безмежжі – без межі, ввись – вись* тощо. До експерименту було залучено чотирьох носіїв української мови та одного носія російської мови (три жінки та однин чоловік носії української мови, та один чоловік – носій російської мови, які були зашифровані як УЖ$_1$, УЖ$_2$, УЖ$_3$, УЧ$_1$, РЧ$_2$).

На першому етапі було запропоновано прочитати текст, потім його переказати. Читання тексту та переказ було записано за допомогою комп'ютерної програми Cool Edit, та була виміряна тривалість довгих та коротких приголосних під час читання та переказу. Таким чином матеріалом дослідження стали фонозаписи.

Слуховий аналіз показав, що довгі та короткі приголосні виділяються, як під час читання так і під час переказу. Кожен диктор виділяв довгі приголосні для розуміння контексту, тому що саме довгота впливала на розуміння тексту. Ці слова можна назвати омофонами, які розрізняються саме з фонологічної позиції.

Результати комп'ютерного аналізу вказують на те, що довгі приголосні (подовжені) вимовляються майже в два рази довше, ніж короткі, вимовляються єдиною артикуляцією, в якійсь мірі відбувається спрощення, тому що не помітили і не виявили будь-яких ознак, які б указували на спад інтенсивності голосового тону в середині приголосних.

Також було виявлено, що у одних дикторів наприклад УЖ$_1$ та УЧ$_1$, УЖ$_3$ довгі та короткі приголосні читалися довше, а під час переказу тривалість довгих та коротких приголосних зменшувалась, а у дикторів УЖ$_2$, РЧ$_2$ спостерігалось навпаки під час читання довгі та короткі приголосні вимовлялися коротше, а під час переказу тривалість довгих та коротких приголосних збільшувалась.

Аудиторський аналіз показав, що в УЖ$_1$ під час читання тривалість довгих та коротких приголосних збільшена, а під час переказу тривалість зменшена. Лише у слові *догана* тривалість [н] була однакова як під час читання так і під час переказу, та у слові *до Ганни* під час переказу була збільшена тривалість [н:] в порівнянні з читанням.

В УЖ$_2$ під час читання тривалість довгих та коротких приголосних була зменшена, а під час переказу тривалість була збільшена. Лише у слові *у суді* тривалість [д] під час переказу була зменшена, та у слові *догани* тривалість [н] зменшена, та у слові *се* тривалість [с] також була зменшена в порівнянні з читанням.

В УЖ$_3$ під час читання тривалість довгих та коротких приголосних була збільшена, під час переказу тривалість була зменшеною. Лише у слові *без межі* тривалість [ж] підчас переказу була зменшеною та у слові *доганц* тривалість [н] була зменшеною, а у слові *ссе* тривалість [с] була збільшеною в порівнянні з читанням.

Чоловіки були не стовідсоткові розпізнані, тому що за тембром були схожі та за хезитацією. Розрізнення заключалось тільки в тім, що РЧ$_2$ читав і вимовляв *льють* замість *ллють* і під час переказу замінив слово *на піддашші – на дахах домів.*

В УЧ$_1$ (вища освіта) під час читання тривалість довгих та коротких приголосних була збільшеною, а під час переказу тривалість була зменшеною. Лише у слові *ссе* тривалість [с] була збільшеною в порівнянні з читанням.

В РЧ$_2$ (вища освіта) під час читання тривалість довгих та коротких приголосних була зменшеною, а під час переказу була тривалість збільшеною. Лише у слові *у суді* тривалість [д] була зменшеною, та у слові *гана* тривалість[н] була також зменшеною в порівнянні з читанням.

Гіпотетично, на початку експерименту, вважалось, що у росіянина тривалість довгих буде збільшеною ніж у українця під час читання, а в ході експерименту вийшло навпаки. Росіянин читав швидше, переказував повільніше. Українець навпаки читав повільніше, переказував швидше.

Таким чином говоріння залежить від того наскільки людина упевнена, в тому, що вона говоре. Впевненість – темп переказу збільшується, невпевненість – темп зменшується.

Написання, графіка впливає на читача, тобто додає впевненості. Читач може читати не задумуючись над змістом, чи задумуючись.

На другому етапі експерименту було запропоновано двадцятьом інформантам прослухати записаний матеріал, і спробувати ідентифікувати дикторів. Спочатку інформанти прослухали текст, потім переказ. Для цього було складено дві анкети за якими б інформанти відзначали та помічали свої спостереження. В анкетах були поставлені такі запитання: 1) Прослухайте текст, який читали п’ять дикторів, потім переказ цього тексту, і визначте дикторів. Кому з дикторів за голосом відповідає переказ. 2) За якими ознаками ви визначили: тембр голосу, особливості вимови певних звуків. 3) Якщо так, то яких саме: голосних чи приголосних. 4) Якщо приголосних, то яких? 5) Чи помітили ви особливість вимови наступних слів: *читання, чи Таня, почуті, почутті, гаддя, Ганна, гана, беззахисна, у суді, піддашші, се, лють, співчуття, ссе, слухання, у судді, безмежжі, від дому, ллють, цвіллю, догани, оббитому, вважала, до Ганни.* 6) Якщо так, то у яких дикторів?

Слухове розпізнання показало, що з двадцяти аудиторів правильно визначили дикторів 11чоловік, тобто 55%. Помилка була у визначенні чоловіків, жінок визначили вірно. УЖ$_1$ – 100%, УЖ$_2$ – 100%, УЖ$_3$ – 100%, УЧ$_1$ – 55%, РЧ$_2$ – 45%. Всі визначали за тембром голосу, за швидкістю говоріння, за наявності хезитації, та за особливостями вимови приголосних звуків у тому числі, та за особливостями вимови слів з короткими та довгими приголосними.

35% звернули увагу на [д:]; 30% звернули увагу на [ж:]; 25% звернули увагу на [в]; 25% звернули увагу на [л:]; 20% звернули увагу на вимову довгого [с:]; 20% звернули увагу на [т:]; 20% звернули увагу на [ш:]; 20% звернули увагу на [з:]; 20% звернули увагу на [н:]; 20% звернули увагу на [с]; 15% звернули увагу на [в:]; 10% звернули увагу на [т]; 10% звернули увагу на [л];

90% звернули увагу на вимову наступних слів як [*ссе, гаддя, безмежжі*] в усіх дикторів.

УЖ$_1$ – 45% звернули увагу на вимову слова [*почутті*], 30% на [*цвіллю*], 25% на [*чи Таня*], 25% на [*вважала*], 20% на [*догани*], 20% на [*почуті*], 10 % на [*читання*], 10% на [*піддашші*], 10% на [*гана*];

УЖ$_2$ – 20% звернули увагу на вимову слова [*до Ганни*], 15% на [*чи Таня*], 10% на [*гана*], 10% на [*у суді*], 10% на [*вважала*], 10% на [*піддашші*]; 5% на [*від дому*];

УЖ$_3$ – 25% звернули увагу на вимову слова [*піддашші*], 20% на [*слухання*], 20% на [*співчуття*], 20% [*догани*]; 10% на [*гана*], 10% на [*у судді*];

УЧ$_1$ – 25% звернув увагу на вимову слова [*ллють*], 25% на [*до Ганни*], 25% на [*читання*], 20% на [*у судді*], 10% на [*лють*], 10% [*безмежжі*];

РЧ$_2$ – 50% звернули увагу на вимову слова [*ллють*, тому, що вимовлялось *льють*], 10% на [*у судді*], 10% на [*піддашші*], 10% на [*почутті*], 10% на [*безмежжі*], 10% на [*Ганна*];

З експерименту видно, що форми слів *ссе, гаддя, безмежжі* у всіх дикторів розпізнавалися. Також була помічена вимова таких слів, як *почутті, почутті, цвіллю, чи Таня, вважала, піддашші, читання, у судді, у суді, до Ганни, догани, гана, Ганна, ллють, лють, слухання, співчуття. Н*айменше привернуло уваги слово *від дому. З*овсім уваги не привернули такі слова як *беззахисна, оббитий* (тобто подвоєні приголосні не привертали уваги).

З експерименту ми побачили, що не носію мови дуже важко переказувати текст. Це пов'язано насамперед з тим, що більшість слів диктор не розумів. Саме переказ показав чи розуміє переповідач значення слів та взагалі зміст тексту. Це було помічено на прикладах, які були підібрані. Той, хто розумів значення, переказував текст, вимовляючи довгі та короткі приголосні і навіть було помічено, УЧ$_1$ давав пояснення [*У судді було багато запитань, тобто суддя ставив багато запитань*].

Довга вимова приголосних виділялась і давала більшу можливість пізнання голосу, ніж короткі. Кожен мовець звуки в тому самому слові промовляє по своєму, залежно від особливостей голосу, індивідуального тембру, манери говоріння тощо, про те фонеми усі сприймають однаково. Довгі фонеми сприймаються довго, короткі – коротко.

В нашому випадку розуміння тексту несли на собі **довгі** приголосні. Якщо науковці досліджували і досліджують лише голосні і говорять про їх вагомість, то наш експеримент вказує на **вагомість** приголосних, і не просто приголосних, а довгих та коротких приголосних, які несуть головну інформацію.

Як показав експеримент, особливості вимови довгих та коротких приголосних відіграють важливу роль у процесі ідентифікації мовця, та можуть стати одним із критеріїв фоноскопічного аналізу мовлення.

Література

1. Баранов А. Н. Введение в прикладную лингвистику. – М., 2000. – С. 8.
2. Бардіна Н. В. : Лінгвістична фоноскопія як новий напрямок прикладної лінгвістики (у друці).
3. Галяшина Е. И. Прикладные основы судебной фоноскопической экспертизы. //Теория и практика судебной экспертизы: Сборник . Питер, 2003. – С. 9–134.
4. Работягов А. В. Некоторые методические аспекты проведения судебной фоноскопической экспертизы. // Теорія і практика судової експертизи і криміналістика. Зб., науково-криміналістичних матеріалів. Вип. 4. Харьків „Право" 2004. – С. 219–227.
5. Русский язык и советское общество: Социол.-лингв. исследование / АНССР. Ин-т русс. яз.; Под. ред. М. В. Панова. – М.: Наука, 1968. – 213с.

В. Близнюк*

Київський національний лінгвістичний університет (Київ)

УДК 81'322

МОДЕЛЮВАННЯ ФОНЕМНОЇ СТРУКТУРИ ЛЕКСЕМ СУЧАСНОЇ ІТАЛІЙСЬКОЇ МОВИ

The analysis of phonemic structures of lexemes in a modern Italian literary language is carried out. The structure of a word is regard as a canonical form. Inductive and deductive linguistic model is a real-life and theoretically possible canonical forms are compared.

Модель (франц. *modèle*, італ. *modello*, від лат. *modulus* – міра, мірило, зразок, норма) в широкому значенні є умовним або мисленнєвим образом або прообразом якого-небудь об'єкта чи системи об'єктів («оригінала» даної моделі), що використовуються за даних умов в якості їхнього «заступника» або «представника» [Булыгина, Крылов, 1990: 304].

* © *В. Близнюк, 2006*

Модель є лише формою вираження, яка може мати свою повну наукову значущість лише з урахуванням того, про що йде мова [3: 15].

Лінгвістичні моделі поділяють на індуктивні (аналітичні) й дедуктивні (синтетичні, породжувальні). Індуктивні моделі виходять з мовного матеріалу і шляхом його аналізу встановлюють класи й підкласи одиниць та їхню типологію. Дедуктивні моделі виходять з абстрактних класів, що є вихідними для моделі, і шляхом заданих операцій синтезу повинні породити, тобто обчислити, одиниці мови й визначити їхню ієрархію [4: 339].

Найпростішою моделлю фонемної структури лексем будь-якої мови є канонічна форма (КФ). Під каноном (від давньогрецького *k¼n3nion* – норма, правило, критерій) розуміється зведення положень, які мають догматичний характер, тобто є твердо встановленими й визнаними зразковими [6: 544]. Отже, канонічна форма є формулою, яка представляє слово в найбільш узагальненому вигляді – як ланцюжок послідовностей голосних (V) та приголосних (C) фонем [8: 284-286].

Предметом нашого дослідження є встановлення канонічних форм сучасної італійської літературної мови. Аналіз проведено на матеріалі механічної вибірки зі словника італійської мови «Dizionario Italiano Sabatini Coletti». – Firenze: Giunti Gruppo Editoriale, 1997, яка складає 20000 слів загальною кількістю 193688 фонем (83592 голосних і 110096 приголосних).

Завдання, що стоять перед нами, такі: аналіз розподілу довжини КФ; з'ясування моделюючої сили КФ певної довжини, встановлення КФ, які описують ядро, основну та периферійну систему лексичного складу сучасної італійської літературної мови; аналіз співвідношення теоретично можливих і реалізованих КФ, тобто зіставлення індуктивних і дедуктивних моделей КФ, і з'ясування причин, з яких можлива або неможлива реалізація тих або інших КФ.

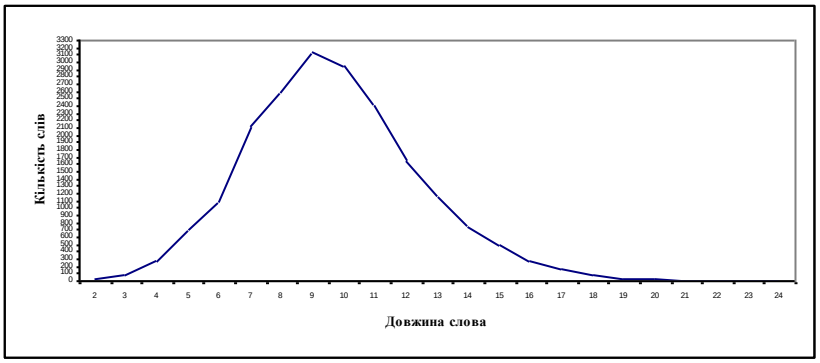
Перед тим, як починати дослідження правил побудови слів із фонем, необхідно встановити, якої довжини слова допустимі в італійській мові. Аналіз розподілу довжини слів у вибірці показує, що найкоротшою є словоформа в 2 фонеми (в італійській мові є й словоформи довжиною в одну фонему, але їх дуже мало і вони не потрапили в нашу вибірку), а найдовшою є словоформа в 24 фонеми. До цього слід додати, що в словниках італійської мови зареєстровано найдовше слово в 26 фонем – /preʃipitevolissimevol'mente/ [9: 65]. Найчастішими є слова довжиною дев'ять фонем. За ними слідують зі зменшенням частоти слова довжиною десять, вісім, одинадцять, сім фонем. Як випливає зі схеми 1, зростання кількості слів довжиною від двох до дев'яти фонем відбувається поступово, а зменшення кількості слів довжиною від десяти до двадцяти чотирьох фонем – більш інтенсивно. Отже, коливання середньої довжини слова в італійській мові становить від 9,64 до 9,72 ($\bar{x} \pm 2\sigma_{\bar{x}} = 9,64 \approx 9,72$).

Встановлений на матеріалі механічної вибірки інвентар КФ нараховує 1276 одиниць. Кожна КФ описує різну кількість слів. Дослідження моделюючої сили КФ показує, що в словах певної довжини виділяється лише одна КФ, яка має найвищу моделюючу силу. КФ з найвищою моделюючою силою описують 4816 слів або 24,08 % вибірки. Серед виявлених КФ форм можна виділити КФ з високою (від 40,85 % до 74,36 %), середньою (від 12,13 % до 43,48 %) та низькою (від 4,97 % до 9,73 %) моделюючою силою. 4 КФ (довжиною від 21 до 24 фонем) зустрічаються спорадично, тому для них не можна зробити статистично вірогідні висновки.

У словах довжиною від 2 до 7 фонем виділяється 1-3 моделі КФ, які описують до 75 % слів заданої довжини, або ядро системи лексики, та невелика кількість моделей, за якими будується до 90 % слів заданої довжини, або основна система лексики, а інші ж КФ, кількість яких є значною, належать до КФ з низькою моделюючою силою [5: 10]. Отже, для 2-фонемних слів сучасної італійської літературної мови 1 КФ (CV) описує 74,36 % слів такої довжини, для 3-фонемних 2 КФ (CVC, CVV) описують 67,61 %, для 4-фонемних 1 КФ (CVCV) – 68,23 %, для 5-фонемних 2 КФ (CVCCV, CCVCV) – 74,93 %, для 6-фонемних 2 КФ (CVCVCV, CCVCCV) – 65,15 %, для 7-фонемних 3 КФ (CVCCVCV, CVCVCCV, CCVCVCV) – 74,80 %.

Для слів довжиною від 8 до 14 фонем кількість КФ, яка необхідна для опису ядра і системи починає зростати, тому що зі збільшенням кількості фонем у слові зростає комбінаторна можливість сполучення фонем, і мова використовує більшу кількість КФ, які не суперечать її законам. Далі спостерігаємо зменшення кількості цих КФ, оскільки зменшується загальна кількість слів, що мають довжину від 14 до 24 фонем.

Схема 1. Розподіл довжини слів сучасної італійської літературної мови



Таблиця 1. КФ слів сучасної італійської літературної мови

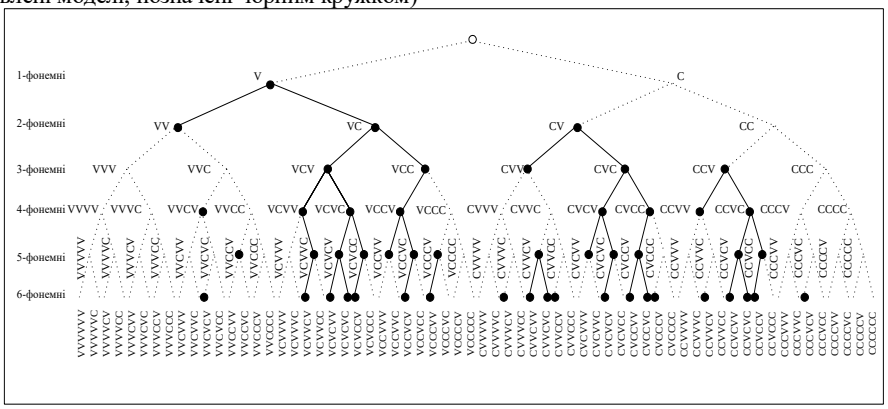
Довжина КФ	Кількість різн. КФ	Кількість слів даної довжини	КФ з найбільшою моделюючою силою			
			КФ	Кількість слів	Доля КФ в словах даної довжини, %	Моделююча сила
2	3	39	CV	29	74,36	висока
3	5	71	CVC	29	40,85	середня
4	8	277	CVCV	189	68,23	висока
5	13	706	CVCCV	445	63,03	
6	20	1102	CVCVCV	555	50,36	
7	32	2123	CVCCVCV	923	43,48	середня
8	47	2591	CVCVCVCV	672	25,94	
9	66	3132	CVCCVCVCV	803	25,64	
10	79	2953	CVCCVCCVCV	522	17,68	
11	115	2390	CVCCVCVCVCV	290	12,13	
12	122	1644	CVCCVCCVCVCV	160	9,73	низька
13	138	1146	CVCCVCCVCVCCV	84	7,33	
14	151	726	CVCCVCVCVCVCCV	41	5,65	
15	140	483	CVCCVCVCCVCVCCV	24	4,97	
16	120	286	CVCCVCVCCVCCCVCV	19	6,64	
17	87	169	CVCVCVCVCCVCCCVCV	15	8,88	
18	62	83	CVCCVCVCVCCVCCCVCV	5	6,02	
19	31	37	CVCVCVCVCVCCVCCCVCV	3	8,11	
20	22	27	CVCCVCVCVCVCCVCCCVCV	4	14,81	середня
21	6	6	CVCCVCVCVCVCCVCVCVCCV	1	–	спорадична
22	5	5	VCCVCCVCVCCVCCVCVCCVCV	1	–	
23	3	3	CVCVCCVCVCCVCVCCVCCCVCV	1	–	
24	1	1	CVCCVCCVCVCCVCVCCVCCCVCV	1	–	

Виділення КФ, що формують основну систему, включаючи ядро, можливе не лише в словах заданої довжини, але й цілком у вибірці. Залежно від моделюючої сили КФ створюється їхня ієрархія: 108 КФ моделюють 15157 слів (75,79 % всієї вибірки) і

становлять ядро основної системи. Основна система, включаючи ядро, охоплює 320 КФ, які моделюють 18174 слів (90,87 % вибірки). Останні 956 КФ моделюють 1826 слів (9,13 % вибірки). Такі КФ становлять периферійну підсистему.

Виходячи з результатів аналізу, можна порівняти кількість теоретично можливих і наявних моделей КФ слів заданої довжини в сучасній італійській літературній мові. Теоретично можливі різновиди фонемної структури слова можна легко встановити, побудувавши відповідне дерево, в якому кожне ліве відгалуження означає дописування голосної, а праве – приголосної до вже наявного ланцюжка фонем (фрагмент такого дерева наведено на схемі 2). Кількість таких структур для слів певної довжини дорівнює 2^n, де 2 – два класи фонем (V і C), а *n* – число фонем у слові. Отже, для однофонемних словоформ кількість різних КФ дорівнює $2^1 = 2$, для двофонемних – $2^2 = 4$, для трифонемних – $2^3 = 8$ і т. д. [4: 256].

Схема 2 Теоретично можливі та реалізовані моделі КФ слів сучасної італійської літературної мови довжиною від 2 до 6 фонем (примітка: суцільними лініями з'єднуються виявлені моделі, позначені чорним кружком)



Аналіз схеми 2 показує, що деякі КФ не можуть бути реалізовані внаслідок обмежень, які накладають закони мови: в італійській мові нема словоформ зі структурою С, СС, ССС, СССС і т. п. (за винятком декількох вигуків, що складаються з двох або трьох приголосних поспіль: /mm/, /ps/, /st/, /brr/, /prm/, /pst/, які зареєстровані у словнику, але не потрапили до нашої вибірки). Не зустрічаються також сполучення чотирьох приголосних на початку слова [7: 66]. За результатами наших досліджень фонемної структури лексики сучасної італійської літературної мови збіг трьох голосних можливий лише в середині слова (за винятків двох вигуків /e'ɔo/, /uo'ɛ/, які не потрапили до вибірки), тому не зустрічаються такі КФ, як VVV, VVVC, CVVV, а також КФ, в яких ми спостерігаємо чотири і більше голосних підряд: VVVV, VVVVC, CVVVV, VVVVV і т. п. Слід зауважити, що є такі КФ, які не суперечать законам мови, але не використовуються нею.

На підставі більш детального аналізу схеми 2 маємо такі результати.

Із 8 теоретично можливих КФ довжиною 3 фонеми використовується 5, тобто 62,50%. Серед невикористаних – VVV, VVC, ССС. В чотирифонемних словах використовується 8 моделей КФ із 16 або 50,00 %. Не використовуються такі КФ: VVVV, VVVC, VVCC, VCCC, CVVV, CVVC, CCCV, CCCC. Сполучення трьох голосних на початку чи в кінці слова або двох приголосних в кінці слова суперечать законам італійської мови. У п'ятифонемних словах використовується лише 13 моделей КФ із 32 теоретично можливих, тобто менше половини (40,63 %). Зростання частки нереалізованих КФ також пояснюється збільшенням кількості теоретично можливих моделей зі сполученнями трьох і більше голосних або приголосних на початку чи в кінці слова.

Частка використаних моделей КФ у словах довжиною шість фонем ще менша, в цьому випадку з 64 теоретично можливих знаходимо лише 20 КФ, тобто 31,25%.

Переважна більшість невикористаних КФ суперечить законам мови, тому що вони містять у собі сполучення трьох, чотирьох, п'яти і навіть шести голосних підряд, а також чотирьох, п'яти й шести приголосних.

Таблиця 2. Теоретично можливі та реалізовані моделі КФ слів сучасної італійської літературної мови

Довжина КФ	Кількість теоретично можливих КФ	Кількість реалізованих КФ	Процент реалізації КФ, %
2	4	3	75,00
3	8	5	62,50
4	16	8	50,00
5	32	13	40,63
6	64	20	31,25
7	128	32	25,00
8	256	47	18,36
9	512	66	12,89
10	1024	79	7,71
11	2048	115	5,62
12	4096	122	2,98
13	8192	138	1,68
14	16384	151	0,92
15	32768	140	0,43
16	65536	120	0,18
17	131072	87	0,07
18	262144	62	0,02
19	524288	31	0,01
20	1048576	22	0,002
21	2097152	6	0,0003
22	4194304	5	0,0001
23	8388608	3	0,00004
24	16777216	1	0,00001
Разом:	33554428	1276	0,004

Як випливає з таблиці 2, кількість різновидів теоретично можливих КФ форм дуже швидко росте зі збільшенням довжини слова. У той же час спостерігається зменшення процента реалізації цих форм у конкретних словах. Така картина пояснюється тим, що зі збільшенням довжини слова зменшується кількість довгих слів внаслідок обмежень сполучуваності двох класів (V і C) фонем. Отже, загальна кількість реалізованих КФ слів сучасної італійської літературної мови становить лише 0,004 % від теоретично можливих.

Висновки:

1. Опис лексем в термінах голосних (V) і приголосних (C) фонем дає можливість виявити КФ з найбільшою моделюючою силою, які описують значну кількість слів сучасної італійської літературної мови.
2. Невелика кількість КФ необхідна для опису ядра та основної системи слів заданої довжини, а також усієї лексики сучасної італійської літературної мови. Це є проявою закону переваги, за яким використовується лише невелика кількість одиниць мови, що формують мінімум, необхідний для її розуміння та запам'ятовування.
3. В сучасній італійській літературній мові, як і в усіх інших, використовується незначна кількість КФ з теоретично можливих. Це свідчить про великий потенціал невикористаних можливостей мови.

Література

1. Ахманова О. С. Словарь лингвистических терминов. – М., 1966. – 606 с.
2. Булыгина Т. В., Крылов С. А. Модель. // Лингвистический энциклопедический словарь. – М., 1990. – С. 304-305.
3. Лосев А. Ф. Введение в общую теорию языковых моделей. / Под ред. И. А. Василенко. Изд. 2-е, стереотипное. – М.: Едиториал УРСС, 2004. – 296 с.
4. Перебийніс В. С. Кількісні та якісні характеристики системи фонем сучасної української літературної мови: Дис. … д-ра філол. наук. – К., 1970. – 510 с.
5. Перебийніс В. С. Модель // Українська мова: Енциклопедія. – К.: Укр. енцикл., 2000. – С. 339-340.

6. Слипченко Л. Д. Фонемная структура лексики английского языка: Автореф. дис. … канд. филол. наук: 10.02.21. – К., 1974 – 21 с.
7. СЭС. Советский энциклопедический словарь. – М.: Сов. энциклопедия, 1982. – 1600 с.: ил.
8. Челышева И. И., Черданцева Т. З. Итальянский язык // Языки мира: Романские языки. – М.: Academia, 2001. – С. 56-90.
9. Hockett Ch. F. A course in modern linguistics. – New York: The Macmillan Company, 1958. – 621 p.
10. Nespor M. Fonologia. – Bologna: il Mulino, 1994. – 348 p.
11. Sabatini F., Coletti V. Dizionario Italiano Sabatini Coletti. – Firenze: Giunti Gruppo Editoriale, 1997. – 3037 p. con illustrazioni

З. Дудник*
УДК 81’322.6

СЕГМЕНТАЦІЯ НА АКУСТИЧНІ СКЛАДИ ДЛЯ КОМПІЛЯТИВНОГО СИНТЕЗУ АНГЛІЙСЬКОГО МОВЛЕННЯ

This paper describes the approach to syllabification currently used in the Laboratory of Computational Linguistics (Kyiv National Linguistic University) for English text-to-speech synthesis. Syllable is a subcomponent of the word model whereas quasi-syllable is a subcomponent of the hierarchic speech model of phonetic word, rhythmical group, and syntagma. The real structure and syntactical position of quasi-syllables enables the TTS-system to retain the naturalness of concatenative speech signals.

1. Відомо, що в теорії складу залишається багато нез’ясованих питань, хоча над ними працюють тисячі дослідників. Тісно пов’язаними між собою є проблема складоподілу і процедура сегментування на склади з прикладною метою. Синтезатори мовлення «текст на озвучення», які підтримують конкатенацію складів, при доборі елементної бази спираються здебільшого на морфемно-фонологічні принципи складоподілу [1]. Процедуру препроцесорного фонетичного аналізу в цьому випадку прирівнюють до процедури автоматичного складоподілу графічного тексту, тобто джерелом правил силабіфікації використовують матеріали, що фіксують результати мовного розвитку і функціонування. Натомість вимогою, яка висувається до конкатенованого мовлення на виході, є природність звучання, досягти якої можна лише з урахуванням реальних властивостей породження мовлення. Несумісність теоретичних підходів і практичних вимог вкотре ставить питання онтологічного підґрунтя теорії складу.

2. Виконання ручної сегментації акустичного сигналу мовлення на відрізки складової розмірності з метою створення акустичної бази складів в системі синтезатора англійського мовлення «текст на озвучення» привернуло увагу до можливостей розв’язання теоретичних проблем загальної фонетики і практичних проблем технології мовлення. Якщо за вихідну точку відліку обрати онтологічний аспект породження, то сегментація на склади має бути процедурою, що орієнтована на сутнісні властивості матерії мовлення, а саме на її психофізіологічно і артикуляторно обумовлену дискретність, а також на значну варіативність якості мовленнєвих подій. У такому разі широкі лінгвістичні узагальнення про склад з позиції виділення в ньому суттєвого, а також залежні від цієї позиції правила складоподілу мають братися до уваги з певною обережністю, адже їх онтологічна орієнтація – це результати мовного функціонування і нормативні вимоги, які не в змозі охопити реальні стратегії мовленнєвої поведінки.

Однак і для дослідника, який має справу з акустичним сигналом, вимовне зусилля розкривається через підручний матеріал, тобто є інструментально фіксованим результатом події. Тут так само виникає небезпека в тому, що особливості репрезентації мовлення за допомогою певного технічного прийому будуть ототожнені із особливостями самого мовлення. Приміром, структуру акустичної події репрезентує часова розгортка сигналу (осцилограма, спектрограма), що відображує зміни тиску повітря та зміни передавальної функції мовленнєвого тракту, яку фіксує мікрофон із власними технічними характеристиками. Тому в акустичній картині немає повної відповідності артикуляційним зусиллям. Наприклад, коли язик сильно притискається до піднебіння і відштовхується від нього для наступного артикуляторного зусилля, акустичний сигнал як такий може бути відсутній, але це не означає, що немає артикуляційної події. Іншими словами, подія

* *© З. Дудник, 2006*

реальна може бути розтлумачена через властивості відображення цієї події фізичним приладом у певний спосіб. Акустичні ознаки складового контрасту, пропоновані Л. В. Бондарко для обґрунтування складової структури, мають бути узгоджені з артикуляційною активністю органів і механізмами породження, інакше залишається небезпека ототожнення підручної репрезентації сигналу із самими мовленнєвими властивостями.

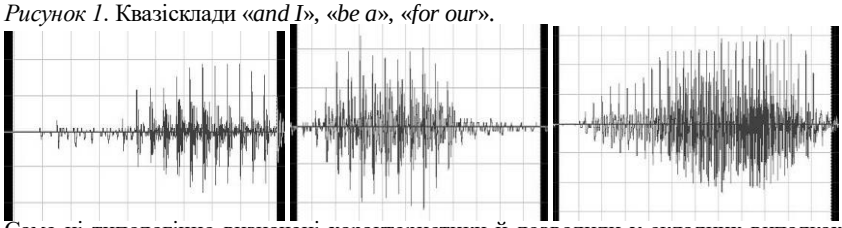
Таким чином, у теорії складу маємо Скіллу (ілюзія першовартості функціонального підходу) і Харибду (ілюзія самодостатності підручного матеріалу), між якими необхідно прокласти шлях і отримати позитивний результат, зокрема для здійснення обґрунтуваної процедури сегментації.

Якщо при аналізі починати відлік не з традиційно прийнятих лінгвістичних категорій та їх ознак, а з явищ і параметрів аспекту породження, то необхідно віддати належне індивідуальному рівню, а саме **психо-фізіологічним умовам звукотворення** і **варіативності** як **сутнісним властивостям живого мовлення. Т**оді через ці сутнісні характеристики можна виходити на функціональний рівень існування складу як мовної категорії, для якої суттєвою є реалізація фонетичної типології (складометричного чи акцентноритмічного типу мовлення). Таким чином, у цілокупне розуміння **складу як актуального вимовного зусилля (сутнісне)** й одночасно як **вихованої артикуляторно-акустичної звички (суттєве)** включені не лише типологічний, а й індивідуальний виміри, де **типологічний визначає реалізацію структури складу**, а **індивідуальний – його конститутивну роль**, і разом вони залежать від виконавського наміру мовця. Поєднання суттєвого й сутнісного у визначенні складу забезпечує адекватний підхід до проблеми реального складоподілу і принципів сегментації.

3. Процедуру сегментації можна звільнити від двох когнітивних ілюзій, нашарованих на явище дискретизації мовленнєвого потоку (широкого лінгвістичного узагальнення явища в категорії ***склад*** та технічно обумовлених властивостей сигналу), якщо залишити в стороні принципово неможливе визначення меж складу (як категорії), потім відмовитися від технічно допустимого встановлення меж акустичних подій. Тоді вивільнену увагу можна зосередити на коректній фіксації змін в координації артикуляторних зусиль мовця.

За цим реальним орієнтиром фізичні ознаки, або параметри, які бралися до уваги при встановленні межі складових зусиль, були визначені за фонетичною типологією англійського мовлення, що належить до акцентноритмічного типу. Для мовлення цього типу притаманний динамічний (силовий) наголос, який посилює контраст між наголошеними й ненаголошеними складами не лише за інтенсивністю й чіткістю вимовляння, а й за зміною тону. Для мовлення акцентноритмічного типу також притаманне злиття кількох ненаголошених складів. В аналізованому матеріалі простежувалася тенденція ритмічного підпорядкування частотних односкладових артиклів, сполучників і прийменників наступному чи попередньому складу: ***and*** *I; be* ***a; for*** *our* (рис. 1).

Рисунок 1. Квазісклади «*and I*», «*be a*», «*for our*».



Саме ці типологічно визначені характеристики й дозволили у складних випадках не шукати обов'язкового поділу на склади в очікуваному нормою випадку, а орієнтуватися на цілісне артикуляційне зусилля. Виявилося, що межа фізіологічних зусиль синхронізується зі зміною просодичних характеристик. Дещо подібне простежується у мовленнєвій поведінці людини: макродискретність мовлення збігається зі зміною жестів руки. У процедурі сегментації бралися до уваги порушення стабільності тонального контуру (зміна висоти та/або інтенсивності, зміна руху тону та/або інтенсивності). З огляду на обмеженість акустичної репрезентації щодо кількості проаналізованих

артикуляційних подій (550 квазіскладів), шукати жорсткий акустичний еквівалент межі виявилося недоцільним, хоча з накопиченням і описом максимально можливої кількості випадків постане необхідність обґрунтувати правила автоматичної сегментації.

Серед завдань сегментації постало питання щодо залучення складів, які виникають на міжсловесних стиках і відповідають за фонетичну цілісність синтагми: *buil**d a** lo**t of*** (рис. 2); *thin**k it***; *brea**k in***; *tariff**s on***; ***if I***; *sor**t out***; *don'**t a**gree*; *ri**ch**, **e**xpanding.*

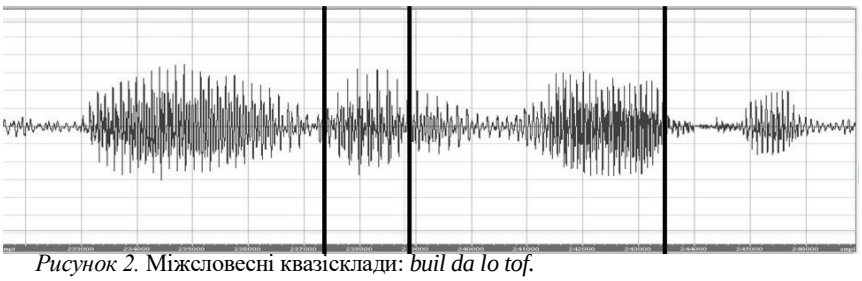


Рисунок 2. Міжсловесні квазісклади: *buil da lo tof.*

Просодичні ознаки також виявилися важливими критеріями при виділенні міжсловесних квазіскладів. На рис. 2 можна навіть помітити зміну частоти основного тону для кожного зусилля.

Важливу роль відіграло урахування при сегментації позиції складу в синтагмі чи акцентній групі. Склади для бази даних індексувалися як початкові, серединні й кінцеві відносно синтагматичної паузи. Початкові й кінцеві склади були почленовані з включенням невеликої паузи. Детальніше про укладання бази даних див. у статті Володкевич О. в цьому ж збірнику. Маючи природно оформлені характеристики інтонації в синтагмі при синтезі такі склади створюють задовільний інтонаційний малюнок звучання. Уже при синтезуванні цих елементів акустичної бази з'ясувалася також і перевага збереження й використання базою даних варіативних реалізацій складу, що додає природності синтезованому мовленню.

4. Склад на функціональному рівні виявляє свої властивості як субкомпонент морфо-фонемної моделі слова, склад на рівні породження (як квазісклад) є субкомпонентом в ієрархічній моделі фонетичного слова, ритмічної групи й синтагми.

Таким чином, складовий синтез англійського мовлення може спиратися на такі реальні характеристики складу, які мають місце як в типології англійського мовлення, так і в індивідуальному мовленні диктора.

Література

1. Lewis E., Tatham M. W // Sixth European Conference on Speech Communications and Technology, September 1999 // Proceedings of the ESCA. – Budapest. – V. 2. – P. 615–618; Möbius B. Word and Syllable Models for German Text-to-Speech Synthesis // SSW3–1998. – P. 59–64.

Н. Наумова, к. філол. н.*
Запорізький національний університет (Запоріжжя)
УДК 81'322.6

ІДЕНТИФІКАЦІЯ КОРИСТУВАЧІВ АНГЛІЙСЬКОЇ МОВИ У ДІЛОВОМУ ДИСКУРСІ

The English language users' identification in Business Discourse is studied diachronically. Socio-pragmatic, discoursive and communicational parameters are considered.

Міждисциплінарний характер сучасних мовознавчих студій уможливлює вивчення когнітивно-дискурсивних аспектів мовлення у різних сферах діяльності людини. Розробка комунікативних і когнітивних засад теорії дискурсу О.М. Ільченко (2002), В.І. Карасиком (2004), К.Я. Кусько (2002), М.Л. Макаровим (2003), М.М. Полюжиним

* © *Н. Наумова, 2006*

(2004), О.О. Селівановою (2004) уможливлює мовно-історичні розвідки на матеріалі окремих дискурсів. Зважаючи на їхню обмежену кількість, об'єктом нашого інтересу є діловий дискурс (ДД) у діахронному вимірі.

Головна мета нашого дослідження полягає у встановленні причин еволюції мовленнєвих засобів ідентифікації мовців у ДД. Діловий дискурс, на нашу думку, є повторюваною функціонально-семантичною єдністю, закріпленою в певному мовленнєвому просторі. Крім мовленнєвого спілкування, ділова комунікація включає в себе й текстове спілкування, при якому одні комуніканти адресують тексти іншим або обмінюються ними.

Оскільки орієнтири дискурсу, на реляційному рівні якого домінує текст, включають адресанта та адресата [9:70], прагмасемантичний аспект дослідження передбачає розгляд фактору адресата і фактору автора. Модель дискурсу, що передбачає „трансакціонну" [2:185] взаємодію адресанта й адресата, відображує навмисний, впорядкований трансфер знання в когнітивній площині.

З метою реконструкції соціальної єдності у певному суспільстві необхідно вивчити соціальний словник (social vocabulary), тобто встановити завдяки яким структурам людина сприймає власне суспільство [13:10–11]. Дискурс ділової комунікації не є винятком. Спостереження над мовним матеріалом свідчать про зміни словарного складу англійської мови ділової комунікації відповідно потребам і прагманастановам її користувачів. Під ***користувачами*** ми розуміємо тих, хто послуговується англійською мовою з професійною метою, тобто мовців, залучених до економіко-правових, господарчих відносин у діловому дискурсі.

Аналіз соціальної приналежності авторів пам'яток, представлених у корпусі ранньоанглійської кореспонденції [14], дає підстави стверджувати, що купці (merchants) та ремісники (professionals) становили окрему сталу групу ділової дискурсивної спільноти, в той час як нестійкий прошарок непрофесійних користувачів англійської мови ділової комунікації, що його вслід за [12:101–102] прийнято найменувати „мобільним", постійно поповнювався за рахунок тих, хто поліпшував власний соціальний статус (повірників, посередників, лихварів, збірників податей, митців, пересічних громадян). Всі зазначені категорії мовців були залучені до економіко-господарчих відносин з метою оформлення листів, рахунків, боргових зобов'язань тощо.

Показовим з точки зору соціальної гетерогенності мовців є діахронна заміна значення давньоанглійського *cleric* ("*clergyman*") на середньоанглійське *clerk* ("*student, scholar*") та новоанглійське ("*secretary in an office*") із одночасним збереженням *clergy* на позначення духівництва. Зауважимо, що завдяки духовним центрам разом із писемністю розповсюджувалися і перші посібники із написання листів, в тому числі й ділових. Розглянемо датований 1678 роком приклад із підручника з укладання листів, представлений у спеціалізованій збірці ранньоанглійської кореспонденції:

[...] your Lingua, my Neighbour tells me that it is Latin and wonder that you would write Latin to me, you know I cannot stand under it and in truth you shall be no zervant of mine; I will have no zuch zervants. […] and vor those shells you zent I know not how you call them I een mended our door with them [14: 22–23].

Незважаючи на деяку схожість з усно-мовленнєвим спілкуванням, ознаками якого виступають підкреслена еліпсизація замість окремого речення разом із використанням *zervant, zent, I een mended*, подібна комунікацією відрізняється цілою низкою ознак: фіксація на матеріальному носії, незалежність від акустичної сфери, більш жорстка регламентація та організація тексту листа. Ці характеристики дають підставу розглядати текстову комунікацію між автором та адресатом стосовно відправлених устриць *those shells you zent* як особливу форму реалізації та існування ділового дискурсу в писемній формі.

Співприсутність адресата уможливлює використання референційних показників *you, my Neighbour, your Lingua, you zent, you call*. У такому випадку ці показники забезпечують у дискурсі ефект співприсутності, автоадресованості, призначеності обмеженому колові читачів [11:252]. У такий спосіб автор вступає у два ряди соціодискурсивних відношень – з об'єктами, що описуються та є відомими обом сторонам, і з адресатом [8:36]. При цьому комуніканти вирізняються рівнем освіти та інформованості.

У попередньому тексті адресат вдавався до латини, а автор, навпаки, попереджує про неможливість спілкуватися нею – *I cannot stand under it.* З огляду на посилання *you shall be no zervant of mine; I will have no zuch zervants* можна впевнено стверджувати про

різницю у соціальному статусі. Необізнаність автора щодо того, що робити з надісланими йому мушлями *I know not how you call them I een mended our door with them* свідчить на користь різного ступеня інформованості учасників комунікації. Емпіричний матеріал дає підстави вважати ділову дискурсивну спільноту комунікантів гетерогенною.

Оскільки екстралінгвальні чинники формують контекст повідомлення та задають прагмаепістемічні параметри процеса ділової комунікації, соціальне розшарування дискурсивної спільноти не могло не позначитися на жанрово-мовленнєвих регістрах ділового дискурсу зміною словникового складу. Так, у середньоанглійській період активності набуває давньоанглійський суфікс *-ere*, трансформований у *-er* на позначення виконувача дії, наприклад: *carter* (візниця) від *cart* (візок), *wever* (ткач) від *weven* (ткати).

Обмеженість норманського впливу на віддалені регіони зумовило функціонування *blacksmith, miller, shoemaker, spinner, weaver* одночасно із запозиченими *merchant, butcher, painter, mason, tailor* на позначення користувачів англійської мови, активно залучених до економіко-правових відносин. Бурхливий розвиток господарчих зв'язків призвів до необхідності запозичення іншомовних одиниць на кшталт *grocer, tailor, carpenter, bocher* франко-норманського походження. Зважаючи на те, що власних ресурсів англійської мови того часу було недостатньо для номінації її користувачів, залучених до економіко-господарчих відносин, цілком вмотивованим було запозичення мовцями вже існуючих іншомовних одиниць разом із словотворчими засобами. Так, на початку новоанглійського періоду збільшилась активність запозиченого суфіксу *-ee*, завдяки якому з'явилися *employee, addressee, trustee, grantee*.

На підставі проведеного аналізу можна виокремити лінгвальні та екстралінгвальні чинники, релевантні для ідентифікації користувачів англійської мови у діловому дискурсі. До перших належить еволюція англійської мови та становлення її як літературної, обізнаність мовців у професійному спілкуванні, здатність використовувати мовленнєві формули ввічливості, соціального підпорядкування і поваги, конвенціалізовані структури з метою прискорення текстотворення. До другої категорії таких чинників слід віднести рівень писемності, професійної підготовки, соціальний статус із відповідною освітою користувачів англійської мови.

Отримані результати дослідження є певним внеском у вивчення впливу чинника особистості на жанрово-стилістичну стратифікацію регістрів дискурсу у діахронічній перспективі. Перспективною видається окреслена можливість мовно-історичних розвідок антропоцентричного характеру на матеріалі інших природних мов у руслі когнітивно-дискурсивної парадигми лінгвістики.

Література

1. Ільченко О.М. Етикет англомовного наукового дискурсу. – К.: Політехніка, 2002. – 288 с.
2. Йокояма О.Б. Когнитивная модель дискурса и русский порядок слов. – М.: Языки славянской культуры, 2005. – 424 с.
3. Карасик В.И. Языковой круг: личность, концепты, дискурс. – М.: Гнозис, 2004. – 390 с.
4. Кубрякова Е.С. Язык и знание. – М.: Языки славянской культуры, 2004. – 560 с.
5. Кусько К. Я. Дискурс іноземномовної комунікації: концептуальні питання теорії і практики // Наукові записки Академії вищої школи України. – Вип. 4. – 2002 – С. 98 – 112.
6. Макаров М. Л. Основы теории дискурса. – М.: Гнозис, 2003.– 280 с.
7. Падучева Е.В. Динамические модели в семантике лексики. – М.: Языки славянской культуры, 2004. – 608 с.
8. Полюжин М.М. Про теоретичні засади когнітивного підходу до дискурсивного аналізу // Studia Germanica et Romanica. – Донецьк: ДонНУ, 2004. – Т. 1., № 3. – С. 32-42.
9. Потапенко С.И. Мовна особистість у просторі медійного дискурсу. – К.: КНЛУ, 2004. – 360 с.
10. Селиванова Е.А. Основы лингвистической теории текста и коммуникации. – К.: Брама, 2004. – 336 с.
11. Шмелев А.Д. Русский язык и внеязыковая действительность. М.: Языки славянской культуры, 2002. – 496 с.
12. Black J. Eighteenth-Century Britain 1688-1783. Houndmills: Palgrave, 2001. – 648 p.
13. Burke P. A Civil Tongue: Language and Politeness in Early Modern Europe // Civil Histories: Essays Presented to Sir Keith Thomas. Oxford: Oxford University Press, 2000. – 496 p.
14. The Corpus of Early English Correspondence // New ICAME Corpus Collection CD-ROM. – Bergen: The HIT-Centre, 1999.

О. Володкевич*
Київський національний лінгвістичний університет (Київ)
УДК 81'322.6

СТВОРЕННЯ Й ВИКОРИСТАННЯ БАЗИ ДАНИХ СКЛАДІВ АНГЛІЙСЬКОГО МОВЛЕННЯ ДЛЯ БЛОКУ ЛІНГВІСТИЧНОЇ ОБРОБКИ ТЕКСТІВ У СИСТЕМІ КОНКАТЕНАТИВНОГО СИНТЕЗУ

The article focuses on the problem of English speech synthesis in the System of Concatenation Synthesis. The core principles and difficulties in creating the database for the synthesis are under our analysis. The article presents the structure of the database and the corresponding concepts of its usage in the application. The database strategies underline the natural sounding of the synthesized speech and saving the individual features of a Speaker.

Імітація людського мовлення машиною цікавить вчених із давніх давен, але з XVIII ст., епохи активного розвитку механіки, збереглися два унікальні пристрої. Пневмопристрій Х. Кратценштейна [1], завдяки збудженню резонансів, міг утворювати голосні звуки, а „мовляча" машина В.фон Кемпелена [2], яка зовні нагадувала піаніно, могла імітувати звучання простих складів, міняти тривалість та інтенсивність звучання. Рівень сучасних електронно-обчислювальних машин і можливості програмування необхідних умов синтезу дозволили перетворити зусилля окремих дослідників у потужну технологічну галузь [4]. Сьогодні мовленнєві технології після випробування формантних синтезаторів, які працювали за правилами фільтрації спектральної картини з максимально збагаченого акустичного сигналу, перейшли до нового етапу й активно впроваджують конкатенативний синтез [3].

Конкатенативний, або компілятивний синтез мовлення – це синтез, що базується на конкатенації (з'єднанні) акустичних сегментів певної фонетичної розмірності, дібраних і проанотованих на матеріалі записів індивідуального мовлення диктора. Вважається, що саме цей вид синтезу генерує найбільш природне мовлення. Одиницями такого синтезу можуть бути алофони, дифони, трифони, склади та навіть цілі слова. Щодо просодії (інтонації), то вона залишається найскладнішим питанням при синтезі мовлення. Зазначені синтезатори алофонного й дифонного синтезу (ті, що моделюють інтонацію) поки що не генерують просодію необхідної якості. Мовлення звучить неприродно, що в решті решт спричинює низький рівень сприйняття тексту. І хоча найпопулярнішими та найефективнішими вважаються дифонні синтезатори штучного мовлення, Лабораторією комп'ютерної лінгвістики КНЛУ було поставлено за мету створення системи складового синтезу англійського мовлення.

За своїм фонетичним виявом склад є найменшою вимовною одиницею. Він поєднує в собі сегментну і просодичну природу. Артикуляційні роботи, виконувані для реалізації такого цілісного вимовного зусилля як склад, одразу налаштовані на скоординовану якість поєднуваних у ньому сегментів – приголосних і голосних, на особливості реалізації наголошеності й ненаголошеності для всіх сегментів складу. Тон, інтенсивність, тривалість мають тенденцію до зміни в поскладовому ритмі або в кількох складах поспіль. У результаті, при складовому синтезі меншу проблему становлять інтонація синтезованого мовлення та правила асиміляції сегментів.

Зазначені властивості складу стали першим лінгвістичним мотивом вибору цієї вимовної одиниці для конкатенативного синтезу англійського мовлення. Другим технічним мотивом стало міркування про те, що складовий синтез дає кращу розбірливість мовлення за рахунок меншої кількості точок конкатенації (у порівнянні з алофонним і дифонним синтезом).

Таким чином, одиницею синтезу було обрано склад, але склад не в системі мови, а в системі мовлення, де він може реалізувати себе як в межах словоформ, так і на міжсловесному шві, включати у свою структуру одну вершину (голосний сегмент) чи декілька вершин з підпорядкованими одному центру якостями тощо. Тому, порівнюючи зі складом у мові, його можна кваліфікувати як **квазісклад** і всі „ненормативно" реалізовані квазісклади включити до бази даних, що значно поповнює десятитисячний список складів англійського мовлення. Відповідний обсяг начитаних диктором текстів має становити 4–6 годин з включенням повторів для різних інтонаційних моделей.

* © *О.Володкевич, 2006*

Ураховуючи, що вибір та анотування такої великої кількості одиниць для бази даних на початковому етапі не автоматизовані і можуть гальмувати реалізацію інших модулів синтезатора, було вирішено опрацювати всі модулі на малому обсязі (550 одиниць) з можливістю поступового долучення приблизно 60 000 квазіскладів.

Як відомо, якість перетворення тесту різної складності й обсягу на синтезоване мовлення в системі будь-якого комплятивного синтезу, і зокрема складового, визначається переважно модулем лінгвістичного оброблення інформації, тобто прямопоропорційна якості й наповненню бази даних, тому що залежить від повноти аналізу різних характеристик текстової графічної інформації. Крім того, рівень розбірливості звучання синтезованого мовлення визначається точністю попередньої сегментації на обрані одиниці, що вносяться як елементи до бази даних, та власною якістю їх звучання.

База даних для аналізованого складового синтезу, що використовується програмою безпосередньо для генерації мовлення, складається з двох частин: інформація, що описує текстове представлення квазіскладу, та дані, що репрезентують його звуковий відповідник. База даних містить таку інформацію:

1. Квазісклад: текстове (орфографічне + пунктуаційне) представлення одиниці, інформація про словесний наголос та паузацію);
2. Позиція конкретного квазіскладу у фонетичному слові або ритмічній групі;
3. Лексичне слово або словоформа, з яких взятий конкретний квазісклад: прив'язка до лексеми чи словоформи є обов'язковою, оскільки саме за допомогою лексичного слова визначається звучання квазіскладу;
4. Фонетичне слово або ритмічна група, з яких взято конкретний квазісклад: вміщує інформацію про логічний наголос;
5. Назва звукового файлу, з якого береться звуковий інтервал конкретного квазіскладу;
6. Початок квазіскладу у файлі (у дискретах);
7. Кінець квазіскладу у файлі (у дискретах);
8. Тривалість квазіскладу (у дискретах).

Цієї інформації достатньо для неускладненого синтезу „текст на озвучення", а обмежена база даних дозволяє використовувати лише ті лексеми, словоформи і фонетичні слова, які представлені в її словнику.

У подібній базі даних складового синтезу англійського мовлення просодія частково вже закладена в самому складі та його позиціонуванні, тому не моделюється спеціально, а саме підбирається потрібний варіант квазіскладу для кожного конкретного випадку на основі реалізованих диктором інтонаційних моделей. На даному етапі інтонаційні моделі диктора враховують три основні типи речень: розповідне, питальне, спонукальне, – а також звертання й вигуки. Квазісклад з необхідними для синтезу просодичними параметрами обирається на основі таких притаманних йому ознак: наявність або відсутність словесного наголосу, наявність або відсутність логічного наголосу, позиція у фонетичному слові й ритмічній групі.

Процес добору необхідних складів для синтезу мовлення передбачає автоматичний складоподіл за законами мовленнєвої реалізації та маркування наголосу на відповідному складі слова з вхідного тексту. Цей перший етап нормалізації вхідного тексту спирається на словник слів із наголосами, що узгоджений з кількістю складових елементів. Отже, здійснюється пошук у базі відповідного текстового представлення складу та визначається його роль в конкретному лексичному слові чи словоформі. Якщо відповідника за цими двома вимогами не знайдено, на цьому етапі склад не відтворюється (на етапі поповнення бази виводиться повідомлення про відсутність інформації про конкретну одиницю у базі даних, а при необхідності прослуховування пошук повторюється для наступного слова без відтворення попереднього чи його частини). Якщо ж необхідний склад знайшовся, потрібна одиниця добирається за такими характеристиками: словесний наголос, позиція у фонетичному слові, логічний наголос. При синтезі інтонаційна модель вхідного тексту узгоджується з формальними ознаками пунктуаційних знаків, що супроводжують текстове представлення складу в базі даних. В ідеальному варіанті передбачається відтворення одиниці з наявністю всіх вище згаданих характеристик, але передбачається, що в багатьох випадках точного відповідника вхідного складу не знайдеться. Тому програма зорієнтована на пошук найкращого варіанту за послідовністю характеристик. З поступовим нарощуванням кількості елементів, які матимуть як мінімум по п'ять

варіантів реалізації, передбачено більше можливостей для створення слів, відсутніх у базі даних, і більш точного їх інтонування.

Окремим програмним етапом синтезу є конкатенація сегментів, маркування додаткових пауз та безпосереднє звукове відтворення послідовності складів.

Оскільки на даному етапі програма синтезу вміщує невелику базу даних, це унеможливлює відтворення власних назв, скорочень, абревіації, відсутніх у начитаному диктором тексті. Через обмеженість інтонаційних моделей можуть виникати проблеми із синтезом деяких видів питань та емоційно чи стилістично забарвлених речень. Має місце також самостійна проблема омонімії. Так, при синтезі мовлення необхідно розрізняти звучання омонімічних слів. Наприклад, англійське слово *read* матиме різне звучання в залежності від вживання у формах теперішнього чи минулого часу. Слово *breathe* може бути іменником або дієсловом і в залежності від цього вимовлятися по-різному. Частотність таких омонімів є достатньо високою.

У цьому синтезаторі, як і в інших відсутній модуль семантичного аналізу тексту. Варто зазначити, що проблема семантичного аналізу тексту може значно поліпшити синтез мовлення. Наприклад, якщо до звичайного розповідного речення '*My sister goes to school every day*.' поставити різні запитання:

What does your sister do every day?
Who goes to school every day?
When does your sister go to school?,

то відповідь кожного разу матиме різну інтонацію в залежності від зміни позиції логічного наголосу. Без блоку семантичного аналізу в багатьох випадках неможливо автоматично визначити місце логічного наголосу у фразі чи реченні, а отже, генерувати просодію, що відповідатиме семантиці природного мовлення. З цією метою необхідним вбачається етап семантичного аналізу тексту, тобто орієнтація на значення.

Література

1. Копелевич Ю. Х., Цверава Г. К. Христиан Готлиб Кратценштейн. 1723 – 1795. – Л., 1989.
2. http://offline.computerra.ru/1999/293/3579/
3. Dutoit T. An Introduction to Text-To-Speech Synthesis . – Springer, 1996. – 316 p.
4. Mark Tatham, Katherine Morton. Developments in Speech Synthesis. – John Wiley and Sons, 2005. – 328 p.

І. Біскуб, к. філол. н.*
Волинський державний університет імені Лесі Українки (Луцьк)
УДК 81'322.5

ЛІНГВІСТИЧНА ПАРАМЕТРИЗАЦІЯ ТЕКСТІВ ДЛЯ ТРЕНУВАННЯ КОРИСТУВАЧІВ У СИСТЕМІ АВТОМАТИЧНОГО РОЗПІЗНАВАННЯ АНГЛІЙСЬКОГО МОВЛЕННЯ DRAGON NATURALLY SPEAKING

The report deals with the complex analysis of the training session texts taken from Dragon Naturally Speaking program for automatic speech recognition. The new type of software – SEGMIX© is introduced in order to provide automatic text segmentation into words with statistics frequency calculations attached. The results of the frequency calculations are a analyzed from the point of view of artificial text vocabulary simplification which helps to optimize man-machine interaction.

Завдяки науково-технічному прогресу людство отримало автоматизовані системи управління, керування якими здійснюється за допомогою природної мови. З цією метою використовуються як усний так і письмовий варіанти мови. Керування різноманітними системами управління за допомогою природної мови є однією з актуальних проблем сучасної лінгвістики. Особової ваги набувають дослідження об'єктів на усіх мовних рівнях, оскільки інформація про інвентар та структуру цих об'єктів використовується у багатьох кібернетичних системах, що здійснюють взаємодію між людиною та електронно-обчислювальними машинами (ЕОМ).

Поява швидких ЕОМ значно збільшила потенціал інтелектуальної діяльності людини. ЕОМ довели свою здатність заміняти обчислювальну працю сотень і тисяч людей, що спричинило появу нових засобів і методів, які отримали назву кібернетичних.

* © *І. Біскуб, 2006*

Кібернетика за своєю ґенезою є суміжною наукою. До її теоретичного арсеналу належать методи і концепції математики, статистичної фізики, логіки, фізіології вищої нервової діяльності, теорії автоматів, теорії зв'язку, теорії інформатики, соціології, гносеології. Кібернетика покликана здійснювати взаємозв'язок між точними і суспільними науками. Вона є однією із наймолодших наук, існуючи близько 40 років.

Кібернетика – це наука про процеси керування і передачі інформації у машинах і живих організмах із застосуванням математичних методів.

Орієнтація на звукове мовлення при розробленні систем, які забезпечують спілкування людини й комп'ютера, спричиняє диференційований підхід до вирішення задач автоматичного розпізнавання ізольованих команд та розпізнавання зв'язного мовлення. Розв'язання цих завдань ускладнюється через існування низки факторів, а саме:

1) варіативність мовленнєвого сигналу, який подається на автоматичний розпізнавач;
2) залежність продуктивності системи від індивідуальних властивостей вимови мовця, його психологічного стану, гучності, темпу та ступеня чіткості вимови;
3) характери акустичних шумів;
4) обсяг лексичної, синтаксичної та семантичної інформації, що потребує оброблення.

Кінцевою метою процесу розпізнання мовлення є ведення діалогу між людиною і комп'ютером із використанням природної мови. Для використовуються спеціальні діалогові системи, наприклад, система KEAL. Ця система передбачає розв'язання цілої низки завдань, а саме запиту інформації, віддавання наказів або автоматичного введення інформації. Режим діалогу реалізується шляхом постановки питання користувачем. Система працює у циклічному режимі. Кожен цикл роботи включає:

1) надсилання усного повідомлення;
2) отримання усного повідомлення;
3) розроблення відповіді;
4) виконання дії, що передбачена відповіддю;
5) перехід до наступного циклу.

Система KEAL складається із наступних частин: фонетичного аналізатора, детектора фонетичних слів, детектора акустичних слів, синтаксичного аналізатора, інтерпретатора аналізу, програми просодичної сегментації, програми фонетико-орфографічної транскрипції, генератора мовлення, діалогового автомату.

Фонетичний аналізатор дозволяє перейти від реального мовленнєвого сигналу до його фонетичної інтерпретації. Кожному сегменту, який виділений фонетичним аналізатором, відповідає певна множина впорядкованих висловлювань, які розташовані згідно із зменшенням вірогідності їх використання. Ця множина висловлювань має назву фонетичного спектру фрази.

Фонетичний аналіз проводиться поетапно: на першому етапі відбувається кодування акустичного сигналу в цифровий формат, яке здійснюється за допомогою вокодера; на другому етапі здійснюється сегментація звукового сигналу; на третьому етапі відбувається ідентифікація фонем.

Центральною частиною усієї системи є діалоговий автомат, який виконує дві функції:

1) визначає, які типи фраз мають найбільшу вірогідність бути вимовленими оператором у відповідності до контексту діалогу;
2) вирішує, яке спрямування буде мати діалог (запит інформації, відповіді на питання та ін.).

Досвід використання цієї діалогової системи підтверджує гіпотезу про те, що спілкування між людиною і машиною допоможе виправити помилки у розпізнаванні.

Важливою проблемою, яка спричиняє значні складності при розпізнаванні мовлення, є проблема адаптованості систем до конкретних дикторів. Нормалізація параметрів або ознак для усунення відмінностей між індивідуальним дикторським мовленням та адаптацією систем до роботи без орієнтації на конкретного диктора відбувається шляхом запровадження універсальної шкали формантних частот. Дана шкала успішно застосовується для ідентифікації стаціонарних голосних звуків. Однак її ефективність значно зменшується при різноманітних ефектах коартикуляції. Можливі шляхи розв'язання цієї проблеми передбачають вивчення формантних дистрибуцій для

ідентифікації голосних звуків із урахуванням довготи голосного, контекстно-зумовлених часових характеристик і спектральних характеристик консонантного оточення.

При створенні систем автоматичного розпізнавання звукового мовлення особливого значення набувають експерименти в галузі сприйняття мовлення. Результати таких експериментів часто покладені в основу функціонування тієї чи іншої системи. ЕОМ, що розпізнає мовлення, часто копіює не тільки функції людського вуха, що покликані аналізувати звуки, а й деякі функції людського мозку, які відповідають за запам'ятовування та генерування логічних зв'язків між елементами мови.

Однак, як відзначають дослідники, відсоток помилок при голосовому введенні інформації все ще складає 10%, у той час, як при введенні тексту він є у 10 разів меншим і складає 0, 1%. Таким чином, постає питання про оптимізацію систем розпізнавання.

Висуваючи гіпотезу про відмінність функціонування природної мови як засобу спілкування між людьми та засобу ведення діалогу із комп'ютером, ми проаналізували лінгвістичні параметри текстів, які використовуються для тренування користувачів у системах автоматичного розпізнавання мовлення. Матеріалом для дослідження стала низка текстів із відомої системи автоматичного розпізнавання англійського мовлення Dragon Naturally Speaking.

За основу в нашому досліджені ми беремо гіпотезу про те, що сучасні системи розпізнавання розглядають мовлення як знакову систему, а комп'ютер, який розпізнає мовлення, виконуватиме функції перетворювача пропускної здатності. Визначення шляхів оптимізації роботи систем базуватиметься на аналізі надлишковості мовленнєвих знаків. Надлишковість у цьому контексті ми розуміємо як наявність для однієї мовної одиниці декількох формальних моделей відповідних джерел інформації, які можна виміряти і описати за допомогою акустичних та лінгвістичних параметрів.

Аналіз наявних теорій автоматичного розпізнавання доводить, що обидва компоненти мовлення – фонетика і просодика – є однаково важливими для розпізнавання фонетичних знаків. Збільшення словника призводить до збільшення надмірності у природному мовленні та до зменшення надмірності при закодованому мовленні. Крім того, збільшення словника веде до експонентного збільшення помилок у природному мовленні та до зменшення помилок у закодованому мовленні. Таким чином, ми вважаємо, що аналіз надлишковості знакових систем допомагає оптимізувати процес розпізнавання мовлення, встановлює кращий розподіл у процесі кодування адитивних та мисленнєвих функцій.

Тексти для тренування користувачів передбачають генерування шаблонів (templates) для їх подальшого співставлення із потоком мовлення на вході до системи, індивідуалізацією та розпізнаванням на основі ймовірнісного моделювання.

Вивчення надлишковості лінгвістичних параметрів текстів у автоматичних розпізнавачах проводилося у декілька етапів. На початку дослідження була створена програма автоматичної сегментації англійських друкованих текстів SEGMIX©, яка дозволяє генерувати списки слів із зв'язного тексту із автоматичним визначенням кількості вживань слів та встановленням абсолютної та відносної частотності усіх одиниць тексту. З метою запобігання двозначності при розпізнаванні слів (наявність тире, крапок з комами, інших знаків пунктуації), програма передбачає визначення довжини слів у літерах для оптимізації побудови списків. Програма SEGMIX© також передбачає інтегрування отриманих даних у середовище Microsoft Excel та Microsoft Access для подальшої статистичної обробки.

За допомогою розробленої програми ми проаналізували 5 текстів для тренування користувачів у програмі автоматичного розпізнавання мовлення Dragon Naturally Speaking. Тематика текстів належить то тем „Інформаційні технології", „Автоматичний синтез та розпізнавання мовлення", „Кіберпростір", „Комп'ютерний гумор". Особливо показовим з точки зору статистики повторення слів є текст „Talking to Your Computer" (Instructional). Цей текст є по суті інструкцією до користування автоматичним розпізнавачем, і одночасно міні базою даних, яка створюється для покращення процесу розпізнавання шляхом утворення індивідуальних кальок для кожного окремого користувача.

Текст „Talking to Your Computer" складається із 404 слів. Фрагмент статистичних результатів застосування програми SEGMIX© для сегментації цього тексту наведений у таблиці 1.

Кількісний аналіз засвідчив, що більшість слів (156 = 39%) мають частоту вживання 2, що свідчить про намагання системи підтвердити звукову модель кожного з цих слів шляхом вживання в позиції всередині речення, а також у фінальній або ж ініціальній позиції. Відносна частота таких слів складає 0, 5 %.

Абсолютну частоту 3 (0, 75 %) мають тематично нейтральні слова, наприклад *first, another, using, just, have clearly* та ін. Слова із частотою 4 (1 %) умовно поділяються на дві групи: теоретично нейтральні (наприклад *this, more, each* та ін), а також слова, що є відносно тематично забарвленими (наприклад *noise, speaker, naturally* та ін.).

Таблиця 1. Результати автоматичної сегментації тексту „Talking to Your Computer" (file: 'Source4.doc').

СЛОВО	АБСОЛЮТНА КІЛЬКІСТЬ	ВІДНОСНА КІЛЬКІСТЬ
We'd	3	0, 75%
like	5	1, 24%
read	3	0, 75%
aloud	2	0, 50%
minutes	3	0, 75%
while	2	0, 50%
computer	14	3, 48%
listens	3	0, 75%
learns	2	0, 50%
speak	6	1, 49%
When	9	2, 24%
you've	2	0, 50%
finished	2	0, 50%
reading	2	0, 50%
we'll	2	0, 50%
make	2	0, 50%
some	2	0, 50%
adjustments	2	0, 50%
then	4	1, 00%
you'll	2	0, 50%
able	3	0, 75%
talk	4	1, 00%
your	4	1, 00%
words	8	1, 99%
appear	2	0, 50%
screen	2	0, 50%

Слова із частотністю 5, 6, 7 (1, 25 – 1, 75 %) також поділяються на дві вищезгадані групи (перша: *like, they, other, that* та ін; друга – *person, recognize, recognition* та ін.).

Найбільшу частотність 8 і більше мають тематично забарвлені слова, які вказують на топікальні характеристики цілого тексту, наприклад *computer* (14 – 3, 48 %), *speech* (12 – 2, 99 %), *people* (11 – 2, 74 %), *words* (8 – 1, 99 %), за винятком *what* (11 – 2, 74 %) та *when* (9 – 2, 24 %), які вживаються для ініціації інтерогативно-інформаційних речень.

Загальна статистика свідчить про те, що переважна більшість слів (79 %) мають надзвичайно низьку частотність (0, 5 – 1 %), що сприяє оптимальному створенню великої кількості кальок на основі відновно невеликого за об'ємом зв'язного тексту, який несе, окрім усього іншого, і велике інформаційне навантаження.

Аналіз словникового складу текстів продемонстрував, що переважна більшість слів (близько 80%) належать до теоретично нейтрального лексикону сучасної англійської мови, а також до лексикону словникових визначень (за інформацією Longman Dictionary of Contemporary English (2005). Як відомо, у словникових дефініціях електронних словників видавництва LONGMAN використовується спеціальна добірка слів *The Longman Defining Vocabulary*, яка включає близько 2000 слів для оформлення пояснення слів у словниках. Слова для цього вокабул яру добираються на основі розрахунків

частотності, проведених на матеріалі електронних текстових корпусів, які входять до складу мережі The Longman Corpus Network. І хоча колекція з п'яти текстів є досить різноманітною за тематикою, словниковий склад текстів є відносно однорідним, що свідчить про штучну симпліфікацію добору слів при спілкування людини і машини.

Комплексний аналіз лінгвістичних параметрів текстів, що вживаються для тренування користувачів у програмах автоматичного розпізнавання мовлення засвідчив, що переважна більшість слів (близько 80%) мають надзвичайно низьку частотність вживання, що можна легко пояснити оптимізованим підбором слів для тренінгу, який досягається шляхом укладання зв'язного тексту із мінімальним повторенням слів, і, відповідно, збільшенням кількості слів, які будуть вимовлені користувачем у процесі підготовки до роботи із системою розпізнавання мовлення. Вокабуляр згаданих текстів на 95 % складається із спрощеного лексикону сучасної англійської мови, що іще раз доводить гіпотезу про спрощення стратегій спілкування при ведені діалогу між людиною і комп'ютером.

Н. Філіппова, к. філол. н.*
Національний університет кораблебудування імені адмірала Макарова (Миколаїв)
УДК 81'33

КОГНІТИВНІ ЗАСАДИ АНАЛІЗУ КОНЦЕПТУ „ПРИКЛАДНА ЛІНГВІСТИКА" ЯК СПЕЦІАЛЬНОСТІ

Тема статті пов'язана з концептуальним аналізом поняття „Прикладна лінгвістика" як основи навчального плану спеціальності "Прикладна лінгвістика" (6.030505 – „бакалавр", 7.030505 – „спеціаліст", 8.030505 – „магістр").

Актуальність зазначеної теми базується на необхідності створення науковообґрунтованої і прагматичної моделі навчання із названої спеціальності.

Як відомо, прикладна лінгвістика з'явилась наприкінці 20-х р.р. XX ст., а як самостійна лінгвістична дисципліна вона виокремилася у другій половині XX ст., хоч проблеми прикладного характеру (наприклад, створення алфавітів, систем письма, навчання іноземній мові, тлумачення й коментування текстів тощо) мовознавство розв'язувало ще з давніх-давен. Прикладна лінгвістика (ПЛ) – це така сфера мовознавства, яка займається лінгвістичним конструюванням і вирішує конкретні завдання машинної обробки мови та її використання. У широкому розумінні – це моделювання мови, мовлення, породження мови и сприйняття мови, мовного впливу.

ПЛ — це діяльність з використання наукових знань про устрій і функціонування мови в нелінгвістичних наукових дисциплінах і в різних сферах практичної діяльності людини, а також теоретичне осмислення такої діяльності.

ПЛ – це комплексна дисципліна, пов'язана з психологією, філософією, фізіологією, математикою, логікою, соціологією, інформатикою, програмуванням та ін.

Основний корпус функціональних досліджень, які проводилися у надрах ПЛ, врешті, зводилися до описання й моделювання фонетичної, граматичної, семантичної структури наукових, технічних, організаційних, усних та письмових документів, створення словників, тому одне з завдань ПЛ – це побудова формальних моделей, тип моделі вказує на шляхи алгоритмізації. Тільки за допомогою формальних моделей можна автоматизувати обробку даних.

Основний предмет ПЛ – розробка й удосконалення структурних та формальних методів аналізу й опису мови, опис (модель) проблемної області для вирішення конкретної задачі, вивчення й розробка засобів оптимізації функцій мови (якісна характеристика методів ПЛ): комунікативної функції (фатична (контактно-встановлювальна), інформаційна, впливу); соціальної функції (соціолінгвістика, мовна політика, орфографія, орфоепія, політична лінгвістика, теорії перекладу, машинний переклад, викладання мов, інформаційно-пошукові системи, створення штучних мов, теорія кодування); епістемічної функції (мовна система як засіб збереження та передачі інформації; відображення національної самосвідомості; лексиграфія, комп'ютерна лексикографія, термінологія, термінографія, корпусна лінгвістика); когнітивної функції

* © *Н. Філіппова, 2006*

(мислення людини, пізнання дійсності, психолінгвістика (афазіологія), комп'ютерна лінгвістика, лінгвістична кримінологія) [1].

Таким чином, кількість та види лінгвістичних задач прикладного характеру важко навіть передбачити, тому що неможливо передбачити всі інтерактивні комунікативні ситуації („людина–людина"; „людина–комп'ютер") та всіх міждисциплінарних контактів лінгвістики з іншими науками.

На жаль, широка семантична й термінологічна невизначеність цього поняття не дозволяє на сьогодні однозначно відповісти на питання: *Як готувати спеціалістів різного рівня?*

Але нам добре відомо, що людина, по-перше, користується мовою для отримання, обробки, фіксації, збереження, організації, накопичення, використання і збільшення інформації про світ. По-друге, людина обробляє всю цю отриману, оброблену, зафіксовану, збережену, організовану, накопичену інформацію мовою. По-третє, людина використовує мову в процесі своєї комунікативної діяльності, тобто у двосторонньому інтерактивному процесі обміну й створення значення. По-четверте, людина використовує мову для того, щоб несвідомо або цілком свідомо впливати на інших людей. Узагалі сучасне розуміння значення мови для людини й людства полягає в тому, що мова – не просто інстинкт інтелекту, оскільки у ній (мові) закодоване людське життя у всій його складності й різноманітті сприйняття.

Таким чином, вже сама роль мови у житті людини вказує на ті блоки знань, які концептуально повинні бути означеними в навчальних планах цієї спеціальності:

1) теорія інформації, математичне моделювання, семіотика, лінгвістика, корпусна лінгвістика;
2) психологія, моделювання штучного інтелекту, філософія, логіка, психолінгвістика;
3) комунікативна лінгвістика, міжкультурні комунікації, переклад;
4) комунікативні технології, психолінгвістична діагностика.

Врешті, прикладна лінгвістика – розділ мовознавства, спрямований на розв'язання практичних завдань різних галузей науки й техніки, повсякденного життя людини, суспільства на підставі теоретичного доробку досліджень мови й мовлення.

Зацікавленість прикладною лінгвістикою, яка досить широко проявляється зараз, безсумнівно, пов'язана з атмосферою розуміння цілісності природних і соціальних явищ, що є базою для розвитку світової науки з 80-х років. „Синергетичність (від гр. synergeia – спільна дія, взаємодія) – принцип дослідження, згідно з яким предметом аналізу є складні, відкриті нелінійні суперсистеми, що функціонують за рахунок взаємодії їхніх підсистем і перебувають у стані більшої чи меншої рівноваги, а об'єктом – механізми, що регулюють динаміку, самоорганізацію та збереження цих систем. Самоорганізація здійснюється шляхом спільного функціонування елементів і підсистем, які кооперуються задля збереження системи" [2: 508].

Сучасна лінгвістка на початку XXI ст. визначає перспективи свого подальшого розвитку. Еволюція нашої науки характеризується як зміна трьох наукових парадигм: 1) генетичної, історичної; 2) таксономічної, системно-структурної; 3) комунікативно-функціональної, прагматичної. Тобто загальною тенденцією сучасного лінгвістичного пізнання є рух від сфери систем до центру всіх цих систем – до людини. Гуманізація мовознавства, поворот до мовної особистості як медіума для проникнення в мову, занурену в життя, зумовили перехід традиційних лінгвістичних дисциплін на новий науково-пізнавальний рівень. Якби ми взялися за мету зробити огляд лінгвістичної еволюції, то могли би побудувати "фрейм", складові слоти якого створили б уявлення про цей концепт "лінгвістика" і показали б його певні системні зв'язки. Так, можна виокремити сім інформаційних блоків:

1. <u>Слово та його значення.</u> Два ракурси аналізу слова: ономасіологічного (номінативно-репрезентативного), що апелює до орієнтації мовної одиниці на ментальний зв'язок із світом дійсності, та семасіологічного (внутрішньоструктурного), який репрезентує слово відносно до інших одиниць мови у лінійному ряді або в парадигматичних ґрунтах (моделювання).

Особливо треба виділити перспективні дослідження конотації: індивідуальний, суб'єктивний аспект, що залежить від <u>особистості мовця,</u> тактики <u>мовленнєвої поведінки,</u> спрямованої до успіху комунікації.

2. Теорія номінації. Творячи слово, номінатор використовує дві мисленнєві операції: від конкретного чуттєвого образу до абстрактного уявлення, а далі на основі пізнання властивостей об'єкта конкретизується структура поняття з огляду на його зв'язки, один із зв'язків включається у мовленнєві відношення і набуває конкретного значення (ментальні механізми творення номінативних одиниць).

3. Психолінгвістика. Когнітологічні моделювання психологічних процесів, що передують мовленнєвій діяльності, мовної людської поведінки, породження мовлення, розуміння мовлення, застосовуючи для дослідження таких когнітивних феноменів, як пам'ять, увага, знання, навчання, методи математичного моделювання і кібернетики.

4. Теорія картини світу. Мовна картина світу як загальна інтеріоризація дійсності носіями однієї мови отримує атрибут національно-мовної. Останнім часом у світі взагалі й в Україні зокрема надзвичайно актуальним стало вивчення різних фрагментів національно-мовної картини світу з огляду на специфіку етнічної свідомості мовців, вплив звичаїв, соціокультурних, фонових знань, традицій, обрядів, вірувань, міфології і т. ін.

5. Функціональний напрям. Функція є основою категоризації концептуальної системи людської свідомості та, зокрема, внутрішнього лексикону мовців і, відповідно, мовних явищ, що проявляються у комунікативному акті.

Нині функціонально-семантична категорія визначається як концептуально інтегрована система різнорівневих мовних та комунікативних засобів, здатних виконувати одну спільну функцію.

6. Теорія тексту / дискурсу. Загальною стратегією сучасної лінгвістики тексту є перехід від вивчення текстової інтегрованої структури до аналізу тексту як комунікативної системи**,** детермінованої такими її складовими, як особистість мовця (автора) у сукупності його психологічних, ментальних, соціально-культурних, етнічних та ін. властивостей; адресат та його рівень сприймання, екстралінгвістичні характеристики двох типів ситуації: того, про що йде мова у тексті, та того, що опосередковує його творення у соціокультурному контексті дійсності. Отже, лінгвістика тексту може найближче підійти до вирішення глобальної мовознавчої проблеми – дослідження мовного відображення менталітету, культури етносу та впливу світової культури на етнічну.

7. Комунікативна лінгвістика. Комунікативна лінгвістика шляхом екстраполяції власного об'єкта — дискурсу — на різні сфери знань склалася у межах теорії комунікації, утворюючи нові мовознавчі дисципліни та їхні суміжні галузі. Формування дискурсу залежить від того, які комбінації виникають під час взаємодії його компонентів, пов'язаних з партнерами комунікації (їхніми комунікативними ролями й інтересами, особистісними характеристиками, міжособистісними стосунками, загальним досвідом та інформованістю), з темою й предметом спілкування, мовленнєвою ситуацією (офіційними та неофіційними обставинами, наявністю чужих, залученістю оточуючих предметів до розмови, супроводженням бесіди діями, можливістю безпосереднього контакту, змогою бачити й чути, а також перешкодами: комунікативним шумом, цейтнотом).

Нарешті, як ми вже зазначили, необхідною умовою для прикладного використання мови є можливість моделювати мовні явища й факти. Відомо, що модель – це логічна конструкція, яка відтворює характеристики об'єкта, модель – це система, яка відображає або відтворює об'єкт, вона може заміщувати його й надавати нову інформацію про цей об'єкт. Модель відображає не тільки реальні якості об'єкта, але й нашу точку зору на об'єкт, ті вимоги до відповідності моделі модельованому об'єкту, які ми завжди пред'являємо. Модель обов'язково потрібно перевірити (верифікувати). Лінгвістичне моделювання обов'язково передбачає використання абстракції та ідеалізації.

Іншими словами, модель – це зразок, який є стандартом для подальшого відтворення, це типова схема якоїсь структури чи відношень між елементами структури, це парадигма організації класу лінгвістичних одиниць або їх сукупності, а конструювання моделі – це не тільки один із засобів відображення мовних явищ і процесів, але й об'єктивний практичний критерій перевірки (верифікації) правильності наших знань про мову.

Системність мови полягає в цілісності взаємозв'язків між мовними елементами, які відбивають найсуттєвіші ознаки внутрішніх та зовнішніх зв'язків мовного об'єкта. Структурність мови полягає в засобі організації мовної системи, відношень і зв'язків між її елементами. Унаслідок того, що мова системна і структурна, можна будувати модель,

тобто такі конструкції, у яких відтворюється та чи інша реальність, які складаються із взаємозв'язаних елементів, функціонуючих як одне ціле.

Мовна система – це система, що поряд із системними фактами та явищами допускає використання несистемних фактів і явищ, оскільки вона пов'язана з іншими системами. Мовна система – це ієрархічно організована й певним чином упорядкована цільність взаємопов'язаних систем і підсистем (фонетичної, семантичної, граматичної), що перебувають у парадигматичних і синтагматичних відношеннях. Системність і структурність дають змогу стверджувати, що у мовній системі можливе моделювання. Природний світ відображається в мові й моделюється тільки тоді, коли він представлений у вигляді мовних творів (текстів). Мова ж проявляється у функціонуванні в мовленнєвій діяльності (звичайне спілкування, переклад, навчання мові), а отже, моделюються і різні типи мовленнєвої діяльності (моделі розуміння, породження тексту, моделі перекладу, інтелектуальні системи, які використовують діалог на природній мові).

Таким чином, особливістю мовного моделювання є те, що моделюється і картина світу, яку мова представляє, і мовна діяльність, в якій мова функціонує, і мовний механізм, що управляє правилами створення слів, побудови речень, тексту, і мовні твори, які є результатами мовної діяльності [3].

Коли визнати, що з функціональної точки зору прикладна лінгвістика – це „наукова дисципліна, в якій вивчаються і розробляються засоби оптимізації функціонування мови" [1], то особливість прикладного опису полягає в тому, що він орієнтується на конкретні умови, а не на всю мову в цілому. Якщо ж визнати, що найважливішим для прикладного використання лінгвістичних знань є оптимізація, тоді оптимізується, тобто моделюється такий мовний компонент, який потрібний для вирішення конкретної поставленої проблеми, а модель як результат представлення конкретного мовного компонента виявляється тільки одним з етапів прикладного опису, і не викликає сумніву той факт, що сьогодні вся сукупність проблем ПЛ вирішується за допомогою комп'ютерної обробки даних. Тому інформатика, програмування, експертні системи, бази даних, лінгвістичне забезпечення комп'ютерних систем – це ті галузі знань, що безпосередньо пов'язані з можливістю обробки мовних даних.

Тому, підбиваючи підсумок, слід зазначити, що концептуально основою для створення ефективної навчальної системи може бути концептуальна інтегральна модель, функціональна за спрямованістю, конструктивна за духом, яка може показати місце мови у вирішенні проблем світорозуміння та її участі в пізнавальній діяльності людини.

Література

1. Баранов А. Н. Введение в прикладную лингвистику. – М.: УРСС, 2001.– С. 7.
2. Селіванова О. Сучасна лінгвістика: термінологічна енциклопедія. – Полтава: Довкілля-К, 2006. – 716 с.
3. Філіппова Н. М. Вступ до прикладної лінгвістики: Моделювання у мові: Навчальний посібник. – Миколаїв: НУК, 2004. – 36 с.

ЛІНГВІСТИЧНІ ПОРТАЛИ, ІНФОРМАЦІЙНИЙ ПОШУК

Н. Дарчук, к. філол. н, ***О. Сірук , В. Сорокін****
Київський національний університет ім. Т. Шевченка (Київ)
УДК 161.2. 81'33

ЛІНГВІСТИЧНИЙ ПОРТАЛ MOVA. INFO: ТРИ РОКИ В МЕРЕЖІ ІНТЕРНЕТ

Лінгвістичний портал MOVA.info в мережі Інтернет є одним із провідних центрів для всіх, хто зацікавлений у швидкому отриманні якісної інформації про українську мову в контексті як культурного та суспільно-політичного, так і "професійного" її життя. Стаття присвячена: звітуванню про здійснену нами останнім часом роботу; підсумкам опрацювання напрямків розвитку, аналізу продуктивності, пошукам нових ділянок роботи.

Коротка історична довідка. Лінгвістичний портал MOVA.info в мережі Інтернет існує вже більше трьох років. 12 січня 2003 року було запущено першу версію

* *© Н. Дарчук, О.Сірук , В. Сорокін, 2006*

лінгвістичного порталу. Вона містила такі основні рубрики, як «Наші проекти», «Новини», «Конференції», «Опитування», «Лінгвокалендар».12 жовтня 2004 – офіційна презентація лінгвістичного порталу MOVA.info в Інституті філології. 27 квітня 2006 – підсумки роботи.

Огляд **основних напрямків роботи** відповідно до тематичної **структури лінгвістичного порталу MOVA.info.** свідчить про **розширення тематичного та кількісного наповнення порталу.**

Структура порталу

Інформаційний блок

- **Анонси конференцій**
- **Новини**
- **Лінгво-календар**
- **Читальна зала**
- **Лінки**

Інтерактивний блок

- **Довідкова служба**
- **Форум**
- **Опитування**
- **Гостьова книга**

Розважальний блок

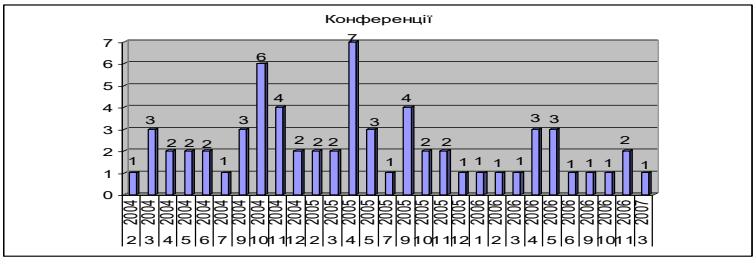
- **Ігри**
- **Цитати**
- **Конкурси**

Науково-освітній блок

- **Електронний підручник**
- **Словники**

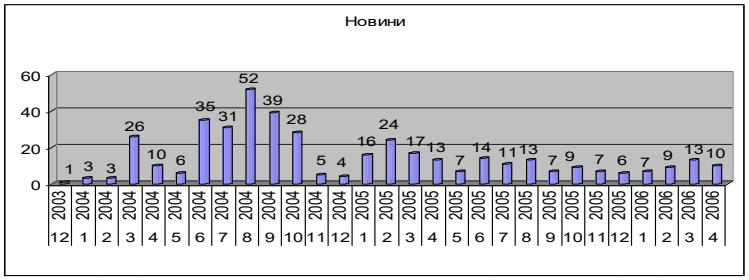
Інформаційний блок

1. Анонси **конференцій** і подій з філологічного життя



Загальна кількість анонсованих конференцій – 65. До рубрики потрапляють оголошення як про українські, так і про зарубіжні конференції та семінари з філологічної тематики.

2. **Новини**



Загальна кількість новин – 426. Стосуються здебільшого подій літературного життя, нових книг та книгодрукування в Україні, політичних питань функціонування української та інших мов в Україні, міжнародні філологічні новини. Наповнення рубрики регулярне. Працює розсилка новин для зареєстрованих користувачів.

3. **Лінгвокалендар** – 476 персоналій та знакових для філології подій (інформація як коротка, так і розгорнута, з фотографіями).

4. **Читальна зала** – регулярне оновлення бази нових надходжень з лінгвістичної та дотичної тематики. Поповнюється з сайтів бібліотек, філологічної періодики, книжкових ярмарків. Також поповнюється науковими статтями з різних галузей мовознавства.

5. **Лінки** – широкий тематичний і кількісний спектр. У центрі уваги – ресурси з лінгвістики, і перш за все – української. Окрім цього – літературознавство і персональні сайти митців, видавнича справа, історія, культура, мистецтво, електронні бібліотеки, здебільшого – українські. Розділ наново впорядковано, додано більш ніж триста нових посилань.

1. Українська мова (сайти, безпосередньо пов'язані з різними аспектами функціонування української мови у світах реальному та віртуальному, мовні ресурси від освітянсько-новинних до розважальних)

2. Українська література (про літературу – як від критиків-теоретиків, так і від митців-письменників. Інформація від сивої давнини до активної сучасності, від лірики до публіцистики, переклади, авторські сторінки та багато іншого на всі смаки.)

3. Українська історія, культура, мистецтво (Сучасне мистецтво України, історія України, віртуальний театральний проект, Українська Спілка, втрачені храми м. Києва. Цікавинки для істориків та усіх любителів культури і мистецтва.)

4. Українські видавництва (віртуальний видавничий калейдоскоп України. Сайти таких найвідоміших видавництв, як "Просвіта", "Афіша", "Літопис", видавництво ім. Олени Теліги, "Дух і літера", "Знання", "Основи", "Смолоскип", "Кальварія" тощо.)

5. Авторські сайти (Подано сайти письменників та поетів, сучасних і не дуже, трохи укарїнських, трохи зарубіжних. Сайти письменників-емігрантів.)

6. Літературні угрупування (сайти літературних угрупувань – від локальних до загальноукраїнських.)

7. Електронні бібліотеки (представлено солідні українські електронні книжкові зібрання типу "Бібліотека світової літератури – оригінали та переклади", "Поетика" тощо. Незамінний помічник для тих, хто шукає інтернет-версії творів українських класиків та сучасників.)

8. Книгарні та бібліотеки (Онлайн-книгарні та бібліотеки України)

9. Для маленьких мовців (Добірка сайтів для дітей. Вірші, скоромовки та всілякі цікавинки.)

10. Громада (Інформація про українців за кордоном та різні товариства України.)

11. Музика (Добірка сайтів українських виконавців. Можна знайти і музику в форматі MP3, і кліпи, і тексти пісень)

12. Вищі навчальні заклади (Добірка сайтів українських університетів, інститутів, вищих навчальних закладів з філолологічними факультетами, які певним чином пов’язані з українською мовою та зокрема – з комп’ютерною лінгвістикою.)

13. Преса (Добірка сайтів українського та зарубіжного радіо, преси, телебачення, загалом передових видань українського та світового інформаційного простору)

14. Україна політична (Сайти офіційних установ, політичних організацій, партій, урядові портали. Теоретично – живі зразки офіційно-ділового стилю сучасної української мови.)

15. Культурні центри та інститути (Добірка наукових культурно-мовних інституцій для тих, хто хоче вивчити іноземні мови або просто хоче поспілкуватися іноземною мовою)

16. Університети Світу (Низка університетів світу, які мають зв’язок з лінгвістикою) тощо.

Інтерактивний блок

1. **Довідкова** служба

Загальна кількість запитань за період з часу презентації – 357, загалом – 392. Тематика більшості запитань – з написання великої-малої літер у власних назвах, відмінювання і транслітерування прізвищ, стандартів офіційно-ділового мовлення, стилістики («чи можна так сказати…»), перекладу термінів з інших мов. Часто уточнення і розгорнуте обговорення запинань та проектів йде на **Форумі. У**се це вказує напрямки і дає матеріал для роботи викладачам, вчителям, термінознавцям і правописній комісії.

2. **Розсилка** новин, інформації про конференції та відповідей довідкової служби – функціонує регулярно, йде відкрите спілкування з користувачами щодо обговорюваних запитань. Для того, щоб отримувати повідомлення про нові конференції та новини на сайті, слід зареєструватися та на сторінці «Мій профіль» відмітити відповідні прапорці. У відповідь абонентам будуть надсилатися заголовки та короткий опис новин або конференцій. Отримувати повідомлення про структурні оновлення на порталі можна також шляхом реєстрації на відстеження певної теми.

3. **Опитування** стосуються переважно проблем соціо- і психолінгвістики. Теми і варіанти відповідей пропонуються як адміністраторами лінгвістичного порталу, так і користувачами.

4. **Гостьова книга. В**ідгуки – це й конструктивна критика та підказки (що змусило нас взятися до редагування підручника і тестів, примушувало пильно вичитувати тексти портальних матеріалів), і позитивні відгуки, які ще більше надихали на роботу усвідомленням того, що це не просто потрібно, а необхідно широкому загалу. Надійшли також пропозиції співпраці й допомоги, що найбільше нас радує, адже, як говорить давня мудрість, добре щось зробити самому, а ще краще – зуміти своєю роботою зацікавити людей так, щоб вони самі захотіли зробити, навчитися, допомогти. Ми охоче надали місце для започаткованих іншими розробниками проектів (напр., Сергій Адамчук запропонував нам розвинути і інтегрувати з mova.info «Відкритий словник неологізмів»).

Науково-освітній блок

Електронний підручник з української мови. Була проведена серйозна робота щодо зміни дизайну, вдосконалення структури, загального редагування. Додано розділ «Синтаксис», з тестами. Кожний підрозділ Електронного підручника **структурований** у такий спосіб:

теоретична частина, де коротко викладено теоретичний матеріал із визначеної теми, обґрунтовано основні терміни та поняття, офоепічні або орфографічні правила;

практична частина, де якій подаються практичні завдання для засвоєння викладеної теорії та набуття практичних навичок.

У кожному з модулів пропонуються практичні завдання двох типів: тести для самоперевірки та вправи. Практичні завдання мають зв’язок з теоретичною частиною: у разі вибору неправильної відповіді у тестах або помилки у вправі висвітлюється правило або теоретичний пункт, що допоможуть виправити зроблену помилку. Вивчення кожного модуля закінчується контрольними тестами, які дозволять оцінити набуті знання. У межах модуля є термінологічний словник тезаурусного типу, де системно подається трактування основних термінів певного мовознавчого підрозділу.

Комплексна навчальна система розрахована на широкий загал – учнів загальноосвітніх шкіл, абітурієнтів, студентів-філологів та інших спеціальностей вищих навчальних закладів і всіх, хто бажає вдосконалити та перевірити свої знання з

української мови. (Апробовану систему планується рекомендувати для впровадження в навчальний процес вищих і середніх навчальних закладів України.)

Розміщення автоматизованого навчального комплексу в мережі Інтернет у режимі “online” надасть вільний доступ до неї як в Україні, так і за її межами.

Окремого розділу заслуговує і **словникарська діяльність. Т**ут є як **статичні** і популярні проекти на кшталт чотиримовного словника наукової термінології, так і **відкриті** для користувача системи («Відкритий словник новітніх термінів»).

Презентовано ряд комп’ютерних словників.

1. Спеціалізований **тезаурус** з **комп’ютерної ідеографії** (онлайн-версія аналогічного комп'ютерного словника) представляє терміносистему комп’ютерної ідеографії, яка на сьогодні налічує 75 термінів. Одиниці є характерними для цього типу словників – це іменники й іменниково-іменникові чи іменниково-прикметникові словосполучення у різних комбінаціях (також у формі абревіатур) обсягом від 2 до 4 слів. Переважна частина одиниць стосуються тільки предметної галузі, але є й спільні з іншими мовознавчими розділами. Макроструктура тезауруса представлена родо-видовим деревом термінів. Максимальна глибина ієрархізації складає шість інтервалів. Словникова стаття ТКІ складається з заголовної одиниці-терміна та дефініції. Автори: О. Сірук (лінгвістична частина), В. Сорокін (програмне забезпечення).

2. **Семантичний словник** української мови є пілотним проектом лабораторії комп'ютерної лінгвістики та кафедри сучасної української мови. На сьогодні у базу введено близько 30 слів назв рослин і показано їх семантичні складові та відношення між ними. Слова можна переглядати як за їх назвою, так і за семантичними групами чи значеннями, що складають ці слова. Лексичне наповнення підготували студенти 3-го курсу спеціальності "Українська мова та література" під керівництвом д.ф.н., проф. Вовк П. С. Проект відкритий для наповнення і розширення. Можна легко додавати інформацію про інші слова. Також, за необхідності, можна змінювати і розширювати саму стуктуру сайту і бази даних.

3. **Лінгвістична енциклопедія –** довідковий філологічний ресурс. Складається з двох великих розділів: "Події", де подано коротку інформацію про важливі для вітчизняної мовно-літературно-освітянської історії дати, та "Особи", присвячену видатним українським мовознавцям, літературознавцям, письменникам. Тепер у списку подій та особистостей можна відбирати записи за місяцями. Довелося також відмовитися від стандартного для бази даних поля дати.

4. Літературний проект **"Вавилон".** На сьогодні у базі даних міститься загальна інформація про 100 письменників. Переважно це лауреати Нобелівської премії. Пошук можна вести за допомогою поля, куди вводиться ім'я автора або назва твору, як мовою оригіналу, так і українською. Також наявний алфавітний покажчик прізвищ авторів.

Інформація про авторів структурована таким чином:

1. загальний блок (повторюється на кожній сторінці про автора). Містить загальну інформацію про автора – фото, дата і місце народження та смерті, ім'я в оригіналі, короткий опис його творчості тощо;
2. список творів. Ця сторінка завантажується за замовчуванням. Містить список основних творів автора. Для деяких творів може бути подана детальніша інформація на окремих сторінках;
3. біографія. Якщо ця сторінка активна, то на ній розміщена біографія автора;
4. критика. Якщо ця сторінка активна, то на ній розміщено загальний огляд творчості автора.

Джерела інформації про авторів – це довідник про нобелівських лауреатів, сайт нобелівського комітету, різні енциклопедії, у тому числі електронні. Проект відкритий для наповнення і розширення. Розробник – Д. Кузьменко.

Розділ словників **поповнився** також такими лексикографічними проектами, як "Глоса" (англійсько-українська частина. Словник створено в Київському національному лінгвістичному університеті), першою версією українсько-сербського словника, відкритим словником проти суржику (дякуємо за ідею користувачеві нашого порталу), чотиримовним словником наукової термінології з фізики, хімії та теоретичної хімії (надалі планується активізувати підсистему виведення наголосу в українській та російській частинах словника). Завдяки співпраці з Інститутом мовознавства ім. О. О. Потебні НАН України зроблено вебверсію частотних словників наукової прози, розроблених у відділі структурно-математичної лінгвістики інституту. **Доопрацьовано**

частотний словник публіцистичного стилю, зокрема додано можливість будувати конкорданс, переглядати частотні характеристики граматичних класів, подано зведений частотний словник, функцію попарного порівняння між підвибірками та зі зведеним словником. О**новлено** функціональність поетичного частотного словника: тепер картку конкордансу можна будувати не тільки для слів, але й для граматичних кодів (з відповідної сторінки статистики).

1. Якщо ви хочете сказати щось українською, а вдалого слова підібрати не можете (в кращому випадку ви знаєте назву цього терміну англійською чи російською або такого слова навіть у словнику знайти не можна – звертайтесь до "**Відкритого словника новітніх термінів**". Проект пропонує придумати, розглянути різні варіанти та вибрати той, який найкраще відповідає контексту та найкраще вписується в українську мову. Тут можна пропонувати українські відповідники, голосувати за різні варіанти та обговорювати терміни.

2. «Відкритий словник виправлень суржику». Проект присвячений виконанню одного з найважливіших лінгвістичних завдань – очищенню літературної мови від ненормативних конструкцій. Словник є відкритим, тобто дає можливість зареєстрованим користувачам порталу вносити помилкові форми з контекстом, правильний варіант відповіді та визначати категорію помилки. Словник відображає особисту думку дописувачів, яка може не збігатися з нормами української мови.

Слово	Фраза	Правильно	Джерело помилки	Посилання/дата	Категорія	Коментар
виплекувати	Галичина виплекує таких дітей	Галичина плекає, ростить таких дітей	Львівське радіо	4.III.2001 14:05	непр.сл.	безграмотна фантазія мовця

Розважальний блок

1. Конкурс «Наймилозвучніше слово української мови». Цей конкурс-опитування започатковано нами для усіх закоханих в українську мову: для батьків і дітей, для академіків і для школярів, для українців з діда-прадіда і для тих, хто тільки починає відкривати для себе барви української мови. Можна або обрати один з представлених у голосуванні варіантів, або додати те нове слово, яке вважаєте наймилозвучнішим в українській мові, з коротким описом його значення. Проголосувати можна окремо в кожній з трьох категорій. Кожен користувач має можливість через два тижні після першого голосування проголосувати ще раз, а також додати не більше 5 слів. Користувачами серед жіночих імен були відзначені (подаємо з ремарками авторів) Богдася, Гафійка, Ганнуся, Даринка («лагідне від "Дарина", ім'я – свідчення того, що Вона є подарунком (Долі?)»), Іванна, Ілона, Леся, Олена («асоціюється з оленем тобто з кимось дуже граційним, рухливим, тендітним та спритним»), Оленка, Олюся, Орися, Софійка тощо. Серед чоловічих імен – Василь, Гриць, Данило, Димитрій, Дмитро, Кирило, Максик, Микитась, Назар, Овсій, Орись, Тиміш, Ярема, Яросик. Серед слів загального вжитку лідерами стали "малятко" та "кохання".

Загальний рейтинг розділів

*Загальна кількість відвідувань – * Назва розділу*

27310 – * Новини
22843 – * Особи
18331 – * Перекладач
16772 – * Чотиримовний науковий словник
14528 – * Анонси конференцій
13810 – * Лінгвістична бібліографія
10926 – * Відкритий словник
8811 – * Електронний підручник української мови
7836 – * Словники
7356 – * Відомі люди про Україну та мову

Що потрібно зробити.

1. Поповнити розважальний блок – лінгвістичні ігри, цитатник типу «страшне перо не в гусака», впроваджується проект лінгвістичних вікторин; – (поки тут тільки конкурс на найбільш милозвучне українське слово та імена – матеріал для роздумів для фонетистів, дослідників лінгвомелодики, а також психо- і соціолінгвістів (цікаво, що найчисельнішими мелодійними словами були лексеми на позначення малих істот, тварин, зменшено-пестливі назви))
2. Редагування тестів до Підручника, дизайн, тестування усієї системи;
3. Інтеграція довідкової служби з підручником тощо.

В. Могилевский, Т. Королева*
УДК 81'322

КРИТЕРИАЛЬНАЯ МОДЕЛЬ СООТНЕСЕННОСТИ ИНТЕНСИВНОСТИ СЕМАНТИЧЕСКОГО ПРИЗНАКА С КОЛИЧЕСТВЕННЫМ ПРОЯВЛЕНИЕМ ЕГО АКУСТИЧЕСКИХ КОРРЕЛЯТОВ

The criteria model, moved forward in the article, reflects the correlation between the degree of semantic indicator being present in either member within the functional and semantic field "degree of responsibility" and acoustic characteristics of this indicator.

Результаты исследования характеристик идеализированных моделей просодического оформления фраз, передающих модальные значения [2], позволили выявить механизмы инвариантности и вариативности интонационных единиц модальной сферы. К числу наиболее существенных могут быть отнесены следующие обобщения экспериментально наблюдаемых фактов:

1) для компонентов функционально-семантических полей модальности определенных уровней существует взаимосвязь с известным инвентарем релевантных акустических признаков;

2) эта связь не является однозначной, что и объясняет невысокую эффективность попыток создания фиксированного "словаря интонем" модальной (а равно и любой иной языковой) сферы;

3) соотнесенность "интонация–смысл" вполне определенно может интерпретироваться как критериальная, суть которой заключается в том, что данной семиологической единице соответствует не фиксированный набор интонационных коррелятов, а совокупность критериев, сформулированных в виде определенных правил (предписывающих, допускающих, запрещающих) и исключений.

Критериальная модель может быть реализована в различных формах в зависимости от цели моделирования (обучение иноязычной речи, создание программных комплексов для систем синтеза и распознавания речи, искусственный интеллект и т.д.). В частности, для реализации "речевых" программ на ЭВМ эффективной формой использования критериальной модели может служить математическая модель:

$\mathbf{P_s(\Phi_i, a_i) \Rightarrow S(g)}$,

где **S (g)** – семиологическая функция;

(g) – интенсивность семиологического признака;

P (Φ,a) – критериальная функция просодии;

$\mathbf{\Phi_i}$ – максимальные изменения i-того акустического признака;

$\mathbf{a_i}$ – интенсивность i-того акустического признака.

При таком подходе характер изоморфизма единиц семантического и интонационного уровней языка может быть вполне определенно интерпретирован, если конкретизирован вид соотнесенности **a =>g**.

Как показал анализ, для большинства функционально-семантических полей существует функциональная связь **a = f(g)**, причем аппроксимация ее в рамках методологии линейного программирования моделью $\mathbf{a = \Sigma\, b_k\, g^k}$ позволяет считать коэффициенты данной модели критериями изоморфизма рассматриваемых языковых единиц. Критериальная модель, даже в случаях снижения до минимума числа критериев в ней, дает возможность вполне адекватно описывать интонационную структуру каждого компонента функционально-семантического поля модальности, являясь в то же время эффективным инструментом обобщения механизма вариативности акустических коррелятов модальности в пределах всего градуального ряда. Указанное свойство делает критериальную модель чрезвычайно полезной для целей автоматического распознавания и синтезирования речи.

Убедительным подтверждением эффективности этого подхода явились результаты экспериментов по синтезу речевых стимулов на базе предложенных критериальных моделей с последующим аудиторским анализом синтезированных фраз. Иллюстрацией этого может служить опыт по синтезу фразы «*Він приїде завтра*" при передаче модальных значений градуального ряда "**степень ответственности**": легкомысленно,

* © *В. Могилевский, Т. Королева, 2006*

бездумно, безответственно, нейтрально, рассудительно, обдуманно, авторитетно, весомо. Синтез проводился в три этапа.

На первом этапе основой для моделирования синтезированной речи явились акустические характеристики, выявленные при анализе естественной речи дикторов.

В целом при градуировании модальных значений функционально-семантического поля «**степень ответственности**» результатами аудиторского анализа подтверждена наибольшая лингвистическая значимость следующих акустических характеристик:

- предударный слог – скорость нарастания частоты основного тона, длительность;
- первый ударный слог – скорость подъема и спада частоты

основного тона, уровень и интенсивность первой и второй формант (**F_1, F_2, AF_1** и **AF_2**);

- главноударный слог – величина перепада частот между второй и третьей формантами (**F_2 и F_3**), интенсивность третьей форманты (**AF_3**);
- заударный слог – частота основного тона, уровень и интенсивность первой и второй и третьей формант (**F_0, F_1, F_2 и F_3,AF_1, AF_2, AF_3**).

Эти выводы нашли отражение в количественных показателях критериальной модели просодии функционально-семантического поля "**степень ответственности**".

Автономное моделирование просодии каждого из компонентов функционально-семантического поля осуществлялись с помощью пакета прикладных программ "Обобщенный портрет" [1].

На втором этапе синтез осуществлялся путем модификации акустических характеристик сегментов исходного (нейтрального в модальном отношении) стимула по правилам, реализованным в критериальной модели исследуемого функционально-семантического поля.

На третьем этапе моделирования синтезированной речи отдельные параметры критериальной модели, сформированной на втором этапе, поочередно деформировались, что приводило к изменению акустических характеристик речевого стимула. Это позволяло посредством аудиторского анализа выявить степень релевантности варьируемых параметров в репрезентации данного модального значения.

Полученные на каждом из трех этапов образцы модально окрашенной синтезированной речи подвергались аудиторскому анализу, с целью распознавания передаваемого ими модального значения. Данные статистической обработки результатов аудиторского анализа естественных и синтезированных фраз, выражающих модальные значения "**степени ответственности**", представлены в таблице. Процент правильно идентифицированных фраз, синтезированных на каждом из этапов, сопоставлялся друг с другом и с вероятностью распознавания естественных речевых реализаций. При этом:

а) сопоставление стимулов I этапа моделирования синтезированной речи с естественными речевыми образцами фактически дает лишь оценку качества воспроизводимости естественной речи синтезатором, что связано со степенью совершенства программных и аппаратных средств используемого синтезатора;

б) сопоставление результатов слухового анализа стимулов I и II этапов позволяет оценить степень совершенства предложенной критериальной модели;

в) и, наконец, сопоставление Ш и П групп синтезированных стимулов позволяет сделать вывод о степени участия каждого из параметров критериальной модели в формировании просодии данного модального значения и представляет доказательства о лингвистической значимости акустических признаков, отобранных экспериментатором при построении критериальной модели.

Опыты III этапа по варьированию параметров критериальной модели, с целью выявления степени релевантности каждого из акустических признаков, представлены в данной таблице выборочно для иллюстрации влияния на степень распознаваемости модальных значений только лишь двух параметров – интенсивности второй форманты первого ударного слога синтагмы (**AF2**) и длительности первого ударного слога (**t**). При этом рассмотрены 4 варианта: *I вариант* – изменялось максимальное значение поправки к **AF2** первого ударного слога на 25%; *2 вариант* – изменялось максимальное значение поправки к **AF2** первого ударного слога на 50%; *3 вариант* – изменялось максимальное значение поправки к **длительности t** первого ударного слога на 25%; *4 вариант* – изменялось максимальное значение поправки к длительности t первого ударного слога на 50%.

Таблица. Распознавание аудиторами естественной и синтезированной речи (%)

Компоненты модально - семантического поля «степень ответсвенности»	Есте-ствен-ная речь N^e	Синтезированная речь					
		I этап N_I^c	II этап N_{II}^c	III этап			
				AF2		**t**	
				N_{III1}^c	N_{III2}^c	N_{III3}^c	N_{III4}^c
Легкомысленно	92	96	92	73	31	87	85
Бездумно	81	90	96	62	48	74	81
Безответственно	72	81	91	71	36	91	87
Нейтрально	92	91	87	52	28	83	80
Рассудительно	88	92	96	81	52	92	82
Обдуманно	80	86	87	65	28	90	88
Авторитетно	83	90	93	78	41	96	82
Весомо	90	94	87	62	33	85	80

Здесь $\mathbf{N^e}$ – абсолютная вероятность правильного распознавания естественных фраз; $\mathbf{N^c_i}$ – относительная вероятность правильного распознавания фраз, синтезированных на I, II и III этапах опыта, соответственно.

Анализ приведенных данных позволил сделать следующие выводы:

а) естественность звучания синтезированных стимулов оценивается аудиторами как вполне удовлетворительная;

б) степень распознаваемости модальных значений на основе предложенных критериальных моделей просодии – хорошая для крайних членов ряда и вполне удовлетворительная – для промежуточных;

в) из двух, отраженных в таблице критериев, интенсивность второй форманты **(AF2)** первого ударного слога обладает высокой дистинктивностью при градуировании модальных значений степени ответственности,и небольшое изменение этого акустического параметра существенно снижает процент адекватной распознаваемости синтезированных фраз (варианты 1 и 2 этапа Ш). Другой критерий – длительность первого ударного слога не может считаться лингвистически значимым признаком, т.к. существенное изменение этого показателя практически не влияет на степень распознаваемости (3 и 4 варианты этапа Ш). Учитывая, что этот показатель в естественных речевых реализациях модально окрашенных фраз все-таки присутствует, в критериальную модель просодии он может включаться в качестве факультативного.

Анализ всех исследованных в работе функционально-семантических полей модальности подтвердил эффективность применяемой критериальной модели, в частности, приемлемость гипотезы об изоморфизме, а во многих случаях – количественной адекватности показателей интенсивности семантического **(g_i)** и акустических признаков **(a_{ij}).**

Выполненный в настоящей работе статистический анализ позволяет ранжировать модальные значения по степени устойчивости их идеализированных критериальных моделей просодии, проследив определенную иерархию модальных полей по этому признаку. Компоненты модально-семантических полей V уровня располагаются в этом ряду следующим образом: "сильные" эмоциональные значения (типа восторженно, яростно, злобно); эмоционально-оценочные значения средней интенсивности (типа огорченно, нетерпеливо, раздраженно); наиболее яркие из собственно оценочных (типа уверенно, удивленно, с колебанием, протестуя) и побудительные модальные значения (типа приказ, мольба); завершают этот иерархический ряд наименее интенсивные и нейтральные эмоционально-оценочные, собственно оценочные и побудительные модальные значения (типа безразлично, неодобрительно, советуя).

Основным достоинством критериальной модели является ее эффективность не только при описании зависимости между степенью выраженности семантического признака каждого члена функционально-семантического поля и акустическими характеристиками этого признака, но и при обобщении механизмов изменения акустических коррелятов модальных значений градуального ряда в целом.

Одним из важных достоинств предлагаемой идеализированной критериальной модели является возможность использования ее для конструирования реальных многофакторных критериальных моделей. С этой целью многофакторная модель должна формироваться как суперпозиция матрицы нейтрального контура фразы и матриц

критериальных моделей, соответствующих различным факторам – коммуникативным, интеллектуально-логическим, модально-оценочным.

Литература

1. Алгоритмы и программы восстановления зависимостей // Ред. В. И. Вапника.- М.: Наука,1984.- 815 с.
2. Королева Т.М. Интонация модальности в звучащей речи. – Киев: Высшая школа, 1989.- 147 с.

КОМП'ЮТЕРНА ЛЕКСИКОГРАФІЯ

Д. Калинчук*
Київський національний лінгвістичний університет (Київ)
УДК 81'374.33

ПРОБЛЕМИ ПЕРЕКЛАДУ ПРИ УКЛАДАННІ НАВЧАЛЬНОГО ІТАЛІЙСЬКО-УКРАЇНСЬКОГО СЛОВНИКА

Composing lerner's dictionary has many difficulties of translation. Hence, not only the lexicographical principles should be used but also the rules of translation.

У сучасному суспільстві дедалі більше зростає потреба у вивченні іноземних мов. Оскільки міжнародні зв'язки стрімко розвиваються, то й потреба у розумінні іноземних, особливо рідкісних мов, збільшується з року в рік. З огляду на ці обставини кількість людей, що вивчають іноземну мову, зростає, а, отже, зростає й попит на навчальну лексикографічну літературу. На сучасному етапі на ринку книжкової продукції пропонується широкий спектр різних підручників, самовчителів та словників, і звичайним споживачам, що не розуміються на методиках викладання іноземних мов дуже легко розгубитись у виборі.

Отже, на що в першу чергу звертає увагу звичайний покупець обираючи той чи інший підручник чи словник? Якщо йдеться про словник, то в першу чергу ми звертаємо увагу на кількість слів та на склад словникової статті. Та, нажаль, ми вважаємо, що чим більше слів в словнику – тим він кращий. Насправді це не завжди так.

На кожному рівні навчання ми вивчаємо та використовуємо певний обсяг лексики, який називається „лексичним мінімумом". Отже, у студента чи учня, на початковому етапі вивчення мови, при користуванні перекладним словником, виникатимуть певні проблеми, а саме труднощі щодо вибору того чи іншого значення слова, якщо воно багатозначне, чи якщо воно має різні морфологічні характеристики, або є омонімом.

Отже, яке рішення цієї проблеми можна запропонувати? Для уникнення цих проблем та спрощення навчального процесу необхідні навчальні словники. Таким чином виникають наступні питання :

Що таке навчальні словники?

В чому полягають відмінності навчальних словників від перекладних словників?

В чому полягають принципи укладання таких словників?

Відповідаючи на перше запитання варто ще раз нагадати, що на кожному етапі вивчення мови учні використовують певний обсяг лексики, який називається „лексичний мінімум". Візьмемо за приклад перший етап вивчення мови, а саме перший рік навчання, скажімо, у вищих навчальних закладах. За перший рік студент має вивчити приблизно дві – дві з половиною тисячі слів, а загалом при вивченні має користуватися приблизно трьома тисячами слів. Причому бажано не виходити за рамки цього лексичного мінімуму, щоб не перенавантажувати студента. На першому етапі навчання студент ще не володіє мовою достатньо для того, щоб відчувати її та орієнтуватися при виборі одного із значень полісемічного слова. Якщо при опрацюванні текстів чи при перекладі учень користується словником, який наводить максимум значень, то він може розгубитись при виборі потрібного йому значення. На відміну від перекладного навчальний словник має

* © *Д.Калинчук, 2006*

наводити лише ті значення, які є необхідними для певного рівня навчання. До того ж словникова стаття має бути не перевантаженою прикладами сполучуваності слів.

Враховуючи вищесказане можна дати наступне визначення: навчальний словник – це словник, реєстр якого складається з найбільш уживаних слів певного етапу вивчення мови. До кожного етапу вивчення мови має бути свій навчальний словник, реєстр та структура словникової статті якого мають враховувати всі особливості навчальної програми даного рівня навчання. Щодо відмінностей між навчальним та перекладним словником то вони полягають в наступному:

Перекладні словники мають за мету навести максимальну кількість значень слова та прикладів сполучуваності слів, щоб користувач мав можливість сформувати повне уявлення про морфологічні та семантичні якості цього слова. Такі словники підходять для тих, хто має достатній обсяг знань, щоб орієнтуватися при виборі значення полісемічного слова.

Навчальні ж словники першого рівня мають за мету ознайомити користувача лише з найголовнішими, або скажімо з найбільш уживаними словами та їх значеннями. Словники другого рівня наводитимуть більше слів, значень та прикладів. Словники третього рівня матимуть ще більший реєстр та ще більше значень та ілюстрацій до кожного значення слова, і так далі. Кожен рівень навчального словника міститиме лексику відповідну до рівня вивчення мови. Наприклад, словник п'ятого рівня матиме в реєстрі велику кількість суспільно-політичної, економічної лексики та технічних термінів і до того ж в словникових статтях буде наведено якомога більше значень слова.

При укладанні італійсько-українського та українсько-італійського словника автори керувалися такими принципами:

1) Реєстр слів відбирається за аналізом підручників, навчальних програм та частотних словників. Для першого рівня беруться дві з половиною – три тисячі найчастотніших слів. Для другого рівня буде відібрано від чотирьох до шести тисяч найбільш уживаних слів.

Для першого рівня буде достатньо декілька найбільш уживаних значень, з кожним рівнем зростатиме й кількість значень та прикладів. Всі приклади мають бути наведені у всіх словникових статтях до слів, які використані у цих прикладах. Розглянемо такий приклад: „Ми повернемося дуже швидко." У цьому реченні використані наступні слова: „ми", „повернемося", „дуже", „швидко". Це означає, що речення „Ми повернемося дуже швидко." має бути наведене як приклад до займенника „ми", дієслова „повертатись", прислівників „дуже" та „швидко".

При створенні української частини укладач не добирає реєстр слів згідно з їх частотністю в українській мові, а бере кожне українське слово, що було наведено як варіант перекладу італійського реєстрового слова. Враховуючи те, що найчастотніші слова, як звичай, то в українській частині реєстрових слів буде в декілька разів більше.

Проблеми перекладу можна умовно поділити на дві групи:

1. Проблеми „багатофункціональності" італійської лексики.
2. Проблеми, що стосуються аспекту перекладності та неперекладності.

Що саме називається „багатофункціональністю" лексики? Лексична система італійської мови дуже багата, її можливості є певною мірою „неосяжними". Перш за все варто зазначити, що одне й те саме слово в італійській мові може одночасно належати до різних граматичних категорій. Як, наприклад слово **tornare**:

tornare *vi вертатися: Torneremo molto presto. = Ми повернемося дуже швидко.*

tornare *vt {віддавати щось комусь} повертати: Gli ho prestato la macchina e non me l'ha ancora tornata. = Я позичив йому машину, а він мені її ще й досі не повернув.*

Залежно від того чи є дієслово перехідним чи неперехідним воно буде використовуватись з прислівником або без нього, також від цього фактора залежить те, з яким допоміжним дієсловом воно буде використовуватись. Крім того категорія перехідності та неперехідності має значний вплив на семантичне значення слова. Це є абсолютно зрозумілим для того хто вивчає італійську мову вже не перший рік, а от для початківців це може бути або ще невідомим, або ж не досить зрозумілим. У такому випадку завданням лексикографа є створення якомога більш прозорої словарної статті.

Доволі-таки часто зустрічаються й омоніми. Деякі з них можуть бути одночасно декількома частинами мови. Як, наприклад, слово **quanto**:

quanto[1] *agg скільки: Quanta gente! = Скільки людей!; Quanti quaderni hai? = Скільки у тебе зошитів?; Quanti amici hai chiamato? = Скільки друзів ти покликав?*

quanto[2] *pron 1. скільки: Quanto Le devo? = Скільки я Вам винен?; Quanto chiede per questo libro? = Скільки він просить за цю книгу?; Devo comprare del pane, ma non so quanto mi serve. = Я повинен купити хліба, але не знаю скільки мені потрібно. 2. все що: Ho fatto quanto potevo. = Я зробив все, що міг; Voglio essere sincero e quindi ti dico quanto penso. = Я хочу бути відвертим, а отже я скажу тобі все, що думаю.*

quanto[3] *avv настільки; як: Quanto sono contenta di vederti! = Як я рада бачити тебе!; Nemmeno posso immaginare quanto è difficile. = Навіть не можу уявити наскільки це важко.*

Таке часто трапляється зі службовими частинами мови, до того ж в італійській мові прикметники та займенники, а іноді й прислівники, мають ідентичні форми і розрізняються лише за функціями, які виконують в реченні. Якщо ж слово є багатозначним, а такими є більшість найчастотніших слів слів, і всі значення є доволі поширеними, то доводиться наводити усі варіанти перекладу. Наприклад:

provare *vt 1. пробувати: provare un piatto = спробувати страву; rovare di essere sereno = спробувати бути спокійним. 2. приміряти: provare un vestito = приміряти сукню/костюм; provare le scarpe = приміряти черевики; 3. випробовувати: provare una macchina prima di comprarla = випробувати машину перш ніж купити її. 4. відчувати: provare amore (odio) = відчувати кохання (ненависть)/кохати (ненавидіти).*

Стосовно аспекту перекладу при укладанні словників, то тут автор має враховувати, що скоріш за все, користувач не знає правил та принципів перекладу, а особливо вирішення проблеми перекладності та неперекладності. Якщо ж учень побачить невідповідність структури речення перекладу до речення оригіналу він може бути збентеженим. Таким чином, укладач словника має знайти настільки коректний еквівалент, наскільки це можливо, оскільки від нього залежить не просто якість перекладу та його адекватність, а розуміння учнем чи студентом, який в першу чергу має зрозуміти та відчути мову.

Розглянемо ретельніше, які саме розбіжності в мовах можуть створювати вищезазначені проблеми. По-перше варто зазначити, що на відміну від української мови, що являється синтетичною, італійська мова – аналітична, і тому, наприклад, не всі прикметники з української мови будуть перекладатись на італійську одним словом. З іншого боку, враховуючи те, що більшість прикметників в італійській мови виражені дієприкметником минулого часу, а в українській мові, як відомо, не від усіх дієслів можна утворити дієприкметник, то не завжди італійський прикметник буде перекладено одним словом на українську. Розглянемо такий приклад:

campagna *f село, сільська місцевість: abitare/vivere in campagna = жити в селі/за містом; andare in campagna = поїхати в село/за місто; di campagna = сільський*

Нас цікавить один з прикладів наведених ло слова „campagna", тобто, село, а саме прийменниковий зворот „di campagna", де "di" – це прийменник, а „campagna" – іменник. У даному випадку зворот перекладається одним словом, до того ж прикметником. Враховуючи те, що слово „campagna" – це іменник і варіанти його перекладу також є іменниками, а в українській мові прикметник „сільський" є незалежною частиною мови, автор має зробити окрему статтю для цього слова. Тому, що користувач шукатиме саме прикметник „сільський", а не іменник „село" і як приклад до нього „сільський":

село ім ***campagna:*** *жити в селі/за містом = abitare/vivere in campagna; поїхати в село/за місто = andare in campagna; сільський f* ***di campagna;*** *сільська місцевість =* ***campagna***

Наведемо інший приклад: ***pomeriggio*** *m час після обіду: domani pomeriggio = завтра після обіду.*

У цьому випадку одне слово перекладається цілим словосполученням. При укладанні українсько-італійського словника автор має виділити головне слово в словосполученні та створити статтю для нього, тобто це слово „час", але також може бути й слово „обід". Словосполучення, що є варіантом перекладу італійського слова має не загубитись, і тому ми використаємо його як приклад до статті до слова „обід". Ми не можемо вказати цей приклад у статті до слова „час" оскільки при перекладі словосполучення „час після обіду" слово „час" опускається.

обід *ім* ***pranzo:*** *час обідати = è ora di pranzo; запросити на обід = invitare a pranzo завтра після обіду = domani pomeriggio*

Таким чином, існує чимало конструкцій, що не мають адекватних відповідників у мові перекладу. До того ж не всі частини мови мають ідентичні форми в двох різних

мовах, а з вищенаведених прикладів стає зрозумілим, що така, на перший погляд, лексикографічна робота, як укладання словника дуже тісно переплітається з основами та правилами перекладу.

Література

1. Зорько Г. Ф., Майзель Б. Н., Скворцова Н. А. Большой русско-итальянский словарь. 3-е издание стереотипное. – М.: Издательство Русский Язык, 1999. – 1018 с.
2. Ковалев В. Ф. Итальянско-русский и Русско-итальянский словарь. 2-е издание, стереотипное. – М.: Русский Язык Медиа, 2005. – Х, 1077 с.
3. Корунець І. В. Теорія і практика перекладу (аспектний переклад). Підручник. – Вінниця: Нова Книга, 2003. – 448 с.
4. Перебийніс В. І. Лексикографічне забезпечення навчального процесу з іноземної мови // Вісник Київського лінгвістичного університету. Київ, 1999. – Серія Філологія, Том 2, №1. – 233 с.

О. Лазарєва*

УДК 81’374.33

ПРО МОЖЛИВІСТЬ АВТОМАТИЧНОГО УКЛАДАННЯ ТЕРМІНОЛОГІЧНИХ СЛОВНИКІВ ЗА ФАХОВИМИ ТЕКСТАМИ

The paper suggests an approach to creation of the vocabulary of terminological dictionaries. Word combinations of definite grammar structures being automatically extracted from the texts of some domain and arranged in a frequency list may constitute the basis for a terminological dictionary. The obtained vocabulary will be excessive and should be edited by an expert.

Термінологічний вибух як одна з проблем інформаційного сьогодення викликав необхідність пошуку шляхів інтенсифікації та інтелектуалізації процесів укладання та оновлення словників різного призначення. Це й паперові фахові та тлумачні словники, й словники, що використовуються в автоматизованих інформаційних системах (класифікатори, рубрикатори, інформаційно-пошукові тезауруси). Традиційно такі словники складалися фахівцями з предметної галузі, можливо у співпраці з лінгвістами, на основі аналізу цілої колекції документів з заданої тематики. І цілком очевидно, що сам процес накопичення масиву лексичних одиниць був і лишається дуже трудомісткою операцією. Крім того, зважаючи на стрімкий розвиток практично всіх галузей науки й техніки, з одного боку, і доволі довгий процес підготовки чи модернізації словників, з іншого, такі словники завжди відстають за складом лексики від поточного стану галузі, тобто не відповідають вимогам повноти.

Всі ці проблеми склали підставу для пошуку методики автоматизації первинного накопичення бази термінологічних одиниць окремої предметної галузі. Основоположним принципом такої методики є припущення (і наша глибока впевненість), що необхідним і достатнім джерелом термінологічної лексики є фахові тексти, тобто достатньо великий масив таких текстів може практично повністю покривати лексичний і термінологічний склад певної тематичної підмови. Другим важливим принципом методики є припущення, що потенційно будь-яке іменне словосполучення може бути терміном або може містити термін. Отже, для вирішення поставленої задачі формування лексичного складу термінологічного словника необхідно розробити методи, які б дозволяли автоматизувати процес виділення з текстів таких іменних словосполучень, структура яких відповідає структурі термінологічних одиниць. Сформований таким чином словник, безперечно, буде надмірним і буде потребувати постредагування. Однак, трудовитрати, пов’язані з постредагуванням, будуть несумірно меншими за трудомісткість формування такого словника традиційними методами. Точне визначення обсягу корпусу текстів, необхідних для складання словників, представляє окрему задачу і не розглядається в цій роботі. Однак приблизну оцінку можна зробити емпірично, а саме – в процесі створення словника спостерігати за швидкістю його зростання. Якщо надходження нових слів або словосполучень практично припиняється, тобто відсоток нових надходжень відносно, наприклад, загальної кількості слів у тексті складає менше деякого досить малого значення, то поповнення словника можна припинити.

Перш за все визначимось, які словосполучення ми будемо шукати в текстах. Дослідники термінології [4] виділяють досить багато синтаксичних структур термінів, які

* © *О. Лазарєва, 2006*

включають, насамперед, найбільш розповсюджену модель «одиничний іменник», а також дво-, три- або багатокомпонентні словосполучення, які, в тому числі можуть містити прийменники, прислівники і дієслова. Однак найбільш продуктивними [2] є субстантивні словосполучення з прикметником або дієприкметником (ад'єктивний тип): *програмне забезпечення, багатозначна суфіксальна морфема*; субстантивні словосполучення з іменником (або їх послідовністю) в родовому відмінку (атрибутивний тип): *моделювання дискретних схем, неповне опрощення основи слова*. Оцінки, надані в роботі [3], показують, що терміни іншої структури, наприклад, з прийменниками або сполучниками складають менш одного відсотка від загального числа термінів. Часто дослідники, що працюють над аналогічними задачами, зокрема [1], вводять обмеження на довжину словосполучення, що буде включатися до словника – не більше 3 слів. З одного боку вони мають рацію – довжина більшості термінів не перевищує трьох. Але існують і набагато довші терміни, наприклад: *багатокомпонентна дрібнодисперсна суха суміш, область керування програмними засобами*, й ігнорувати їх вважаємо недоцільним.

Схематично типи найбільш продуктивних термінологічних конструкцій можна представити наступним чином:

S_1:=[<ад'єктив>] ... <іменник> або [A_i] ... N;

S_2:=S_1 {S_1}^{p}... ,

де квадратні дужки містять необов'язковий елемент, а фігурні – обов'язковий.

Цілком очевидно, що не всі словосполучення такої структури є термінами: це й слова загального вжитку, й вільні словосполучення. Але що є критерієм для включення чи не включення виділеної лексичної одиниці до словника? Це, перш за все, її галузева приналежність, і ціла низка інших критеріїв, що визначають властивості терміна як одиниці терміносистеми. А це вже екстралінгвістична інформація, формалізувати яку на сьогодні не представляється можливим. Тому ми ставимо задачу виділення всіх слів і словосполучень певної структури за формальними ознаками, а потім пропонуємо надати фахівцеві можливість редагування цього списку.

Якщо проаналізувати розташування термінів у тексті з урахуванням їхньої структури, то виявляється, що термін (зокрема, структури типу S_1 або S_2) може бути як самостійною синтаксичною одиницею речення, так і входити до складу словосполучення такої ж структури. Тому пропонується поділення виявлених у тексті початкових словосполучень на компоненти і формування з них усіх можливих комбінацій – **похідних дескрипторів**, синтаксична структура яких також відповідає структурі терміна.

Похідні дескриптори для конструкції S_1 формуються шляхом генерації усіх можливих комбінацій ад'єктивів з іменником. При цьому свідомо виключаються інвертовані ланцюжки типу $A_i+_kA_iN$. Це пояснюється тим, що порядок слідування ад'єктивів носить не випадковий характер, а зумовлений семантичними відношеннями актантів. Отже, наприклад, з ланцюжка A_1A_2N буде сформоване 3 похідних дескриптори: N; A_1N; A_2N, які разом з початковим будуть занесені до словника.

Для атрибутивних словосполучень (тип S_2) процес генерації дескрипторів має дещо інший характер, виходячи з того, що відношення між компонентами цієї конструкції є лінійними, тобто напрямок синтаксичного зв'язку спрямований від попереднього компонента до наступного без стрибків: $S_1 \rightarrow S_1 \rightarrow ... \rightarrow S_1$. Генерація полягає у формуванні всіх допустимих лінійних ланцюжків довжиною від 1 до n, де n – число ад'єктивних компонентів атрибутивної конструкції. Отже з початкового словосполучення цього типу, наприклад *A1N1+A2N2+A3N3*, буде сформоване 5 похідних одиниць: *A1N1; A2N2; A3N3; A1N1+A2N2; A2N2+A3N3*, і вони теж будуть занесені до словника. Крім того, за правилами обробки ад'єктивних словосполучень до словника додатково будуть включені елементи *N1; N2; N3; N1+A2N2; N2+A3N3*.

Отримані в результаті описаних перетворень словосполучення додатково піддаються процедурі інвертування, тобто розташування елементів словосполучення таким чином, щоб головне слово, тобто перший іменник, опинилось на першій позиції. Це дає можливість при сортуванні кінцевого списку словникових одиниць за алфавітом отримати цікаву картину лексико-семантичних гнізд, об'єднаних загальним головним словом.

Крім того, для запобігання повторення одних і тих те слів і словосполучень, які в текстах зустрічаються у різних відмінково-числових формах, дескриптори потрібно нормалізувати, тобто привести до початкової форми головний іменник і узгоджені з ним ад'єктиви.

Важливим показником при відборі дескрипторів у тематичні термінологічні словники є показник частоти вживання даного слова чи словосполучення. Цілком очевидно, що ймовірність приналежності деякого дескриптора до складу термінів певної предметної галузі прямо пропорційна його частоті й тим вища, чим більшою є кількість його слів. Отже показник

K = (F-1)*Q,

де K – імовірнісний коефіцієнт, F – частота дескриптора, Q – кількість слів у словосполученні, можна використовувати як оціночний для прийняття рішення щодо включення певного слова чи словосполучення до словника. Сортування складеного словника за частотою, або за отриманим показником, незважаючи на уявну простоту цього методу, дає непогану вершину списку на достатньо великих обсягах текстів. Звичайно, найбільш частотними будуть однослівні дескриптори, попри все – загальновживані слова і загальнонаукові слова типу *система, метод* тощо. Але ж ми вже зазначали, що дана методика не враховує ані семантичні, ані прагматичні властивості слів, а використовує лише формальні методи для відбору претендентів на включення до словника, а вирішальне слово завжди надається експертові.

Література

1. Антонов А.В. Информационно-поисковая система Galaktika-ZOOM с элементами анализа на гипермассивах информации // НТИ. Сер. 1. – 2001. – № 8. – С.12–21.
2. Лазарєва О.Я. Про деякі продуктивні моделі науково-технічних термінів. // Українська термінологія і сучасність: Зб.наук.праць. Вип.4. – К: КНЕУ, 2001. – С.185–188.
3. Лукашевич Н.В. Автоматизированное формирование информационно-поискового тезауруса по общественно-политической жизни России // НТИ. Сер.2. – 1995. – №3. – С. 21–24.
4. Суперанская А.В., Подольская Н.В., Васильева Н.В. Общая терминология. Вопросы теории. – М.: УРСС, 2003.

В. Шульгіна, к. філол. н.*

УДК 81'322.373

ПРОБЛЕМИ КОМП'ЮТЕРНОЇ ЛЕКСИКОГРАФІЇ У ГАЛУЗІ ІНФОРМАЦІЙНОЇ ДІЯЛЬНОСТІ

Метою статті є аналіз можливостей створення електронного забезпечення лексикографічними виданнями такої сфери, як документознавство. Проблема забезпечення навчального процесу на факультетах документознавства необхідною літературою включає в себе і підготовку термінологічних словників цієї потрібної сфери. Термінологічний словник документознавчої галузі є першою спробою узагальнення термінологічних понять, необхідних в роботі документознавців.

Прикладна лінгвістика як один з новітніх сучасних напрямків розвитку мовознавства виокремилася з традиційної лінгвістики не так давно. Серед проблем, якими займаються дослідники в галузі прикладної лінгвістики, проблема комп'ютерної лексикографії посідає одне з чільних місць [1]. Це пов'язано, насамперед, з активним розвитком комп'ютерної техніки, з її використанням у різних сферах науки та народного господарства. Комп'ютерна лексикографія є складовою частиною лексикографії як розділу традиційного мовознавства. Тому досвід лінгвістів-лексикографів, силами яких створювалося словникове багатство будь-якої мови, є надійним фундаментом для подальшого його розвитку у межах комп'ютерної лексикографії. Важливу роль у розвитку сучасної комп'ютерної лексикографії відіграють і новітні лексикографічні розробки, які є безпосередньою реакцією на вимоги сьогодення.

Останнім часом українська лексикографія поповнилася цілою низкою нових термінологічних словників. Поява цих видань зумовлена у певній мірі оновленням змісту багатьох наукових сфер українського суспільства, яке переживає часи набуття і зміцнення незалежності нашої держави, часи переходу від соціалістичної економіки до ринкової. Економіко-соціальні зміни, що відбуваються в Україні, не могли не позначитися і на розвиткові такої наукової галузі, якою є документознавство. Адже сама підготовка фахівців-документознавців для України стала справою зовсім новою – до розпаду СРСР таких спеціалістів готували лише в Московському історико-архівному інституті. Зараз, з урахуванням потреби в грамотних документознавцях, яку відчувають як державні установи, так і кожна приватна фірма, факультети документознавства та

* © *В. Шульгіна, 2006*

інформаційної діяльності відкрито в багатьох вищих навчальних закладах України як державної, так і недержавної форми власності (Східноєвропейський університет економіки і менеджменту, Київський університет культури, Харківський університет культури, Гуманітарний інститут Національного авіаційного університету та ін.).

Тому метою даної статті є аналіз можливостей створення електронного забезпечення лексикографічними виданнями такої сфери, як документознавство. Проблема забезпечення навчального процесу на факультетах документознавства необхідною літературою включає в себе і підготовку термінологічних словників цієї потрібної сфери. Термінологічний словник документознавчої галузі є першою спробою узагальнення термінологічних понять, необхідних в роботі документознавців. Враховуючи необхідність сучасного документознавця мати міцні знання не лише у межах власне документознавства, нами було залучено до словника ще і терміни суміжних з документознавством дисциплін: діловодства, архівознавства, менеджменту, інформатики, комп'ютерного забезпечення діловодства, інформаційна діяльність та ін.

Терміни, представлені в словнику, подаються в алфавітному порядку з обов'язковим посиланням на джерело, де тлумачення цього терміна може бути подано в більш широкому викладі. Певні терміни у різних джерелах мають різне тлумачення. Такі поняття і в словнику подаються у кількох варіантах, що передбачає можливість студентові порівняти існуючі уявлення про документознавче явище, обрати найбільш оптимальне для конкретної виробничої ситуації. Список використаних при укладанні словника джерел подається в кінці словника.

Вважаємо, що поява термінологічного словника документознавчої галузі не лише допоможе в формуванні високопрофесійних фахівців з документознавства у межах вищих навчальних закладів, а й знадобиться в практичній роботі багатьом працівникам управлінської сфери. У цілому, вміння спеціаліста-документознавця користуватися словниками, в тому числі і термінологічним, вчасно знайти потрібну інформацію повинні розглядатися як одна з його найважливіших професійних якостей.

Термінологічний словник документознавчої сфери доцільно розглядати як перший крок для створення комп'ютерного тезауруса галузі документознавства. У такому тезаурусі матеріал, зібраний у словнику, має бути систематизованим за певними напрямками, його компоненти мають підпорядковуватися однин одному в залежності від тих ієрархічних відношень, у яких вони перебувають саме у терміносистемі документознавчої галузі.

Створення такого тезауруса відповідає тенденціям, які простежуються на сучасному етапі розвитку нашого суспільства, яке поступово набуває статусу суспільства інформаційного. Адже для сучасного розвитку проблем інформації та інформованості характерним є те, що на одне з перших місць виходить проблема удосконалення інформаційних технологій, комп'ютеризація, створення новітніх баз даних та нових електронних видань, які мають значні переваги перед виданнями паперовими.

До таких електронних видань можна віднести і такі одиниці комп'ютерної лексикографії, як електронні тезауруси, створення яких є нагальною потребою сьогодення. Під терміном "тезаурус" можна розуміти таку одиницю системології, складові елементи якої входять у певні відношення між собою. При такому розумінні тезауруса особливої актуальності набуває проблема дослідження як самих компонентів тезауруса, так і тих відношень, зв'язків, які між ними виникають.

Статтю тезауруса можна гіпотетично ототожнити з моделлю хімічної речовини з шкільного курсу органічної хімії, де речовина, що має свої фізичні і хімічні властивості, складається з окремих молекул різнорідних хімічних речовин, що в свою чергу мають власні фізичні і хімічні ознаки за рахунок своєї внутрішньої структури – атомів та їх складових частин.Усі компоненти складної речовини при цьому перебувають у складних відношеннях не лише з сусідніми компонентами, а й з найбільш віддаленими. Окрім цієї аналогії напрошується аналогія побудови структури Всесвіту, де його компоненти – планети та інші космічні тіла, перебувають у постійних специфічних зв'язках між собою. Тож основні принципи побудови тезауруса є природними і універсальними.

Електронний тезаурус документознавчої галузі, на нашу думку, повинен відповідати таким основним вимогам:

1. тезаурус має бути багатоаспектним електронним виданням, кожний аспект якого повинен нести інформацію про концепти певної галузі;
2. послідовність словникових статей має бути на алфавітна, а тематична;

3. концепт – це узагальнене найменування для цілого класу явищ реальної дійсності;
4. вершиною словникової статті має бути артефакт;
5. кожна стаття концептного тезауруса має будуватися на принципі ієрархічності, в основу якого покладено гіпо-гіперонімічні, родо-видові відношення;
6. найдрібнішим елементом статті словника концептів може бути максимально конкретне найменування;
7. між артефактом і максимально конкретним найменуванням назви вибудовуються у межах послідовної шкали наростання деталізації за принципом: від найбільш загального до найбільш конкретного;
8. у межах комп'ютерного тезауруса концептів слід забезпечити можливість динамічних процесів між елементами однієї словникової статті як в бік узагальнення, так і в бік конкретизації.

Сфери застосування тезауруса можуть бути досить різноманітними.

Однією з основних сфер їх використання вважається інформаційно-пошукова робота, при якій відбувається " заміна лексичних одиниць тексту стандартизованими словами і висловлюваннями (дескрипторами) при індексуванні документів і використанні родо-видових і асоціативних зв'язків між дескрипторами при автоматизованому інформаційному пошуку документів" [7: 506 – 507].

Поряд з цим електронний тезаурус може активно використовуватися в навчальному процесі, в тому числі при застосуванні дистанційної форми навчання, адже кожний концепт у такому тезаурусі має супроводжуватися цілим „деревом" конкретизації його певних сторін та аспектів, що може сприйматися як свого роду схема, за якою можуть бути структуровані знання з певної проблеми.

Електронні тезауруси, що грунтуються на потужних можливостях сучасної комп'ютерної техніки для зберігання інформації, здатні містити значно більше інформації, ніж окремий паперовий словник специфічного спрямування. Йдеться про можливість реалізації критеріїв комплексності і вичерпності у подачі та зберіганні інформації, яка стосується кожного окремого поняття реальної дійсності.

Системи концептів, які функціонують у природному людському мовленні і являють собою систему понять у їх понятійному, концептуальному аспекті, є тим спільним, що об'єднує мову людини і мову машини. Тому можна вважати, що порушена проблема може бути цікавою для розробників діалогових систем "людина – ПЕОМ", а значить створення електронних тезаурусів концептів певних галузей можна розглядати лише як перший крок на шляху до створення штучного інтелекту.

Література

1. Богданов В.В., Бондарко Л.В., Буторов В.Д., Герд А.С. Прикладная лингвистика и теория языка // Структурная и прикладная лингвистика. – Вып 3 Л: ЛГУ, 1987. .
2. Клименко Н.Ф. Нові підходи до укладання *комп'ютерних словників*// Мовознавство, 1996, №4-5.
3. Морфологический анализ научного текста на ЭВМ. – Киев, 1989.
4. Грязнухіна Т.О. Лінгвістичні засади автоматичного синтаксичного аналізу// Мовознавство. – 1994. – №2.
5. Апресян Ю.Д. О языке толкований и семантических примитивах// Избран. Труды. В 2-х т. – М.: Школа «Языки русской культуры», 1995. – Т.2: Интегральное описание языка и системная лексикография.
6. Шайкевич А.Я. Тезаурус // Лингвистический энциклопедический словарь. – М.: Сов. энциклопедия, 1990.
7. Чурсин Н.Н. Популярная информатика. – К.: «Техника», 1982.

В. Перебийніс, д.філол. н., ***С. Назаров,*** к. філол. н., ***Т. Бобкова*** к. філол. н.*
Київський національний лінгвістичний університет (Київ)
УДК 81'374.33

КОМПЛЕКС ЕЛЕКТРОННИХ НАВЧАЛЬНИХ СЛОВНИКІВ „ГЛОСА"

Learner's dictionaries are intended for students and teachers of a foreign language and serve the purpose of selecting and activizing vocabulary to be studied. Lexical minimums are aimed at selecting most frequent words to be taught at different stages of studying a foreign

* *© В.Перебийніс, С.Назаров, Т.Бобкова, 2006*

language. They are to help teachers and authors of text-books in determining the syllabus and making the process of teaching more effective.

Серед усіх словників, що використовуються в процесі навчання іноземній мові, в окрему групу виділяють навчальні словники. На відміну від інших типів словників основне призначення навчальних словників полягає у відборі й активізації лексичного матеріалу. Поряд з довідковою літературою навчальні словники є основою лексикографічного забезпечення навчального процесу. За формою подання матеріалу навчальні словники бувають трьох видів: друковані, електронні й комбіновані – з додатком у вигляді CD. Прикладом алфавітних електронних словників може слугувати створений у лабораторії комп'ютерної лінгвістики КНЛУ комплекс електронних навчальних словників „ГЛОСА", який включає шість словників для навчання англійської, італійської та японської мові.

Як свідчить практика укладання навчальних словників, на початкових етапах основна увага укладачів концентрується на відборі лексичного мінімуму для різних рівнів навчання іноземної мови. При цьому в основу відбору лексики покладається, насамперед, принцип тематичний або такі принципи, як здатність слова виконувати службові функції, стилістична нейтральність, сполучуваність слова, його частотні характеристики та ін. Таким чином, на початкових етапах визначення лексичного мінімуму залежить від того, яким принципом керується укладач при відборі лексичного матеріалу. Так були укладені словники-*довідники лексичного мінімуму* [7; 8; 9]. Значні відмінності в реєстрах цих словників говорять про необхідність наукового обґрунтування визначення лексичного мінімуму на основі максимально формалізованих, об'єктивних критеріїв, що не залежать від суб'єктивних оцінок укладача.

Досвід укладання навчальних словників комплексу „ГЛОСА" показує, що найбільш доцільним є відбір лексичного матеріалу на основі аналізу навчальних програм, підручників і частотних словників, укладених з навчальною метою [3; 5; 6]. Критерієм відповідності реєстру навчальній меті можна керуватися й при відборі значень багатозначного слова, які включаються в лексичний мінімум.

Відомо, що ефективність сприйняття лексичного матеріалу, особливо на початкових етапах навчання іноземної мови, залежить від того, наскільки навчальний словник відповідає певним методичним і дидактичним вимогам. При укладанні словників комплексу „ГЛОСА" автори керувалися такими принципами побудови навчальних словників [4: 15–16]:

1. Реєстр навчального словника першого рівня має включати невелику кількість слів. Так, реєстр словника для початкового етапу навчання нараховує 1, 5–3 тисячі лексичних одиниць. Англо-український словник містить 3145 статей, італійсько-український – 1500 і японсько-український навчальний словник – понад 1800. Основу реєстрів навчальних словників становить лексичний мінімум, відібраний відповідно до навчальних програми першого етапу навчання.

2. Навчальний словник має містити рекомендації стосовно послідовності вивчення лексики: на якому етапі навчання доцільніше вивчати певне слово. Відповідно до цієї вимоги створюють або серію градуйованих навчальних словників (окремий словник для кожного етапу навчання), або один словник, що охоплює весь курс навчання [2; 10]. У другому випадку слово супроводжується спеціальною позначкою щодо терміну вивчення лексики. Укладачі комплексу навчальних словників „ГЛОСА" орієнтувалися на користувача, який не має досвіду роботи зі словниками, тому найбільш доцільним виявилося створення серії концентрично організованих словників. У цьому випадку реєстр словника наступного етапу навчання включає лексику попереднього, а його розширення визначається вимогами навчальних програм і частотою вживання слів. Таким чином, забезпечується концентрація уваги користувача на засвоєнні найбільш частотної лексики при доступності користування словником. Крім того, визначення конкретного адресата навчального словника дає можливість оптимально організувати словникову статтю без перевантаження її зайвим лексичним матеріалом.

3. Словникова стаття навчального словника повинна бути короткою, добре структурованою, прозорою. На першому етапі навчання іноземної мови не доцільно подавати всі існуючі значення слова. У навчальних словниках комплексу „ГЛОСА" поступово з поглибленням вивчення мови словникова стаття ускладнюється за рахунок інших, менш частотних значень слова, прикладів, граматичного коментарю, ідіом і т. д.

4. Словникова стаття навчального словника має містити не тільки переклад слова, а й показувати його вживання в певному лексичному оточенні. Сполучуваність кожного слова повинна бути представлена найбільш частотними словосполученнями, в яких реалізуються відібрані для лексичного мінімуму значення даного слова. Так, англо-український навчальний словник комплексу „ГЛОСА" містить понад 10600 прикладів, італійсько-український – понад 2000, а японсько-український – понад 3600 прикладів. Приклади таких словосполучень сприяють не тільки засвоєнню значень слова, але й формують навички перекладу зв'язного тексту. Причому приклади-речення включалися в словникову статтю тільки в тому випадку, коли переклад словосполучень не розкриває значення слова й особливостей його функціонування в мовленні. Однак на початковому етапі навчання іноземній мові кількість прикладів-речень варто обмежити. Основну увагу користувача доцільно сфокусувати на найбільш частотних моделях словосполучень.

5. Як приклади-словосполучення, так й ілюстративні речення повинні включати тільки реєстрові слова, щоб уникнути неконтрольованого розширення реєстру навчального словника. Відомо, що системне знання краще несистемного, тому в аспекті подання інформації навчальні словники комплексу „ГЛОСА" являють собою замкнуту систему, кожний елемент якої можна витлумачити за допомогою інших елементів. Насамперед необхідно звернути увагу на відповідність цій вимозі навчальних словників для початкового етапу навчання, тому що їхній користувач не володіє необхідним лексичним мінімумом. Надалі тлумачення слова й словникова стаття в цілому можуть бути розширені за рахунок нових елементів.

6. Доцільним уявляється також включення в навчальний словник різного типу додатків. Причому ця вимога є обов'язковою відносно навчальних словників, видаваних в Англії й США. Можливості електронних словників комплексу „ГЛОСА" дозволяють представити у додатках найуживаніші суфікси й префікси англійської мови, а також географічні назви, похідні від них і власні імена.

7. Крім цього, на початкових етапах навчання іноземної мови ефективність сприйняття на слух і оволодіння новими артикуляційними рухами насамперед залежать від належної організації звукового супроводу електронного навчального словника. [1]. Звуковий супровід комплексу навчальних словників „ГЛОСА" демонструє нормативну вимову реєстрових слів англо-українського, італійсько-українського та японсько-українського словників у початковій формі. Причому під час створення звукового супроводу навчального англо-українського словника першого рівня укладачі свідомо відмовилися від представлення двох варіантів вимови (британського й американського) через необхідність формування у користувача образа нормативної англійської вимови.

8. У навчальних словниках, призначених для наступних етапів навчання іноземної мови, доцільними є граматичні й методичні коментарі. Включення коментарів, що містять рекомендації про правила вживання слова в мовленні, передбачено в англо-українському й українсько-англійському словнику другого рівня комплексу „ГЛОСА".

9. Кожен українсько-іноземний словник укладається на основі перекладів реєстрових слів і прикладів відповідного іноземно-українського навчального словника, тобто відбиває особливості слововживання в мовленні, а не систему базової мови, отже є ситуативним. Оскільки іноземні реєстрові слова можуть мати кілька перекладів, то кількість статей в українсько-іноземному словнику більша, ніж у відповідному іноземно-українському. Так, наприклад українсько-англійський навчальний словник містить понад 6000 статей.

10. В електронних навчальних словниках програмне забезпечення повинне бути простим, доступним для учнів. Вимоги до апаратного забезпечення, з огляду на умови навчання в середніх школах, не повинні бути дуже жорсткими. Єдине для всіх навчальних словників програмне забезпечення комплексу „ГЛОСА" дозволяє переглядати реєстр словника, статтю окремого реєстрового слова зі звуковим супроводом чи без нього, здійснювати пошук статті окремого слова і його вживання в інших статтях реєстрових слів словника.

Комплекс електронних навчальних словників „ГЛОСА", призначений для широкого кола користувачів: викладачів і вчителів англійської мови, укладачів навчальних посібників та підручників, а також усіх, хто вивчає англійську мову, буде в пригоді як для аудиторної роботи, так і для самостійного вивчення іноземної мови.

Література

1. Бобкова Т. Принципи створення звукового супроводу електронних навчальних словників // Українське мовознавство. – 2004. – № 32-33. – С. 175-178.
2. Загородний Е. С. 8 уровней английского языка. Словарь-самоучитель. – М.: Изд-во Эксмо, 2002.
3. Перебийніс В. І., Рукіна Е. П., Хідекель С. С. Англо-український навчальний словник-мінімум. – К.: Наукова думка, 1995.
4. Перебийніс В. І. Лексикографичне забезпечення навчального процесу з іноземної мови // Вісник Київського державного лінгвістичного університету. Серія „Філологія", 1999. – Т. 2. – № 1. – С. 12-19.
5. Перебийніс В. І., Рукіна Е. П., Тіменко Т. В., Хідекель С. С. Англо-український та українсько-англійський словник. – Харьків: Фоліо, 2001.
6. Перебийніс В. І., Рукіна Е. П., Хідекель С. С. Англо-український навчальний словник з методичними коментарями та граматичними таблицями. – К.: Вежа, 2002.
7. Суша Т. Н. и др. Лексический минимум по английскому языку как воторой специальности с учетом соответствий в немецком и французском языках. – Минск: Вышейшая школа, 1981.
8. Фалькович М. М. Лексический минимум по английскому языку (для языковых вузов) //Учебн. Пособие для студентов I-IV кк. ин-тов и ф-тов иностр. Языков. – Изд. 2-е. – М.: Высшая школа, 1964.
9. Хидекель С. С. Гроссман Е. Я., Шарапова А. Н. Лексический минимум по английскому языку. – В 2-х ч. – М.: Высшая школа, 1965.
10. Macmillan Essential Dictionary for Learners of English. + CD-ROM, 2003.

Б. Шуневич, к. філол. н., ***О. Паплик****
Національний університет „Львівська політехніка" (Львів)
УДК 81'374.322

ПРОЕКТ АНГЛІЙСЬКО-УКРАЇНСЬКОГО КОМП'ЮТЕРНОГО СЛОВНИКА З РОБОТОТЕХНІКИ

У статті обґрунтовано актуальність і принципи створення англійсько-українського комп'ютерного словника з робототехніки, проаналізовано лексикографічний досвід у цій галузі, описано способи користування таким словником.

За останні десятиріччя виник цілий ряд нових галузей техніки, зокрема космонавтика, робототехніка та інші. Це зумовило значний ріст термінологічної лексики, а саме: виникнення нових термінів, розширення значень вже наявних, запозичення іншомовних термінів, головним чином з англійської мови. Усі ці терміни повинні бути належним чином зафіксовані в базах даних, словниках.

Донедавна взагалі не існувало фахової літератури з робототехніки українською мовою, оскільки більшість її публікувалася англійською, а в Радянському Союзі – російською мовами.

Останнім часом модернізується українська промисловість, а отже виникає потреба володіти найновішою закордонною інформацією з різних галузей науки й техніки, і в тому числі з робототехніки, а також представляти свій досвід закордонним науковцям.

Укладені в 1980 р. англійсько-російський і англійсько-український частотні словники з робототехніки [7: 117–126], [4: 157–159] базувалися на науково-технічній літературі 1975–1979 рр. Для цього було використано 211 джерел наукової літератури: журнальні статті, матеріали конференцій і патентні описи. В результаті було отримано список термінів обсягом 7 000 лексичних одиниць.

З 1981 року ці словники доповнювалися термінами (до 10 000 л.о.) із опрацьованої англомовної літератури [8: 9–10], а також закордонних [9] і виданих в СРСР словників [1] з цієї тематики, обговорювалися на різного рівня конференціях із вітчизняними та закордонними спеціалістами в галузі робототехніки [3: 111–112], [4: 157–159] і детально досліджувалися [6].

У 1987 році вийшов друком словник найчастотніших загальнотехнічних і спеціальних термінів з робототехніки обсягом 2000 лексичних одиниць [5].

Останнім часом „Англійсько-український словник з робототехніки" доповнюється нами за рахунок опрацьованої сучасної науково-технічної літератури з цієї тематики і за допомогою програмного забезпечення укладається як комп'ютерний словник.

* © *Б. Шуневич, О. Паплик, 2006*

Для укладання словника використовується програмне забезпечення з вільним кодом (open source), а саме PolyDic, версія 1.0, яке укладено під керівництвом Романа Мисака (Національний університет "Львівська політехніка") [2].

Система укладання та перегляду електронних словників PolyDic складається з двох програм: PolyDic Editor – для набирання, форматування та редагування словникових баз даних та PolyDic Viewer – для перегляду електронних словників. Перша версія системи PolyDic є електронним варіантом двомовного або тлумачного термінологічного словника. Програма дає змогу супроводжувати словник з фіксованою базою слів і словосполучень та з одним напрямком перекладу. Усі терміни можна уводити із наголосами, що є суттєвим для кращого розуміння перекладу, змісту термінів та їх усного відтворення. Для цього в програму введено спеціальний шрифт із наголосами *PolyDic Sans Serif*.

Центральну частину основного вікна програми PolyDic Editor розділено на дві частини: ліва – перелік усіх термінів словника, права – відповідні терміностатті (з тлумаченням, перекладом тощо).

Пошук терміна у PolyDic Editor відбувається за першими уведеними літерами в інтерактивному режимі у вікні пошуку. Програма запам'ятовує історію (почерговість) переглянутих статей, в якій зберігається послідовність до десяти переглянутих статей. Особливістю PolyDic та її істотною перевагою над іншими електронним словниками є механізм фільтрів, який можна застосувати до термінів, тексту статей, або до них обох одночасно. Наприклад, можна залишити видимими слова, які починаються на "абр", або слова, що мають закінчення "ан". У статтях, у разі потреби, можна встановити зв'язки з іншими статтями, які можна переглянути просто натиснувши курсором миші по відмітці зв'язку. Переклади або тлумачення можна супроводжувати короткими поясненнями, наприклад, щодо галузі застосування. Пояснення виконано у формі прихованих підказок. Щоб увести словникову базу, розроблено програмний модуль PolyDic Editor, який забезпечує основні інструменти для створення, наповнення та редагування словникових баз

У програмі PolyDic Editor передбачено можливість набору словникової бази частинами з подальшим злиттям цих частин в єдину базу. Ця функція корисна під час розподілення праці з уведення інформації в базу між різними операторами-користувачами.

З використання системи PolyDic 1.0 укладено електронні варіанти двох словників з термінографічної серії *СловоСвіт* "Російсько-український та українсько-російський словник термінів будівництва й архітектури" (№ 8) та "Російсько-український словник з інженерних технологій" (№ 9), випущено компакт-диски з цими словниками.

Словник призначений для наукових та інженерно-технічних працівників в галузі робототехніки, перекладачів, викладачів, студентів та учнів, а також студентів та учнів середніх навчальних спеціальних закладів.

Література

1. Петров А., Масловский Е. Англо-русский словарь по робототехнике. – М.: Рус. яз., 1989. – 427 с.
2. Система укладання та перегляду електронних словників PolyDic v.1.0 (www.lp.lviv.ua)
3. Шуневич Б. Англо-русский словарь по робототехнике. Пути его пополнения новой терминологией // Тез. докл. Всесоюзной н/т конф. «Робототехнические системы в текстильной и легкой промышленности», Ленинград, 1984. – С. 111–112.
4. Шуневич Б. Англо-український словник з робототехніки // Тези 1-ої міжн. конф. «Проблеми української науково-технічної термінології», Львів, 1992. – С. 157–159.
5. Шуневич Б. Методическая разработка и учебные задания для чтения текстов по робототехнике. ЛГПИ им. Герцена, Ленинград, 1987. – 70 с.
6. Шуневич Б. Структурные и функциональные характеристики английской терминологии по робототехнике. Дис. ... к.філол.н. – Львов, ЛДУ, 1991. – 532 с.
7. Шуневич Б. Частотный англо-русский словарь по робототехнике // Учен. зап. Тарт. ун-та. Тарту, 1985, вып. 711. – С. 117–126.
8. Шуневич Б., Тихомирова В. Рекомендуемый перевод с английского языка нових терминов // Экспресс-информация, серия «Иностранная техника и экономика средств связи». – М., 1987. – Вип. 14. – С. 9–10.
9. Robotertechnic: English-Deutsch-Französisch-Russisch. – Berlin: VEB Verlag Technik, 1986. – 273 s. (близько 7000 термінів)

О. Сірук*
Київський національний університет ім. Т. Шевченка (Київ)
УДК 161.2.81'374.322

КОМП'ЮТЕРНИЙ ТЕЗАУРУС ДІЄСЛІВ УКРАЇНСЬКОЇ МОВИ: ПРОБЛЕМИ ТА МЕТОДИКА УКЛАДАННЯ

У статті здійснено огляд наукової розробленості проблеми створення комп'ютерних тезаурусів, зокрема дієслів, з'язовано актуальність таких розвідок, визначено мету, етапи роботи, а також її значення з погляду комплексного підходу.

Одним з найважливіших завдань сучасної лексикографії є проектування таких словників, які б на рівні світових стандартів задовольняли велику потребу сучасної інформатизованої спільноти в систематизованій лінгвістичній інформації. З огляду на це **тезауруси** як словники, які не лише інвентаризують, а й систематизують лексичні одиниці у межах певної мовної підсистеми, потрапляють у поле підвищеної уваги фахівців. Рівень розвитку інформаційних технологій в Україні дозволяє, а потреби користувача вимагають зосередитися на розробленні саме **комп'ютерних тезаурусів** (КТ) різних типів: як загальномовних, так і вузькогалузевих термінологічних. Історія укладання ідеографічних словників має довгу традицію: однією з найдавніших писемних пам'яток тезаурусного типу є створений ще в II-III століттях до н. е. санскритський словник "Амара-коша". Серед наукових праць, актуальних і на сьогодні, найбільш відомі такі тезауруси: словник П. Роже для англійської мови, П. Буассьєра – для французької, Ф. Дорнзайфа – для німецької, Х. Касареса – для іспанської. Вагомий внесок у розбудову тезаурусної і дотичної до тезаурусної проблематики зробили Ш. Баллі, Л. В. Щерба, Н. Ю. Шведова, В. В. Морковкін, Ю. М. Караулов, Ю. Д. Апресян, О. С. Баранов, І. О. Мельчук, М. А. Кронгауз, М. Я. Гловінська, Л. Г. Бабенко. Серед українських дослідників можна відзначити роботи таких авторів, як Н. П. Дарчук [1], В. В. Дубічинський [2], А. Я. Середницька [3], О. Синиченко [4], Ж. П. Соколовська [9], Н. В. Сніжко та М. Д. Сніжко [8]. У комп'ютерній мережі Інтернет успішно функціонують і розвиваються тезаурус WordNet, багатомовна електронна лінгвістична база EuroWordNet, створені за аналогією до них бази GermaNet, BalkaNet, RusNet та багато інших. Співробітники лабораторії комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка працюють над укладанням таких комп'ютерних тезаурусів, як КТ дієслів, КТ іменників, Спеціалізований тезаурус з комп'ютерної ідеографії (онлайн-версія доступна для користувачів на сайті лінгвістичного порталу MOVA.info у розділі «Словники» http://www.mova.info/Page3.aspx?l1=188&vocid=1) тощо [5; 6; 7]. Успішна практика створення і використання КТ зумовила виникнення комп'ютерної лексикографії й поставила нові завдання її розбудови, зокрема створення **формалізованих методик конструювання** словників тезаурусного типу.

Проблеми. У зв'язку зі станом лексикографічних досліджень з тезаурусотворення, зокрема українських, терміносистема такої відносно «молодої» мовознавчої галузі, як комп'ютерна ідеографія, досі практично не розроблялася. Тому під час роботи над укладанням комп'ютерного тезауруса дієслів української мови виникла потреба чітко визначити і систематизувати терміни цієї лінгвістичної ділянки.

Під час аналізу тезаурусної продукції мережі Інтернет не завжди можна було отримати доступ безпосередньо до проектів у повному обсязі. У процесі роботи над дієслівною базою даних траплялися ситуації, коли досить важко було встановити типи відношень між лексико-семантичними варіантами: синоніми виявлялися надто далекими за значеннями і семантичними відтінками, тобто належали до різних груп семантичної спорідненості (наприклад, рід – вид, дія – початок дії); асоціативні відношення часто дублювали гіпонімію та синонімію тощо. Виникала потреба звертатися до основних засад теорії поля та її відображення у словниках, розглядати лексикографічні питання та супутні теоретичні проблеми перш за все таких розділів лінгвістики, як лексикологія і семантика, застосовувати різні види аналізу, що допомогло розв'язанню цих проблем, зокрема формалізованим способом. У процесі укладання бази даних постали також проблеми лінгво-методичного та технологічного характеру: спосіб відбору й

* *© О. Сірук, 2006*

упорядкування дієслівної лексики у синоптичній схемі, визначення та називання концептів ЛСП мовлення, відбір методів для аналізу наявної у тлумачному словнику лінгвістичної інформації про дієслівні лексеми, вибір методів побудови синоптичної схеми тезауруса, визначення особливостей лексикографування дієслів, зокрема похідних і непохідних, визначення мікро- і макроструктури комп'ютерного тезауруса, лексичного формування словникової статті тощо.

Як у процесі створення лінгвістичної бази даних для тезауруса дієслів, так і на етапі експерименту над художніми текстами Л. Костенко і В. Стуса, де тезаурус застосовувався як дослідна система, актуальною була проблема розмежування омонімії й полісемії вживаних у текстах дієслів ЛСП мовлення, що на сучасному рівні знань практично не піддається формалізації. Загальними і для розроблення дієслівного тезауруса, і для експериментальної частини роботи були пошуки шляхів оптимального представлення результатів, прийнятного як для звичайного користувача, так і для дослідника-лінгвіста.

Теоретичні постулати дослідження були трансформовані у низку **завдань,** послідовне виконання яких привело до поставленої мети. У підсумку вимальовалася чітка **послідовність** укладання КТ дієслів, описана у вигляді **чотирьох блоків. У підготовчому блоці** здійснюється *підготовка теоретичної бази, відбір матеріалу* (у нашому випадку – суцільна вибірка зі «Словника української мови» дієслів з семою мовлення кількістю близько двох тисяч одиниць) та *вибір концептів* (наприклад, ЛСП дієслів мовлення).

Основне опрацювання підготованого мовного матеріалу відбувається у **блоці аналізу. У**кладання бази даних (БД) комп'ютерного тезауруса дієслів здійснено за допомогою *комплексної процедури*, яка включає елементи морфемно-словотвірного аналізу лексем ЛСП, синтаксичного аналізу ілюстративного речення, синтаксичного аналізу тлумачення, компонентного аналізу тлумачення, ступеневої ідентифікації лексики. Результатом такого комплексного аналізу було *встановлення відношень між одиницями* (гіпонімія, синонімія, антонімія, деривативні та інші типи відношень). Послідовність розподілу ЛСВ дієслів за лексико-семантичними групами і лексико-семантичними полями було здійснено алгоритмічно представлено в комп'ютерному тезаурусі у вигляді блок-схем. Завершальним етапом укладання КТ дієслів є **блок синтезу,** у межах якого спроектовано мікро- і макрорівні комп'ютерного ідеографічного словника. У цьому блоці були виконані такі завдання, як *формування макроструктури КТ дієслів* (побудована синоптична схема тезауруса, спроектовані входи у словник); *визначення мікроструктури тезауруса* (розроблено словникові статті, висвітлено структуру міждієслівних відношень) та розроблені принципи *інтеграції дієслівної частини* КТ у комп'ютерний тезаурус української мови (на базі міжчастиномовних зв'язків дієслова).

Комп'ютерний тезаурус дієслів української мови виступає у **трьох** основних іпостасях: як теоретична логіко-семантична модель лексики; як багатопланова довідкова система та як інструмент для проведення лексико-семантичних досліджень. **Експериментальний блок** присвячений застосуванню КТ дієслів для порівняльного аналізу авторських художніх текстів Л. Костенко і В. Стуса за допомогою тезаурусного поля дієслів мовленнєвої діяльності в комплексі з іншими лінгвістичними програмами з представленням кількісних і якісних характеристик дієслівних складових означеного ЛСП. Використання тезауруса у поєднанні зі статистичними методами дало змогу помітити деякі лексичні особливості двох вищезгаданих поетів, які не могли бути виявлені простим зіставленням частотних характеристик лексем обох вибірок [5]. Подібні дослідження роблять внесок у вивчення авторського стилю письменників.

Результати дослідження свідчать, що методика аналізу **дієслівної лексики** суттєво **відрізняється** від аналізу інших частин мови, зокрема іменників як основних представників предметної лексики. Це пов'язано з референтним характером семантики дієслів. Саме тому аналіз дієслівних значень необхідно проводити перш за все на основі індуктивного методу, через уточнення окреслених дедуктивним способом концептів. Встановлено, що особливості дієслівної семантики зумовлюють також кількісні та якісні параметри словникової статті КТ дієслів. **Індуктивний підхід** показав, що можна деталізувати спільний для усіх дієслів мовлення концепт "говорити" шляхом виділення **чотирьох** концептів другого рівня, які уточнюють умови та способи перебігу процесу мовлення: «вимовляти певним чином», «висловлювати думку чи почуття», «повідомляти інформацію», «розмовляти між собою». Ці чотири основні способи мовленнєвої передачі інформації стали базовими для **синоптичної схеми ЛСП мовлення** (чотири

словосполучення *особливості вимови, висловлення думки, повідомлення інформації, обмін думками* представляють другий рівень ієрархізації у тезаурусі), а дієслова *вимовляти, висловлювати, повідомляти* і *розмовляти* з тотожними до концептів значеннями стали їхніми репрезентантами. Детальна інвентаризація диференційних лексико-семантичних ознак, які у значеннях дієслів мовлення уточнюють категоріально-лексичну ознаку "говорити", дозволила виділити низку диференційних ознак, з яких 42 стосуються способу вимови, 21 – висловлення певної думки чи почуття, 24 — способу повідомлення інформації, 4 — обміну інформацією. Поєднавши отримані диференційні ознаки з категоріальною ознакою "говорити", ми отримали зміст концептів третього рівня**,** на яких базуються словникові статті. Під час наступного етапу аналізу у межах цих двох груп відбувається подальша конкретизація синоптичної схеми як через співвіднесення узагальнюючих схем лексико-семантичних груп, так і шляхом порівняння схем лексико-семантичних варіантів у межах підгруп і виділення диференційних сем глибших рівнів.

Розроблення методики укладання КТ дієслів сприяло **вирішенню дотичних лінгвістичних проблем.**

Для конструювання бази даних мовна інформація повинна бути **формалізовано представлена. П**ерш за все це стосується поданих тлумачним словником визначень. У процесі роботи суттєвою проблемою стала відсутність їх чіткої стандартизованої структури (відсутність номерів значень ЛСВ, замкненого кола тощо). Стандартування визначень тлумачного словника й укладання ідеографічного словника необхідно проводити паралельно, що дає можливість оптимізувати процес роботи.

Дослідження показало, що більшість семантичних зв'язків між дієсловами характеризуються **підрядністю**, сурядність, тим більше у чистому вигляді, трапляється насправді рідше, ніж прийнято вважати. Так, значна частина слів, які у синонімічних та фразеологічних словниках визначаються як синоніми, при уважнішому аналізі виявилися поєднаними іншими типами зв’язків; зокрема, родо-видовим. Стилістичні та семантико-стилістичні синоніми визначені у КТ як родо-видові пари з семами стилістичного чи емоційного забарвлення. Було конкретизовано поняття родо-видових, синонімічних й антонімічних відношень, які стали основними структуруючими чинниками семантичного поля мовлення.

Результати аналізу дієслівних ЛСВ дають підстави стверджувати, що лексема у певному значенні може належати до **декількох полів** за своєю природою, маючи при цьому різний статус, перебуваючи на різних рівнях віддаленості від центру ЛСП. Наприклад, і до ЛСП мовленнєвої діяльності (ядра), і до ЛСП емоційної характеристики мовлення (периферії) відносяться дієслова типу *огризатися “відповідати у різкій, грубій формі”*.

Одним з важливих етапів розроблення ідеографічного словника є вибір відповідних форм **назв** для його концептів. Оскільки визначальним при доборі назви концепту є намагання досягнути її максимальної однозначності і прозорості, чіткого розкриття змісту, найбільш коректною формою називання концептів комп’ютерного тезауруса дієслів мовленнєвої діяльності є словосполучення, у складі яких кожна семантична ознака передається окремою лексичною одиницею.

Запропонована методика дозволяє порівняно точно визначити місце дієслова у лексико-семантичній системі мови завдяки застосуванню основних засад методики компонентного та ступеневого аналізів ЛСВ. Таким чином, синоптична схема тезауруса розробляється не накладанням масиву лексики на аналізовані значення, а формується на основі аналізу лексико-семантичних значень, що дозволяє уникати суб’єктивних впливів розробника на результати дослідження. І, що особливо важливо для нашого дослідження, чітко визначена послідовність кроків дає можливість формалізувати роботу з укладання комп’ютерного тезауруса, допомагає знайти оптимальну форму представлення результатів та застосування їх у комплексі інших лінгвістичних продуктів.

Зонний спосіб є найбільш прийнятним для розташування мовної інформації у КТ. На підставі лексико-семантичних відношень дієслова було виділено 4 зони для дієслівного КТ (зони гіпонімії, синонімії, антонімії та відношення роду дії) та ще 3 зони, які базуються на міжчастиномовних зв’язках дієслова (зони субстантивів, атрибутивів та адвербативів), що важливо для інтеграції КТ дієслів у загальномовний КТ української мови. Кожний тип відношень формує відповідну зону, яка на екрані позначається своїм фоновим кольором. Загалом, за такого підходу можна говорити про конвергенцію синонімічного, антонімічного, тлумачного та ідеографічного словників, завдяки чому КТ дозволяє систематизувати різнобічну лінгвістичну інформацію. Зональна структура

допомогла уточнити за формальними показниками межу між центром і периферією ЛСП, що безпосередньо відображається на розміщенні ЛСВ дієслів у синоптичній схемі КТ (місце у межах одного чи декількох концептів; рівень ієрархізації; кількість заповнених зон у мікростатті), а також у вигляді зони додаткової семи у мікростатті КТ дієслів, де фіксуються ЛСВ з ядра інших лексико-семантичних полів, що містять у своєму складі диференційну сему мовлення. Як результат інтеграції КТ дієслів у КТ української мови спроектовано **загальну структуру комп'ютерного тезауруса**, що складається з **трьох** базових компонентів: синоптичної схеми, власне ідеографічної частини (мікро- і макростатті) та пошукової системи з функціями алфавітного та пермутаційного покажчиків.

Розроблення основних теоретико-методичних засад створення тезауруса дає можливість зробити внесок у дослідження таких аспектів мовознавства, як парадигматика й синтагматика, теорія поля, синонімія, антонімія і полісемія у лексичній семантиці, показує органічний взаємозв'язок синтаксичних і семантичних характеристик дієслова. Аналіз і систематизація наявних, а також конструювання нових електронних словників тезаурусного типу сприяють становленню поняттєвого апарату, методології та структури такого якісно нового підрозділу української лексикографії, як комп'ютерна.

Отримані результати, відповідно, можуть знайти відображення в лекціях, спецкурсах і спецсемінарах з проблем комп'ютерної ідеографії. Вони важливі для розвитку прикладної лінгвістики в цілому як логічне продовження її теоретико-методологічних засад, заснованих на глибокому симбіозі класичних лінгвістичних теорій і новітніх комп'ютерних технологій.

Література

1. Дарчук Н., Денисенко І., Сірук О., Сорокін В. Теоретичні питання моделювання ідеографічного тезауруса української мови // Українське мовознавство. Міжвідомчий науковий збірник. Вип. 24. — К.: Видавничо-поліграфічний центр "Київський університет", 2002. — С.107 – 118.
2. Дубичинский В. В. Искусство создания словарей конспекты по лексикографии / Харьковский политехнический ун-т. — Х., 1994. — 102 с.
3. Середницька А. Я. Мовна категоризація світу і її відображення в ідеографічному словнику // Семантика, синтактика і прагматика мовленнєвої діяльності. Матеріали Всеукраїнської наукової конференції. — Львів: Літопис, 1999. — С. 87 – 90.
4. Синиченко О. До проблеми створення словника понять української мови // Мовознавство: Доп. та повідомл. IV Міжнародного конгресу україністів / Відп. ред. В. Німчук. — К.: Пульсари, 2002. — С. 31 – 35.
5. Сирук Е., Сорокин В. Использование компьютерного тезауруса в статистическом анализе лексики художественных текстов (на материале украинского языка) // Texts and Contexts: the Movement of Language. Selected Papers. Vilniaus universiteto Kauno humanitarinis fakultetas. — Kaunas VU Press, 2005. — P. 601 – 608.
6. Сірук О. Два підходи до побудови комп'ютерного тезауруса дієслів української мови // Українське мовознавство. Міжвідомчий науковий збірник. Вип. 31. — К.: Видавничо-поліграфічний центр "Київський університет", 2004. — С.84 – 87.
7. Сірук О. Систематизація світових комп'ютерних тезаурусів як підґрунтя для укладання тезауруса української мови // Мовні і концептуальні картини Світу: Збірник наукових праць. — Вип. 16. Кн. 2. — К.: Видавничий Дім Дмитра Бураго, 2005. — С.189 – 193.
8. Сніжко Н. В., Сніжко М. Д. "Ідеографічний тезаурус" як інформаційно-довідкова система при вивченні закономірностей структурно-функціональної організації лексики // Мовознавство. — Київ, 1996. — № 4–5. — С. 23 – 28.
9. Соколовська Ж. П. Картина світу та ієрархія сем // Мовознавство. — Київ, 2002. — №6. — С.87 – 91.

І. Васильєва*
Інститут філології КНУ імені Т. Шевченка (Київ)
УДК 81'33

ВИКОРИСТАННЯ КОМП'ЮТЕРНОГО ТЕЗАУРУСА В ДОСЛІДЖЕННІ МОВИ ПОЕТІВ

In our research we propose the ways for using computer-based Ukrainian thesaurus and the frequency dictionary of Ukrainian poetry for creating poetry language picture. We consider lexico-semantical group "Human" in Ukrainian poetry discourse of 90-th.

* *© І.Васильєва, 2006*

Сьогодні, на щастя, уже не стоїть питання про доцільність застосування ЕОМ у лінгвістичних дослідженнях [2]. Використання комп'ютерних технологій у сучасній науці не лише полегшує роботу дослідників, оптимізує дослідницькі процеси, а й дозволяє ефективно використовувати отримані результати, головним чином як інструмент у дослідженнях суміжних галузей. Прикладами такої плідної "співпраці" у лінгвістиці можуть бути розвідки з галузі термінології, мови поетичних і художніх текстів, національних мовних картин світу тощо, проведені на базі комп'ютерних лексикографічних систем, різноманітних корпусів і комп'ютерних (частотних та ін.) словників [4; 1]. Запозичення інструментарію та даних лексикографічних систем має безліч переваг:

– дослідник отримує у своє розпорядження достовірну інформацію і готовий інструмент дослідження (для пошуку нового наукового результату йому, як правило, вже не потрібно винаходити велосипед);

– це, в свою чергу, позначається на економії ресурсів і часу й дозволяє досліднику оптимізувати роботу над об'єктом, зосередившись на головному;

– залежно від структури й логічної впорядкованості комп'ютерної лексикографічної бази даних або словника можна досліджувати декілька параметрів одночасно (наприклад, частотні, лексико-семантичні, граматичні й стилістичні характеристики тощо).

У цій доповіді ми розглянемо можливість використання комп'ютерного тезауруса української мови, а власне, іменників лексико-семантичної групи "Людина", для дослідження мови поетів. На основі отриманих результатів спробуємо побудувати фрагмент мовної картини світу українських поетів ХХ століття. Є чимало спроб створення тих чи інших картин світу, серед яких і чітко обґрунтовані методи із застосуванням семного аналізу [8]. Але нас зараз більше цікавить художня картина світу (частиною якої є поетична), яка відіграє суттєву роль в сприйнятті дійсності. З. Попова та І. Стернін представляють її як вторинну, що виникає в свідомості читача при сприйнятті художнього твору [5]. Картина світу в художньому тексті відбиває індивідуальну картину світу в свідомості автора і створюється тими ж мовними засобами, але своєрідно. Ця своєрідність полягає у використанні окремих тематичних груп мовних одиниць, різній (підвищеній чи зниженій) частотності окремих одиниць та їхніх груп, в індивідуальному використанні образних засобів (системи тропів) тощо.

Створення тезауруса, як відомо, може відбуватися декількома шляхами: згори донизу (накладанням заздалегідь розробленої синоптичної (класифікаційної) схеми, яка покликана відображати мовну картину світу, на мовний матеріал) або знизу вгору (виведенням синоптичної схеми шляхом аналізу лексичних значень). Найдоцільнішим, на нашу думку, є поєднання цих двох підходів – самостійна розробка синоптичної схеми шляхом узагальнення мовних понять і порівняння її з відомими класифікаціями. Стосовно ж побудови тезауруса поетичної мови і створення поетичної мовної картини світу, то видається прийнятним варіант використання загальномовного тезауруса шляхом перетину поетичної і загальномовної баз даних. Проте, не маючи повного загальномовного тезауруса української мови, зараз можна говорити лише про дослідження окремих аспектів художніх поетичних явищ (напр., ідіостилю, тропіки, лексики різних семантичних груп) і побудову фрагментів поетичної мовної картини тощо [6].

Інструментом дослідження послужив створений нами на базі тлумачного словника комп'ютерний тезаурус, що представляє собою ієрархічно впорядковану структуру, одиниці якої об'єднано в семантичні поля.

Стаття тезауруса має ономасіологічну спрямованість – складається з реєстрової одиниці (концепту) та лексичних одиниць, що передають це поняття у мові. До складу останніх входять також синоніми, антоніми й деривати. У ролі лексичних відповідників концепту виступають лексеми, у яких категоріальна та диференційна ознаки змісту концепту відповідають категоріально-лексичній та диференційній семам значення слова. Таким чином узгоджуються поняття й значення, логічна та мовна частини словника.

Зауважимо, що лексичними одиницями тезауруса є окремі лексико-семантичні варіанти (ЛСВ), а не лексеми, оскільки лексичній семантиці іменників притаманна багатозначність. Різні ЛСВ одного іменника на позначення людини/особи можуть називати різні характеристики і властивості людини. Наприклад, різні ЛСВ лексеми *безштанько* співвідносяться з такими поняттями: *безштанько 1* – маленький хлопчик,

безштанько 2 – бідна людина. Саме тому при ідеографічній класифікації лексики враховувалися всі ЛСВ іменників, які позначають поняття людини/особи.

Джерелом відбору лексичного матеріалу до тезауруса був словник української мови в одинадцяти томах [7]. Масив лексики, що підлягає ідеографічному опису в тезаурусі, становить близько 15 тис. ЛСВ похідних (переважна більшість) і непохідних іменників, значення яких містять категоріально-лексичну сему "людина/особа". Аналізуючи визначення іменників у тлумачному словнику, методом суцільної вибірки було відібрано іменники, субстантивовані прикметники та дієприкметники, що вживаються в значенні іменників, на позначення людей.

Оскільки мова служить не тільки засобом об'єктивізації, репрезентації світу предметів та ідей, але й забезпечує також сфери мовленнєво-мисленнєвої та емоційної діяльності людини, то крім нейтральної ми вважали за доцільне включити до вибірки емоційно та стилістично забарвлені, рідковживані, діалектні та застарілі (архаїзми та історизми) ЛСВ слів на позначення людей/осіб. Хоча тезаурус і представляє синхронний зріз лексики, такий вибір зумовлений як суто дослідницьким інтересом, так і тим, що однією з функцій тезауруса є необхідність задоволення пошуку "поняття – слово" і навпаки. До того ж, за запитом користувача до результатів пошуку в електронному тезаурусі можна не включати стилістично марковану лексику.

Далі методом компонентного аналізу виділялися родова (класифікаційна) і додаткові семи значення. За певними ознаками семи порівнювалися, і за результатами аналізу відповідні ЛСВ об'єднувалися в групи і семантичні поля.

Тезаурус представлений у вигляді індексованої електронної бази даних Access, де кожна одиниця має набір міток-параметрів: найближче родове поняття, тип семантичного зв'язку з іншою одиницею, стилістичну інформацію тощо.

Матеріалом для побудови поетичного тезауруса було обрано електронний частотний словник поетичної мови зі статистичними характеристиками, розроблений і створений на творах 15 авторів (періоду 1975–1995) колективом лабораторії комп'ютерної лінгвістики Інституту філології Київського національного університету імені Т. Шевченка (Н. П. Дарчук, Л. А. Алексієнко, О. М. Зубань та ін.) [3]. Словник створено на корпусі текстів обсягом 300 000 слововживань із творів поетів-шістдесятників (Л. Костенко, І. Драч, Б. Олійник, В. Стус, В. Коломієць, Т. Коломієць, Д. Павличко, М. Вінграновський, М. Воробйов), поетів сімдесятих років (І. Жиленко, С. Йовенко, Л. Скирда) і поетів дев'яностих (Ю. Андрухович, О. Забужко, І. Малкович). Довжина вибірки з текстів кожного поета – 20 000 слововживань. Частотний словник надає інформацію про закономірно повторювані лексеми, які є ключовими для індивідуального авторського словника та поетичного в цілому.

Побудова фрагмента тезауруса поетичної мови відбувалася в напівавтоматичному режимі: кожне слово із бази даних тезауруса іменників на позначення людини порівнювалося зі словами з п'ятнадцяти підвибірок поетів. Якщо знаходилася відповідна словоформа, їй приписувався номер лексеми тезауруса і номер значення, у якому вона зустрілася в тексті. Із частотного словника бралася відповідна статистична інформація про частоту лексеми і частоту словоформи. Результати цього етапу наведені у табл. 1. В індивідуальних тезаурусах не враховані поки що авторські неологізми на позначення людей, які не засвідчені в тлумачному словнику.

Таблиця 1. Частотні характеристики словника іменників окремих авторів

№	Автор	Кількість різних:				Кіл-ть ЛСВ на позна-чення людини	Відношення кількості лексем-назв людей до загальної кількості іменників	Відношення кількості словоформ на позначення людини до загальної кількості іменникових словоформ
		лексем	лексем-назв людей	словоформ	словоформ на позначення людини			
1	Андрухо-вич Ю.	3122	421	4933	670	445	13,5%	13,6%

2	Вінгранов-ський М.	1751	118	3252	339	132	6, 7%	10, 4%
3	ВоробйовМ.	1608	139	3193	382	155	8, 6%	12, 0%
4	Драч І.	2404	251	3818	597	269	10, 4%	15, 6%
5	Жиленко І.	2219	196	3802	576	221	8, 8%	15, 1%
6	Забужко О.	2712	182	4181	375	192	6, 7%	9, 0%
7	Йовенко С.	1888	206	3186	513	221	10, 9%	16, 1%
8	КоломієцьВ.	2549	263	4188	666	298	10, 3%	15, 9%
9	КоломієцьТ.	2034	186	3536	434	204	9, 1%	12, 3%
10	Костенко Л.	2606	333	4040	706	357	12, 8%	17, 5%
11	Малкович І.	1982	175	3159	279	165	8, 8%	8, 8%
12	Олійник Б.	2097	281	3463	840	316	13, 4%	24, 3%
13	Павличко Д.	2201	236	3830	496	253	10, 7%	13, 0%
14	Скирда Л.	1993	145	3338	431	161	7, 3%	12, 9%
15	Стус В.	2268	253	3664	512	270	11, 2%	14, 0%
Усього		13233	1548	29150	7816	3659	11, 7%	26, 8%

Процентні відношення частоти лексем на позначення людини до загальної кількості іменникових лексем (11, 7%) та частоти словоформ на позначення людини до загальної кількості іменникових словоформ (26, 8%) у частотному поетичному словнику дозволяють зробити припущення, що українська поезія другої половини ХХ ст. була певною мірою антропоцентричною. Однак це припущення вимагає додаткових досліджень і розрахунків: для більш точного аналізу необхідно знати кількість та наповнення інших ЛСГ ("Тварини", "Рослини", "Їжа", "Напої", "Засоби пересування", "Одяг" тощо), а залежно від способу поняттєвої класифікації різні дослідники виділяють неоднакову їх кількість від 3 до 100 і більше. Ще точнішими дані могли б бути, якби в розпорядженні дослідника був ще й корпус поетичних текстів.

Найбільший тезаурус назв людей і поезія у Ю. Андруховича – 421 різна лексема і 670 їхніх словоформ на позначення людини. Цікава ситуація у підвибірці Б. Олійника: його тезаурус не надто великий – 281 лексем, проте вживає він їх найчастіше серед усіх поетів – 840.

Отже, можна зробити висновок: чим більша кількість лексем у поезії автора і висока частота вживання їхніх словоформ, тим багатший тезаурус поета. А невисока частота лексем поряд з високою частотою вживання їхніх словоформ може свідчити про усталеність світогляду автора, мислення усталеними категоріями, концептами.

Проте не всі лексико-семантичні поля і підгрупи однаково представлені: із близько 15 тис. різних ЛСВ іменників на позначення людини загальномовного тезауруса лише 1714 наявні в поетичному словнику.

Найчастотнішими лексемами в поетичному словнику (див. табл. 2) виявилися слова на позначення загальної назви сукупності – **люди** (270).

Цікаво, що на другому місці за частотою лексема **жінка** (229), а далі – **мати** (203).

Підсумовуючи, можна припустити, що антропоцентризм української поезії має виражену ознаку статі: роль **жінки-матері** в ній займає провідні позиції. Це підтверджує і побудова синоптичної схеми поетичного тезауруса, яка засвідчила, що найчисельнішим є семантичне поле *Сімейних і Родинних стосунків* (чи не в повному обсязі представлене лексемами **мати, мама, діти, син, чоловік, брат, батько, дід, баба, дядько**). Другою за чисельністю і значущістю є група *Соціальних стосунків* **(друг, гість)** і теми **людини** і **народу**, а також тема мистецтва (частота лексеми **поет** може свідчити про пильну увагу та вимоги авторів до себе та інших митців).

Для більш точних висновків до аналізу можна залучити статистичні дані щодо частоти значень, у яких вжито найчастотніші слова, та їхні контексти. У цьому дослідженні проілюстровано приклад застосування тезауруса і наведено результати апробації методики укладання тезауруса на поетичних текстах.

На основі отриманих результатів проведено порівняння фрагментів мовних картин світу: загальномовної та індивідуальної (поета). Отримані результати дозволяють зробити

припущення щодо антропоцентричності української поезії другої половини двадцятого століття (проте це припущення вимагає додаткових досліджень) із важливими темами родинних і суспільних стосунків, з провідною роллю жінки в поетичному дискурсі (про що свідчать кількісні характеристики ЛСГ "Людина" поетичного тезауруса).

Таблиця 2. Найчастотніші лексеми на позначення людини в частотному поетичному словнику

	Лексема	**Автори**															**Усього в словнику поетів**
		Андрухович Ю.	**Драч І.**	**Йовенко С.**	**Коломієць Т.**	**Коломієць В.**	**Костенко Л.**	**Малкович І.**	**Олійник Б.**	**Павличко Д.**	**Скирда Л.**	**Стус В.**	**Вінграновський М.**	**Воробйов М.**	**Забужко О.**	**Жиленко І.**	
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
1.	люди	3	17	7	24	44	50	5	22	16	14	18	10	17	16	7	**270**
2.	жінка	10	12	40	18	6	15	4	5	5	13	5	13	27	24	32	**229**
3.	мати	3	22	10	22	14	8		32	24	4	18	14	27	3	2	**203**
4.	діти	10	9	12	9	22	9	4	14	9	24	3	5	23	11	8	**172**
5.	син	7	24	4	17	20	8	4	25		10	28	4	4	1	16	**172**
6.	людина	2	9	14	15	19	11	1	11	14	33	9	8	9	1	15	**171**
7.	народ	7	8	2	7	34	17	1	25	11	15	3	21	1	5	1	**158**
8.	мама	1	17	21	11	16	2	10	16	2	21	6	9	3	9	12	**156**
9.	поет	10	9	13	4	9	20	7	29	4	10	7	6	2	11	10	**151**
10.	друг	1	7	8	6	8	7		7	10	33	18	6	5	3	21	**140**
11.	чоловік	4	12	12	1	5	7	3	22	3	11	6	4	25	14	8	**137**
12.	брат	5	5	9	2	7	3	5	16	13	9	5	4	18	3	7	**111**
13.	дитина		13	19	2	7		8	8	12	7		7	2		6	**91**
14.	хлопець	5	4		16	10	7	1	25	2	1	3	1	4	4	1	**84**
15.	дитя	2	4	19	2	4	5		2	6	18		6			14	**82**
16.	батько	1	8	6	5	14	3		11	9	5	1	5	5	3		**76**
17.	дід	1	18		4	11	3	2	16	1	2		7	5		4	**74**
18.	баба	1	20		2	1	6		4			1	7	4		9	**55**
19.	дівчина	3	9		3	8	1		4	12	1	3	4	5	1	1	**55**
20.	гість	4	3		7	2	4	2			2	1	1	8	1	16	**51**
21.	дядько		15		3	6			5	3			13	2		4	**51**

Література

1. Алексеенко Л. А., Дарчук Н. П., Зубань О. Н., Сорокин В. В. Параметризированная база данных поэтической речи как источник и инструмент филологических студий // Материалы международной конференции "Компьюерная лингвистика без границ", Санкт-Петербург, 25–26 марта 2004 г. – СПб., 2004.
2. Баранов А. Н. Автоматизация лингвистических исследований: корпус текстов как лингвистическая проблема // Руссистика сегодня. – 1998. – №1–2. – С. 179–191.
3. Дарчук Н. П. Структурно-статистическая база данных современного украинского языка на основе частотных словарей // Слово и словарь = Vocabulum et vocabularium: Сб. науч. тр. по лексикографии. – Гродно: ГрГУ, 2005. – С. 194–197.
4. Ляшевская О. Н., Плунгян В. А., Сичинава Д. В. Национальный корпус русского языка как инструмент лексикографа // Слово и словарь = Vocabulum et vocabularium: Сб. науч. тр. по лексикографии. – Гродно: ГрГУ, 2005. – С. 197–202.
5. Попова З. Д., Стернин И. А. Язык и национальная картина мира. – Воронеж, 2002. – 59 с.
6. Сирук Е. Б., Сорокин В. М. Использование Компьютерного тезауруса в статистическом анализе лексики художественного текста (на материале украинского языка) // Материалы международной конференции "Тексты и контексты: движение языка", Каунас, 6–7 мая 2004 г. – Каунас, 2004.
7. Словник української мови: В 11 т. – Київ, 1970–1980.
8. Соколовська Ж. П. Картина світу та ієрархія сем // Мовознавство. – Київ, 2002. – №6. – 87–91.

С. Бук, к. філол. н.*
Львівський національний університет імені Івана Франка (Львів)
УДК 161.2.81'373.374.322

СТАТИСТИЧНІ ХАРАКТЕРИСТИКИ ЛЕКСИКИ ОСНОВНИХ ФУНКЦІОНАЛЬНИХ СТИЛІВ УКРАЇНСЬКОЇ МОВИ: СПРОБА ПОРІВНЯННЯ

Résumé: In the article, for the first time, the main statistical characteristics, such as the variety, exclusiveness, concentration indices are compared for the vocabulary of six modern Ukrainian language functional styles: belles-lettres, poetry, journalistic, colloquial, scientific and official. The research has been done at the material of the 6 appropriate frequency dictionaries and the number of words and wordforms is given as well. The result is the complete correlation of all mentioned statistical characteristics.

Key words: frequency dictionary, (absolute and relative) word frequency, word occurrence, sampling, statistical description of vocabulary.

У лінгвістиці зламу XX–XXI століть особливо актуальними стали міждисциплінарні дослідження мови. Таким напрямком є й статистична лінгвістика. Як зауважив німецький лінгвіст Г. Альтманн, "Шлях дисципліни вглиб рано чи пізно наштовхується неминуче на обмеженість якісних методів, на безпорадність неточного способу вираження, на відсутність гіпотез, а також на відсутність теорії" [1: 5].

Мова як складна система дискретних одиниць має окрім якісних (які є об'єктом вивчення фонетики, дериватології, лексикології, синтаксису, теорії тексту, комунікативної лінгвістики тощо), також й кількісні характеристики. Останні властиві усім рівням мовної системи, та особливо виразні для лексичного.

Перші статистичні дослідження української мови було здійснено Інституті мовознавства ім. О. Потебні [11; 10]. Пізніше з метою виявлення лексичних особливостей різних стилів укладено їх частотні словники (далі — ЧС). Так, зараз в українській лексикографії існують ЧС для п'яти функціональних стилів, які більшість підручників зі стилістики визначають основними: для художньої прози [12], публіцистики [7], розмовно-побутового [4], наукового [3], офіційно-ділового [5]; а також для поетичного мовлення [6]. Аналіз залежності ранґ-частота для текстів, що становлять джерельну базу цих словників, здійснено в [16]. Цікаво порівняти й базові статистичні характеристики їх словникового складу: обсяг словника словоформ ($V_ф$), обсяг словника лексем (V), багатство словника (Б), індекс винятковості (Е), середня повторюваність слова у тексті (А) тощо.

Існують різні підходи до способів здійснення цієї процедури. Так, наприклад, В. Перебийніс з цією метою запропонувала поняття "нульовий стиль" [13], М. Арапов зі

* © *С. Бук, 2006*

співавторами висунув теорію визначення “кількісної відстані” між словниками [2] тощо. Проте незаперечним є факт, що питання, пов’язані із методами порівняння ЧС, тісно переплетені з принципами їхнього укладання, “відмінності, інколи вельми відчутні, у методиці побудови словників сильно ускладнюють їх коректне порівняння” [15: 300], іншими словами, для отримання достовірних результатів важливо, щоби порівнювані словники були створені на одинакових засадах. Здійснимо короткий опис згаданих ЧС.

1. ЧС сучасної української художньої прози [12] укладено вручну на вибірці текстів 25 письменників загальним обсягом 500 000 слововживань.
2. В основу ЧС публіцистики [7] лягли тексти всеукраїнських газет за 1994 р., розрахованих на читачів різних регіонів, різних вікових та соціальних груп. Загальний обсяг вибірки — 300 000 слововживань.
3. ЧС розмовно-побутового стилю [4] укладено на вибірці 300 000 слововживань. Оскільки в українській мовознавчій практиці Національного корпусу текстів усного мовлення ще не існує, то матеріал дослідження склали 45 текстових уривків з прозових драм (хронологічні межі 1982–2002 рр.) приблизно по 6 000 слововживань кожен. Аналогічні джерела для відповідних словників використовували, наприклад, автори публікацій [17], [9].
4. ЧС наукового стилю [3] побудовано на 104 уривках приблизно по 3 000 слововживань кожен. Загальний обсяг — понад 300 000 тис. Тут представлено майже у рівних пропорціях такі галузі наук: біологія та хімія, психологія і педагогіка, фізика та математика, техніка, географія та геологія, історія та мовознавство. Зауважимо, що в ЧС наукового стилю [8], віднедавна доступному в Інтернеті, опрацьовано тільки гуманітарні науки: філософію, літературознавство та мовознавство.
5. Джерелами ЧС офіційно-ділового стилю [5] стали тексти підручників ділової мови та офіційних документів загальним обсягом 300 000 слововживань: Конституція України, кодекси, українські та міжнародні закони, міжнародні договори, конвенції, меморандуми, економічні й адміністративні документи та ін.
6. На текстах 1975–1995 рр. п’ятнадцяти поетів загальним обсягом 300 000 слововживань побудовано й ЧС поетичної мови [6].

Усі названі вище ЧС було укладено за майже ідентичними принципами (в тому числі застосовано практично однакові схеми лематизації слів) з використанням комп'ютера (окрім першого, який укладався вручну). Також видно, що величина вибірки усіх словників — 300 тис. слововживань, окрім ЧС художньої прози, який створено на текстах довжиною 500 тис. слововживань. Тому, для правомірності порівнянь словників, частоти усіх слів в ньому було відповідно прошкальовано за формулою:

$$n_2 = n_1 / 500 \times 300, \qquad (1)$$

де n_2 — частотність одиниці у вибірці 300 тис. слововживань, n_1 — частотність одиниці у вибірці 500 тис. слововживань.

Обчислення параметрів для словника художньої прози. При укладанні пробного зошита ЧС художньої прози було помічено, що “... подовження тексту неоднаково впливає на збільшення кількості слів і словоформ: кількість словоформ зростає дещо швидше, ніж кількість слів” [12: 14], тому показник багатства словника, обчисленого для текстів різної довжини, зіставляти не можна — простої формули для шкалювання частот (1) недостатньо. Проте для ЧС художньої прози відома кількість слів W та словоформ F для двох точок: 100 тис. і 500 тис. [12: 14], тому можна обчислити таку кількість для деякої третьої точки — 300 тис. слововживань. Згідно з Ю. Тулдавою [14: 62], частотні характеристики підлягають так званій алометричній залежності на зразок

$$y = Ax^b,$$

де A, b — константи (їх треба рахувати у нашому випадку), а x, y — певні величини.

Якщо W — кількість слів (для 100 тис. — 13 258, для 500 тис.— 33 391), F — кількість словоформ (для 100 тис. — 26 275, для 500 тис.— 86 284, а N — обсяг тексту (відповідно 100 і 500 тис.), то відношення F / W для деякого третього значення (N = 300 000) на підставі даних для двох інших (при N = 100 000 і N = 500 000):

$$\frac{F}{W} = AN^b \qquad (2)$$

[14: 64]. Виконавши обчислення, матимемо $A \approx 0.297$, $b \approx 0.165$, звідки для $N = 300\,000$ за цією ж формулою маємо відношення кількості словоформ до кількості слів 2, 38.

Враховуючи, що величини F і W також пов'язані подібною залежністю [14: 63]:

$$W = CF^d \qquad (3)$$

з параметрами $C \approx 4.89$, $d \approx 0.777$ (які також розраховано на ПК на підставі даних для масивів 100 тис. і 500 тис. слововживань), можна обчислити F і W для $N = 300\,000$. Справді, підставляючи (3) в (2), матимемо:

$$\frac{F}{C\,F^d} = AN^b; \quad F^{1-d} = AC\,N^b; \quad F = \left(AC\,N^b\right)^{1/(1-d)}.$$

Вийде значення кількості слів 24 906, словоформ — 59 161 (дані заокруглено до цілих).

Кількість hapax legomena (слів, що зустрілися в досліджуваному масиві текстів лише один раз), яку позначимо H, пов'язана з обсягом тексту подібним законом:

$$H = KN^q,$$

де K і q — деякі константи (див., наприклад [18: 81]). Для частотного словника художньої прози значення $K \approx 26,7$, $q \approx 0,48$, що дає при $N = 300\,000$ кількість hapax legomena 11 342. Далі отримані вказаним способом значення параметрів словника художньої прози позначено зірочкою (*).

Порівняння основних статистичних характеристик лексики функціональних стилів української мови найзручніше показати в таблиці:

Таблиця 1. Зведена порівняльна таблиця основних частотних характеристик п'яти ЧС української мови

	V	$\boldsymbol{V_ф}$	$\boldsymbol{V / N}$	$\boldsymbol{V_1 / N}$	$\boldsymbol{V_1 / V}$	$\boldsymbol{V_{10т} / N}$	$\boldsymbol{V_{10} / V}$
ЧС поетичної мови	31194	69012	0, 103	0, 052	0, 495	0, 789	0, 098
ЧС худ. прози	24906* (33391)	59161* (86284)	0, 083* (0, 067)	0, 038* (0, 029)	0, 455* (0, 430)	— (0, 821)	— (0, 149)
ЧС публіцистики	20824	43363	0, 070	0, 031	0, 450	0, 789	0, 161
ЧС розм.-поб. стилю	20541	42106	0, 073	0, 034	0, 465	0, 804	0, 121
ЧС наукового стилю	19367	42802	0, 059	0, 025	0, 427	0, 890	0, 189
ЧС офіційно-ділового стилю	9045	24263	0, 030	0, 0085	0, 280	0, 935	0, 303

Тут $\boldsymbol{V}$ — кількість різних слів (обсяг словника лексем), тобто слів у словнику, $\boldsymbol{V_ф}$ — кількість словоформ (обсяг словника словоформ), $\boldsymbol{V / N}$ — індекс різноманітності, $\boldsymbol{V_1 / N}$ — індекс винятковості тексту, $\boldsymbol{V_1 / V}$ — індекс винятковості словника, $\boldsymbol{V_{10т} / N}$ — індекс концентрації тексту, $\boldsymbol{V_{10} / V}$ — індекс концентрації словника.

Обсяг словника словоформ ($V_ф$) та **обсяг словника лексем** (V) — це величини, що показують, скільки окремих лексем (лематизованих, тобто зведених до початкової форми, слів) чи словоформ вжито у відповідних проаналізованих корпусах текстів конкретних функціональних стилів. Як видно із порівняльної таблиці, найбільше як слів, так і словоформ є в поетичному мовленні, далі за зменшенням ідуть художній, публіцистичний, розмовно-побутовий та науковий стилі. Найменше слів і словоформ (у два рази (!) менше, ніж в усіх інших) зафіксовано в офіційно-діловому. Це експериментальне підтвердження апріорно зрозумілого факту, що при написанні офіційно-ділового тексту користуються усталеними зворотами, обмеженою лексикою, уживають менше лексичних засобів, ніж в інших функціональних стилях.

Багатство словника (*Б*), тобто індекс різноманітності — це відношення обсягу словника лексем (*V*) до обсягу тексту (*N*), що обчислюється за формулою:

$$Б = V / N.$$

Підставивши відповідні дані, ми отримали, що в поезії ця величина становить 0, 103, в художній прозі — 0, 083* (0, 067), в публіцистиці — 0, 070, в розмовно-побутовому стилі — 0, 069, в науковому — 0, 065, в офіційно-діловому — 0, 030.

Індекс різноманітності абсолютно точно корелює із даними обсягу словника: найвищий він в поетичному мовленні, найнижчий — в офіційно-діловому, посередині за спадом — художній, публіцистичний, розмовно-побутовий та науковий.

Середня повторюваність слова у тексті (*А*), тобто відношення обсягу тексту (*N*) до обсягу словника лексем (*V*) — величина, обернена до індексу різноманітності, — обчислюється за формулою:

$$A = N / V.$$

Цей показник для поезії становить 9, 6, для художнього стилю — 12, 1* (14, 9), для публіцистики — 14, 3, для розмовно-побутового — 14, 5, для наукового — 15, 5, для офіційно-ділового — 33, 17. Іншими словами, кожне слово в середньому зустрілося в конкретному досліджуваному корпусі текстів 10–16 разів (виняток — офіційно-діловий стиль).

Індекс винятковості *для тексту* (у нашому випадку для тексту 300 000 слововживань) ($E_т$), тобто відношення кількості слів із частотою 1 — hapax legomena — (V_1) до обсягу тексту (*N*), обчислюється за формулою:

$$E_т = V_1 / N,$$

Індекс винятковості *для словника* (тобто для загальної кількості окремих слів, зведених до початкової форми *V*) ($E_с$) — за формулою:

$$E_с = V_1 / V,$$

Результати обчислень показують, що найбільший індекс винятковості і для словника, і для тексту, як і слід було очікувати, — в поезії, найменший — в офіційно-діловому. Далі за зменшенням цього показника ідуть художній, розмовно-побутовий, публіцистичний та науковий стилі. Цей параметр є показником варіативності лексики.

Індекс концентрації — величина, протилежна до індексу винятковості, що вказує, яку частку *тексту* (*N*) або *словника* (*V*) займає високочастотна лексика (з абсолютною частотою 10 і більше). Позначається відповідно $V_{10т}$ і V_{10}. Обчислюється за формулами $V_{10т} / N$ — індекс концентрації тексту та V_{10} / V — індекс концентрації словника. Показники цієї величини є найнижчими для поезії і, логічно, — найвищими для офіційно-ділового стилю.

Отже, порівняння статистичних даних лексики п'яти основних функціональних стилів і поезії вперше в українській лінгвістиці дає конкретні кількісні дані до тези, що найбагатша лексика — в художній поезії та прозі, далі — в публіцистиці, розмовно-побутовому, науковому та офіційно-діловому стилях[1]. Про це свідчить кореляція відповідних величин: кількості слів в ЧС, індекси різноманітності та винятковості, середня повторюваність слова. Суттєво, що наведені факти виявлено на однаковій за розміром вибірці текстів згідно з вимогами до порівнянь ЧС. Це також може слугувати підтвердженням, що вже така невелика з погляду сучасної корпусної лінгвістики вибірка — 300 000 слововживань — репрезентує основні відмінності функціонування слів у різних стилях української мови.

Література

1. Альтман Ґ. Мода та істина в лінгвістиці // Проблеми квантитативної лінгвістики.— Чернівці: Рута, 2005.— С. 3–11.
2. Арапов М. В., Тер-Гаспрян Л. И., Херц М. М. Сравнение частотных словарей // Научно-техническая информация. Серия 2. Информационные процессы и системы.— 1978.— №4.— С. 20–29.
3. Бук С. 3 000 найчастотніших слів наукового стилю української мови.— Львів: Львівський національний університет імені Івана Франка, 2006.—192 с.
4. Бук С. 3 000 найчастотніших слів розмовно-побутового стилю української мови.— Львів: Львівський національний університет імені Івана Франка, 2006.— 180 с.

[1] У попередньо виданій монографії [11], аналізуючи статистичні параметри лексичного рівня, автори зосередили свою увагу на характеристиці особливостей вживання дієслівних форм у різних стилях.

5. Бук С. Частотний словник офіційно-ділового стилю сучасної української мови // Лінгвістичні студії: Зб. наук. праць.— 2006.— Випуск 14.— С. 184–188.
6. Дарчук Н. П. (ред.) Частотний словник сучасної поетичної української мови // Лінгвістичний портал MOVA.info.— 2003–2006.— [Цит. 11 березня 2006].— Доступно з <http://www.mova.info/Page2.aspx?l1=89>; Дарчук Н. П. (ред.) Частотний словник сучасної поетичної української мови.— [Цит. 01 грудня 2001].— Доступно з <http://www.philolog.univ.kiev.ua/WINS/chast/chast.htm>.
7. Дарчук Н. П. (ред.) Частотний словник сучасної української публіцистики // Лінгвістичний портал MOVA.info.— 2003–2006.— [Цит. 11 березня 2006].— Доступно з <http://www.mova.info/Page2.aspx?l1=91>.
8. Критська В. І., Колєнов Г. В., Недозим Т. І. та ін. Частотний словник наукового стилю // Лінгвістичний портал MOVA.info.— 2003–2006.— [Цит. 11 березня 2006].— Доступний з: <http://www.mova.info/article.aspx?11=176&DID=1700>.
9. Морковкин В. В. Сравнительный список наиболее употребительных русских слов (на материале шести словарей) // Лексические минимумы русского языка / Под ред. П. Н. Денисова.— М.: МГУ, 1972.— С. 16–74.
10. Муравицька М. П., Олексієнко Л. А. (відп ред.) Структура мови і статистика мовлення.— К.: Наукова думка, 1974.— 176 с.
11. Перебийніс В. С. (ред.) Статистичні параметри стилів.— К.: Наук. думка, 1967.— 260 с.
12. Перебийніс В. С. (ред.) Частотний словник сучасної української художньої прози.— К.: Наук. думка, 1981.— Т. 1.— 863 с.; Т. 2.— 855 с.
13. Перебейнос В. С. Методы и уровни моделирования нулевого стиля // Вопросы статистической стилистики.— К.: Наука, 1974.—331 с.— С. 16–35.
14. Тулдава Ю. П. Проблемы и методы квантитативно-системного исследования лексики.— Таллин: Валгус, 1987.— 204 с.
15. Якубайтис Т. А. О статистических пластах лексики // Вопросы статистической стилистики.— К.: Наукова думка, 1974.— С. 299–314.
16. Buk, S. N., Rovenchak, A. A. Rank–Frequency Analysis for Functional Style Corpora of Ukrainian // Journal of Quantitative Linguistics.— 2004.— V. 11, No. 3.— P. 161–171.
17. Juilland A., Brodin D., Davidovitch C. Frequency Dictionary of French Words.— The Hague; Paris, 1970.
18. Kornai A. How many words are there? // Glottometrics.— 2002.— V. 4.— P. 61–86.

Н. Андрейчук, І. Волошиновська*
Національний університет „Львівська політехніка” (Львів)
УДК 81’373.322.324

ДІЄСЛОВА, ЩО ВИЗНАЧАЮТЬ ГЛИБИНУ ОПРАЦЮВАННЯ ПРЕДМЕТА ДОСЛІДЖЕННЯ, ТА ЇХ СТАТИСТИЧНІ ХАРАКТЕРИСТИКИ У НАУКОВО-ТЕХНІЧНИХ ТЕКСТАХ

The list of key verbs denoting the stage of scientific research is suggested and their frequency analysis when used in scientific texts is viewed as means of determining of the depth of the subject of research cognition.

Вступ

У процесі пізнання людиною довкілля одним з основних моментів є збереження набутої інформації та її передача наступним поколінням. Для забезпечення поступу у певній науковій галузі корисним, а в більшості випадків навіть необхідним, є аналіз робіт попередників у напрямку запланованих та споріднених досліджень. Систематизація напрацьованого матеріалу, встановлення критеріїв та розробка методів пошуку і відбору базової інформації сприяють ефективній передачі знань, що є особливо важливим в умовах високого темпу розвитку сучасних технологій. Передові розробки передбачають діяльність науковців у міжждисциплінарних сферах, що, разом з неналежним опрацюванням базової інформації, доволі часто зумовлює невідповідність хронології робіт їх науковій новизні. Такі обставини ускладнюють пошук джерел з інформацією про новітні досягнення у заданому напрямку. З іншого боку, новизна не завжди виступає критерієм відбору інформації, необхідної для успішного вирішення завдань, тому потрібно прослідкувати також початкові та проміжні етапи розвитку досліджень. Таким чином, вартою уваги є проблема визначення стадії досліджень, описаних в роботі, що піддається аналізу.

* © *Н. Андрейчук, І. Волошиновська, 2006*

Таблиця 1. Дієслова “дослідницької” групи.

Підгрупа **I** *(початковий етап)*	Підгрупа **II** *(завершальний етап)*
assume	accomplish
estimate	achieve
expect	assign
intend	claim
plan	classify
predict	confirm
presume	determine
presuppose	develop
propose	discover
suggest	elaborate
suppose	establish
surmise	find
suspect	implement
think	prove

1. Формування набору ключових дієслів

В даній роботі розглянута можливість встановлення етапу наукового дослідження на основі аналізу дієслів “дослідницької” групи. Так умовно можна назвати групу дієслів, що вживаються в текстах, зокрема, науково-технічного спрямування для відображення різноманітних станів дослідженості предмета, виражаючи: припущення, очікування, наміри, прогнозування, досягнення, виявлення, підтвердження, створення, завершення. Відбір таких дієслів здійснено на основі статистичного аналізу наукових праць з журналу “Physical Review B” (тематика: Конденсовані системи та фізика матеріалів) Американського фізичного товариства (APS) [1; 2]. Вибір даного журналу пояснюється наявністю в галузі конденсованих систем та матеріалознавства фундаментальних доробок поряд із новітніми розробками, значна частина яких перебуває на початкових стадіях. Таким чином було забезпечено широкий діапазон стадії досліджень (описаних у роботах, що піддавались статистичному аналізу) з метою вибору повного набору дієслів, які можуть характеризувати глибину розробки тематики робіт. Синонімічні множини вибраних дієслів були враховані для розширення числа елементів дослідницької групи, а її остаточне формування та виділення підгруп проводились на основі аналізу семантики кожного елементу. В результаті, виділено дві підгрупи дієслів, які характеризують початковий та кінцевий етапи досліджження (табл. 1). В процесі узагальнення статистичних характеристик відповідних підгруп виділені ключові дієслова, що найчстіше вживаються науковцями: 1) assume, estimate, expect, suggest; 2) achieve, determine, develop, find.

2. Визначення глибини досліджень

Запропонована група дієслів застосовувалась для аналізу дисертаційних та дипломних робіт наукофо-технічного спрямування на предмет глибини проведених досліджень. До розгляду були взяті праці, захищені протягом 2003-2004 рр. за результатами дослідницьких робіт [3; 4] наукового центру DESY (Deutsches Elektronen- Synchrotron, Німеччина).

Графічне відображення результатів проведеного аналізу представлене на рис. 1. Аналізуючи розподіл праць відповідно до вкладів дієслів, що характеризують початковий та завершальний етапи дослідження, можна виділити *три множини робіт*, в яких представлені результати досліджень на різних стадіях розробки.

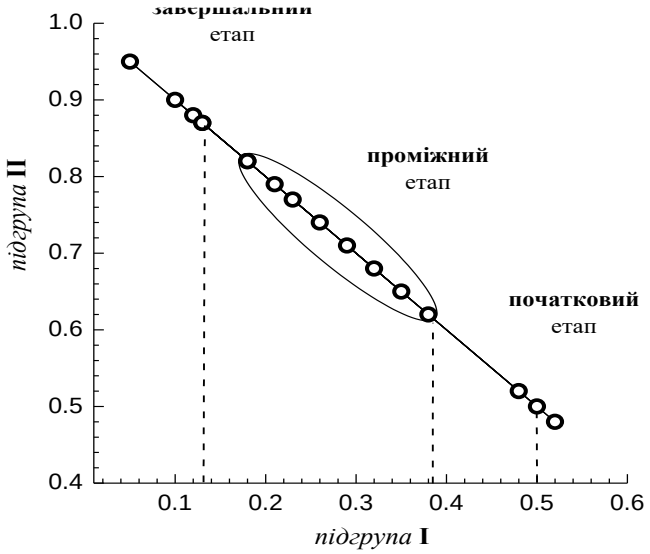
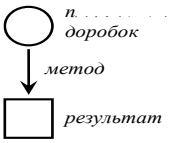


Рисунок 1. Розподіл праць відповідно до вкладів дієслів, що характеризують початковий (*підгрупа* **I**) та завершальний (*підгрупа* **II**) етапи дослідження.

Отриманий розподіл праць на вище згадані множини можна пояснити на основі наступних міркувань. Науково-дослідницька праця являє собою представлення трьох основних складових (див. рис. 2):

1) *предмету* досліджень та *базового доробку* у заданій та спорідненій тематиках;
2) можливих або вже наявних *методів* досліджень;
3) *результатів*.

Рисунок 2. Узагальнена схема науково-дослідницької праці.



У цій роботі встановлено, що етап наукового дослідження може бути визначений на основі аналізу дієслів запропонованої дослідницької групи, які домінують в описі другої (*методу*) та третьої (*результату*) складових дослідницької праці. В описі першої складової завжди домінують дієслова *другої підгрупи,* відображаючи наявність *предмету*, а, можливо, і *базового доробку* на певному етапі дослідницької роботи.

3. Аналіз отриманих результатів

У випадку *початкового етапу* досліджень *друга підгрупа* дієслів задіяна для опису лише *предмету* опрацювання, а опис *методів* та *результатів* проводиться з використанням дієслів *першої підгрупи*, відображаючи відсутність реальних результатів і наявність лише їх оцінок, а також припущень щодо можливих методів та засобів їх досягнення. При цьому *відсутні,* або ж ігноруються розроблені *методи,* що

застосовувались попередниками до предметів дослідження. Випадок дослідження на *початковому етапі* відображений на рис. 3а, де пунктиром позначені складові науково-дослідницької праці (метод, результат), в описі котрих домінують дієслова першої підгрупи. Відносний вклад дієслів першої підгрупи сягає свого максимуму в околі значення 0.5 (див. рис. 1), де зосереджені праці відповідають початковим етапам дослідження. Близькі значення відносного вкладу дієслів першої та другої підгруп обумовлені необхідністю розробки індивідуального методу для кожної з ознак предмету опрацювання у випадку початкового етапу досліджень, та проявляються без перерахунку завдяки одинаковій кількості дієслів у підгрупах.

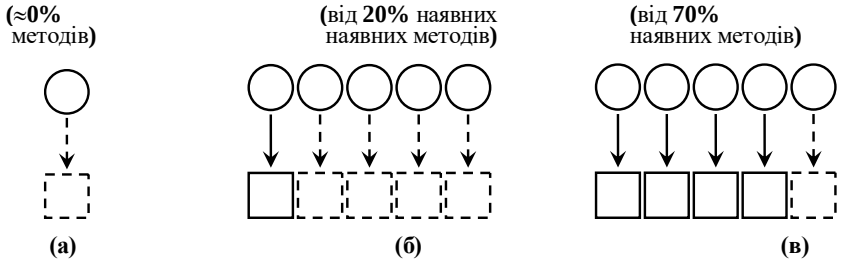


Рисунок 3. Схематичне зображення складових науково-дослідницької праці на початковому (а), проміжному (б) та завершальному (в) етапах та відповідний процентний вклад наявних методів.

Якщо дослідження перебування на *проміжному етапі*, то *друга підгрупа* дієслів вживається для опису *предмету* опрацювання, а також частини *наявних методів* та отриманих *результатів.* Складові науково-дослідницької праці, в описі яких домінують дієслова другої групи, зображені на рис. 3б суцільною лінією. В описі дослідницьких робіт на цьому етапі присутня все ще значна кількість припущень щодо допустимих методів та оцінок можливих результатів, що відображається на рис. 3б пунктиром. Максимальне значення відносного вкладу дієслів першої підгрупи в описі робіт *проміжного етапу* складає 0.38 (див. рис. 1), що дозволяє нам оцінити характерне значення вкладу *наявних методів* (від 20%).

На *завершальному етапі* досліджень методи є детально розробленими і результати подаються, в основному, у явному вигляді. Тому для опису переважаючої більшості складових науково-дослідницьких робіт цього етапу вживаються дієслова *другої підгрупи. Дієслова першої підгрупи* вживаються з метою опису, наприклад, подальших перспектив роботи. Значення максимального вкладу дієслів *першої підгрупи* на цьому етапі становить 0.13 (див. рис. 1), діапазон характерних значень вкладу *наявних методів* – від 70%.

Висновки

Запропонований метод аналізу глибини опрацювання предмета дослідженнь можна використовувати для пошуку інформації, керуючись її науковою цінністю та новизною, які часто не відповідають її хронологічним характеристикам. Сформований набір дієслів можна застосовувати для аналізу наукових праць природничого спрямування. Подальшу роботу в даному напрямку варто спрямувати на розширення набору ключових дієслів, вдосконалення методики їхнього поділу на підгрупи.

Список посилань

1. http://prb.aps.org/info/authors.html
2. Phys. Rev. B 69 (2004).
3. http://www.physnet.uni-hamburg.de/services/fachinfo/dissfb12.htm
4. http://msk.desy.de/Publikationen/

В. Балог, к.філол.н.*
Інститут української мови НАН України (Київ)
УДК 81'33

ПЕРЕДУМОВИ СТВОРЕННЯ КОРПУСУ СЛОВНИКІВ

У статті досліджено лексикографічний аспект у створенні корпусу мови. Мета проекту полягає у створенні локальної джерельної бази для будь-яких досліджень на основі словників. Вважаємо, що корпусний підхід до вирішення поставленого завдання дозволить не тільки уніфіковано формалізувати дані словників, але й ввести в систему будь-який словник, незалежно від ступеня його готовності.

Сьогодення ставить проблему економії часу. Ця теза не нова, проте вимагає створення умов для ефективності робочого процесу в будь-якій сфері суспільної діяльності. Ця ж теза провокує розвиток корпусної лінгвістики як шлях до створення „робочого місця" лінгвіста, полегшення процесу накопичення фактичного матеріалу та його систематизації. Необхідність створення корпусу текстів природної мови не потребує додаткових доказів, оскільки не тільки значно полегшує дослідницький процес, а й у суто лінгвістичному аспекті є репрезентантом мови: її синхронії та діахронії. На наше глибоке переконання корпус має бути публічним об'єктом, загальнодоступним для користувача.

Окремим аспектом функціонування корпусу вважаємо лексикографічний. Створення корпусу словників – це науково-дослідний проект, втілення якого зумовлене потребою зібрання всього масиву лексикографічних джерел української мови в електронному вигляді в одній площині з однаковим доступом до зафіксованої в них інформації. Мета проекту полягає у створенні локальної джерельної бази для будь-яких досліджень на основі словників. Корпусний підхід до вирішення поставленого завдання дозволить не тільки уніфіковано формалізувати дані словників, але й ввести в систему будь-який словник, незалежно від ступеня його готовності (це дуже актуально, оскільки деякі українські словники мають незавершений вигляд через історичні обставини), а також у співпраці з програмістами забезпечити зручну пошукову систему за всіма параметрами, представленими у словнику. Особливим завдання ставимо введення в корпус усіх видань окремо взятої праці, що дозволить накопичити фактичний матеріал для досліджень діахронії в межах одного словника.

Сучасна лексикографія поступово вступає в якісно новий етап свого розвитку і в цьому процесі перебуває в перехідному періоді, коли паралельно співіснують так звані традиційні (паперові) словники, їх комп'ютерні варіанти, та власне комп'ютерні словники. Також у цей період вчені застосовують різні – від традиційних до сучасних комп'ютерних – методики укладання словників. Безперечно, майбутнє словникарства за новітньою методикою з огляду не лише на зручність процесу, а передовсім на досконалий результат, динамічність словника, а отже і постійну адекватність стану мови.

Паралельно досить поширеним методом є використання текстів існуючих словників для формування баз даних, які в свою чергу є основою для формування електронних варіантів словників[1]. „Комп'ютерний словник і комп'ютерний варіант традиційного словника становлять результати двох напрямків роботи в сучасній комп'ютерній лексикографії..." [3: 53]. Ми вважаємо таку практику якраз свідченням перехідних процесів, що неодмінно виведуть лексикографію на новий, корпусний рівень. Це необхідний етап розвитку словникарства, на якому закріплюються в електронному вигляді надбання лексикографічної теорії та практики, а також закладаються і вибудовуються основи поняттєвого та процедурного апарату для створення справді електронних словників. У руслі останнього методу в комп'ютерній лексикографії, на нашу думку, повинні мати місце дві загальні тенденції на шляху до створення електронного масиву словників: синхронна та діахронна. Синхронний аспект комп'ютерного оброблення лексикографічних джерел полягає у опрацюванні із застосуванням комп'ютерних технологій сучасних словників різних типів. Вони функціонують або окремо у вигляді електронної „книжкової полиці" – кожен словник має

* *©В.Балог, 2006*

[1] Докладніше про це див: Карпіловська Є.А. Вступ до комп'ютерної лінгвістики. – Донецьк: ТОВ „Юго-Восток, Лтд", 2003. – С.53–72.

окрему оболонку[1], або у вигляді інтегрованої лексикографічної системи[2], яка має зведений словник з інформацією про різні характеристики слова (морфологічні, правописні, лексико-семантичні, тощо), або як база даних для створення власне електронних лексикографічних об'єктів. Діахронний аспект передбачає створення масиву словників, що вже стали історичним надбанням національної лінгвістичної традиції.

В українській науці розвиток має поки що лише синхронний напрям комп'ютерної лексикографії. Основну увагу приділяють створенню лексичних або лексикографічних масивів з використанням словників II половини XX століття (насамперед це академічні словники, створені на основі лексичної картотеки Інституту української мови НАН України). На таку роботу націлені, зокрема, співробітники відділі структурно-математичної лінгвістики Інституту мовознавства ім. О. О. Потебні, де створено Морфемно-словотвірний фонд української мови. Генеральний реєстр слів української мови зведений за матеріалами 5 словників: *Словника української мови в 11-ти томах* (К., 1970-1980 рр.), словника-довідника І. Т. Яценка у 2-х томах „*Морфемний аналіз*" (К., 1980-1981), „*Частотного словника сучасної української художньої прози*" у 2-х томах (К., 1981), „*Словника іншомовних слів*" за ред. О. С. Мельничука (К., 1974) та орфографічної частини „*Словника-довідника з правопису та слововживання*" С. І. Головащука (К., 1989). Сучасні українські словники також лягли в основу досліджень з прикладної (комп'ютерної) лінгвістики науковців і студентів Інституту філології КНУ ім. Тараса Шевченка, Київського національного лінгвістичного університету, Національного університету „Львівська політехніка", Донецького національного університету та інших навчальних закладів України, результатом яких є функціональні навчальні словники різних типів і посібники, відкриті словники, тезауруси, представлені в мережі Інтернет[3].

Важливим процесом у цьому напрямі сучасної лексикографії є теоретичне обґрунтування та практичне втілення теорії лексикографічних систем, який здійснює колектив співробітників Українського мовно-інформаційного фонду НАН України. Базовим здобутком цієї роботи є видана в електронному вигляді Інтегрована лексикографічна система „Словники України", яка також функціонує в режимі on-line на сайтах Український лінгвістичний портал (http//ulif.org.ua) та Нова мова (www.novamova.com.ua), в основу якої покладено „*Орфографічний словник української мови*" (К., 2002), „*Орфоепічний словник української мови*" (К., 2001-2003); два видання „*Фразеологічного словника української мови*" (К., 1993; 2-ге вид. – К., 1999); „*Словник синонімів української мови*" (К., 1999-2000); „*Словник антонімів української мови*" Л. М. Полюги (К., 1999); у перспективі – залучення *Словника української мови у 11-ти томах* (1970-1980 рр.). Звичайно, вказана лексикографічна система не зовсім відповідає всім запитам користувача (передовсім, у частині семантичної характеристики слова), проте зорієнтована на розбудову та вдосконалення.

У вигляді електронної „книжкової полиці" українські словники можна знайти на сайтах www.novamova.com.ua, w. (обіжний аналіз використання інтернет-ресурсу для функціонування українських словників показав, що робота упорядників лексикографічного матеріалу на сайтах, де представлені лексикографічні джерела української мови спрямована на сучасний період українського словникарства[4]. На пошуковому сайті domivka.net у розділі „українська мова" подано наступні адреси: **Нова мова** Освітньо-інформаційний Інтернет-проект оновлення української мови http://novamova.com.ua/, **Весна** Багатомовний словник, багатомовна перекладачка та словники, підбірка програм і двійкових файлів для/з підтримкою української мови, художня, публіцистична та інша література http://vesna.org.ua/, **Лінгвістичний портал** http://www.mova.info/, **Проект англо-українського словника технічних термінів**

---

[1] Як приклад можна привести функціонування сайту http//www.slovari.ru, де представлені словники російської мови різних типів, або Асоціації лексикографів LINGVO (http//www.lingvo.ru), яка також переймається накопиченням та створенням електронних словників різних типів, насамперед, перекладних.

[2] Див. Широков В.А. Інформаційна теорія лексикографічних систем. – К., 1998. – 331 с.; Інтегрована лексикографічна система „Словники України", версія 1.03. – К.. 2001-2003, – CD-видання

[3] Див., наприклад, портал www.mova.info; сайт Технічного комітету стандартизації наково-технічної термінології (http://lp.edu.ua), www.slovnyk.net .

[4] Досить розлога інформація про українські словники в Інтернеті подана в статті перекладача-фрілансера Дмітрієвої М.М.(mailto: xmas@ukr.net)18 лютого 2003 року.

Проект англо-українського словника технічних термінів http://dict.linux.org.ua/, **Уроки державної мови** Українська літературна мова давно виробила власні вимовні, наголосові, граматичні, лексичні, стилістичні норми. Уміщуємо поради, рекомендації, які саме слова чи словосполучення найдоцільніше вживати, щоб передати потрібний зміст, робимо застереження про неправильне або небажане, невдале використання тих чи інших лексем у певних значеннях, у конкретному контексті. h, **СЛОВНИК. НЕТ – Тлумачний словник** Великий тлумачний словник сучасної української мови онлайн, містить понад 207 000 словникових статей та близько 18 000 фразеологізмів. h

Наші наукові інтереси зорієнтовані на інший напрям комп'ютерної лексикографії – діахронний. Питання збереження в електронному вигляді лексикографічних джерел, укладених до 30-х років XX сторіччя, саме української мови має принципове значення: XX століття було не лише періодом розвитку та вдосконалення мовознавчої науки, зокрема лексикографії, часом виходу найбільших словників української мови, але й періодом штучного коригування розвитку мови. Таким чином, словники, що були укладені в дорадянську добу України, є об'єктом не лише для етимологічних, діалектологічних, народознавчих розвідок, реконструкції мови того періоду, але й джерелом збагачення власного словникового запасу. „Староукраїнська лексикографія – органічна частина історії української літературної мови" [4: 285]. Вибір часових меж не випадковий, зумовлений історико-політичними обставинами, оскільки саме з цього часу в історії українського словникарства почалася нова, радянська епоха.

На початковому етапі роботи наповненням для лексикографічного корпусу слугуватимуть найбільші словники, укладені з початку XIX до 30-х років XX сторіччя. У процесі вибору об'єктів лексикографічного корпусу ми спиратимемося на праці сучасних дослідників історії українського словникарства, передовсім В. В. Німчука та Б. К. Галаса, а також мовознавців XIX – початку XX ст. Цінність цих досліджень – насамперед у окресленні тенденцій та здобутків українського словникарства зазначеного періоду, ґрунтовній хронологічній, типологічній, аналітичній характеристиці словників, що значно полегшить етап аналізу формальних даних словників.

Серед українських словників дорадянського періоду найвизначнішим, знаним словником є „Словарь української мови" у 4-х томах (1907-1909 рр.) за ред. Б. Грінченка. Саме цей словник і стане першим об'єктом лексикографічного корпусу. У подальших статтях плануємо подати результати підготовчої роботи до введення словника в корпус.

Оброблення текстів лексикографічного корпусу української мови здійснюватиметься засобами SGML (Standard Generalized Markup Language) – стандартної мови узагальненої розмітки у форматі TEI (Text Encoding Initiative) – системи кодування текстів, адаптованими до особливостей текстів українських словників на основі попередніх досліджень параметричних даних.

Література

1. Борис Галас. Ф. С. Шимкевич як лексикограф і українське словникарство (кінець XVIII – початок XX ст.). – Ужгород, 1995. – 300 с.
2. Демська-Кульчицька О. М. Основи національного корпусу української мови. – К., 2005. – 219 с.
3. Карпіловська Є. А. Вступ до комп'ютерної лінгвістики. – Донецьк: ТОВ „Юго-Восток, Лтд", 2003. – 184 с.
4. Німчук В. В. Староукраїнська лексикографія в її зв'язках з російською та білоруською. – К., 1980. – 303 с.
5. Українська мова у XX сторіччі: історія лінгвоциду: Док. і матеріали / Упоряд.: Л. Масенко та ін. – К.: Вид. дім „Києво-Могилянська акад.", 2005. – 399 с.
6. Широков В. А. Феноменологія лексикографічних систем. – К., 2004.–327 с.

ЛЕКСИКОГРАФІЯ, ЛЕКСИКОЛОГІЯ: ТЕОРІЯ ТА ПРАКТИКА

Т. Цимбалюк-Скопненко, к. філол. н.*
Інститут української мови НАН України (Київ)
УДК 161.2.81.'374

КОНТАМІНАЦІЯ ЯК СТРУКТУРНО-СЕМАНТИЧНИЙ СПОСІБ ТРАНСФОРМАЦІЇ ФРАЗЕОЛОГІЧНИХ ОДИНИЦЬ (НА МАТЕРІАЛІ ПЕРЕКЛАДІВ МИКОЛИ ЛУКАША)

У статті проаналізовано специфіку фразеологічної контамінації у творчій спадщині відомого вітчизняного перекладача Миколи Лукаша, установлено особливості схрещування різних частин фразеологізмів на тлі традиційно вживаної фразеологічної одиниці в перекладних поетичних і прозових текстах.

Тривалий час фразеологічні одиниці (далі – ФО) розглядали як монолітні й семантично нерозкладані формули. Проте згодом узяла гору думка про те, що фразеологізми в індивідуально-авторському вживанні можуть підлягати найрізноманітнішим трансформаціям, які є оказіональним явищем і супроводжуються зміною денотативно-сигніфікативного та стилістичного аспекту значення. Унаслідок таких змін фразеологічний зворот отримує, „окрім закладених у ньому самому (а іноді й усупереч їм), нові естетичні та художні властивості" й уживаний „як засіб для створення найрізноманітніших художньо-стилістичних ефектів" [10: 205]. При цьому треба мати на увазі, що можливість оновлення ФО закладена в самій моделі стійкості компонентного складу й структури фразеологізму як єдності константного (інваріантного) та змінного (варіантного). Історичні зміни, еволюція мовної системи та вплив екстралінгвальних чинників так само забезпечують реалізацію широких трансформаційних потенцій ФО [8: 144]. Можливість будь-яких оказіональних перетворень ФО зумовлена насамперед їхньою структурною нарізнооформленістю, що є „властивістю фразеологізму, яка постійно викликає актуалізацію його компонентів, повертає їм у мовленні статус слова" [6: 23], а також найважливішою передумовою відхилення ФО від нормативності її форми та семантики[1]. Саме ці фактори й дозволяють майстрам слова звертатися до різноманітних стилістичних оновлень ФО „не тільки шляхом створення нових зв'язків і контекстів, а й при відносній стабільності змісту, " модифікувати, видозмінювати „форму традиційного мовного звороту" [3: 107]. Поза всяким сумнівом, не кожні перетворення, уживані в мовленні, мають потенціал до переходу в категорію нормалізованої та кодифікованої сфер. Лише за певних умов уживання модифікованої одиниці (багаторазове відтворення носіями мови, часте використання в писемній та усній літературних практиках) „може привести до закріплення нового сполучення слів у ролі „повноправної" ФО" [9: 173].

До одних з найуживаніших стилістичних трансформацій фразеологізмів у мові Лукашевих перекладів належать структурно-семантичні (*фразеологічна контамінація, фразеологічна алюзія, фразеологічний еліпсис*), які ґрунтуються на оказіональних змінах у кількісному або якісному складі ФО. Унаслідок творення названих трансформацій порушуються традиційні асоціативні зв'язки й виникає ефект невиправданого сподівання, породжений суперечністю між традиційно вживаною ФО та несподівано новим фразеологічним зворотом. Структурні, семантичні й стилістичні відмінності між ФО та її оказіональною похідною залежать від амплітуди відштовхування модифікованого звороту від прототипного зразка.

---

* © *Т. Цимбалюк-Скопненко, 2006*

[1] Мовна норма – це „сукупність мовних засобів, що відповідають системі мови й сприймаються її носіями як зразок суспільного спілкування у певний період розвитку мови і суспільства" [4: 420]. Певна річ, що структурно-семантичні перетворення ФО можна розглядати як у площині єдино можливих, обов'язкових мовних норм, що стали репрезентантами релевантних рис ідіому, так і в плані реалізації на тлі системи літературної мови конкретного періоду. Проте на теперішньому етапі розвитку фразеографії завдання опису індивідуально-авторських трансформацій у зіставному аспекті з діалектною фраземікою *всього* українського етномовного простору наштовхується на багато перешкод, передусім через брак належного фразеографічного опрацювання матеріалу з різних ареалів [Див.: 5]. Тому доводиться вдаватися до відносного означення некодифікованих джерел поповнення фразеологічної підсистеми (часто без конкретизації говору або навіть наріччя (чи іншого різновиду національної мови)).

Фразеологічна контамінація [< лат. contaminatio – змішування] полягає у схрещуванні в одному виразі двох (іноді й більшої кількості) фразеологізмів. При цьому деякі вихідні форми фразеологічних зворотів можуть цілком зберігати всі свої компоненти й структуру в модифікованій одиниці, приймаючи при цьому деякі елементи з іншої (інших) ФО, що підлягають контамінації. У разі застосування контамінації першої частини одного вихідного звороту й другої іншого кожен фразеологізм у складі трансформованої ФО якісно змінюється, утрачаючи один зі своїх складників (іноді – декілька) унаслідок поєднання. В. Ужченко, розглядаючи цей тип структурно-семантичних перетворень, зазначає: „У „класичному" варіанті контамінації відбувається не просто поєднання ФО, а взаємопроникнення словоформ, своєрідне накладання і злиття значень різних ФО внаслідок мимовільних чи свідомих зближень, смислова компресія, лаконізація – як результат опущення певних компонентів" [8: 151].

Виникає чимало питань стосовно того, у межах мови чи мовлення треба здійснювати кваліфікацію модифікованої одиниці. Деякі з дослідників уважають, що під час контамінації не утворюється нової ФО, а виникає мовленнєва одиниця, що має оказіональний характер. Вона не входить до системи мови, а є лише авторською знахідкою [7: 105]. Їхні опоненти твердять, що „фразеологічна контамінація – це об'єднання частин семантично схожих фразеологізмів, яке утворює нову ФО з тим самим або іншим подібним значенням" [2: 9]. О. Ахманова зазначає, що контамінація – це „взаємодія мовних одиниць, що стикаються в асоціативному чи в синтагматичному ряду, яка приводить до їхньої семантичної чи формальної зміни або до утворення нової (третьої) *мовної одиниці*" (курсив наш. – Авт.) [1: 206]. На думку Е. Ханпіри, уживання фразеологічних оказіоналізмів у конкретній художньо-мовленнєвій ситуації виводить їх з площини аналізу входження/невходження до системи мови, оскільки „мірилом їхньої вдалості може бути тільки відповідність завданню" перекладного твору [11: 253].

На підставі аналізу фактичного матеріалу[1], який був дібраний з мови Лукашевих перекладів, можна зробити цілком обґрунтований висновок, що контамінація ФО в розгляданих текстах здійснювана здебільшого за принципом формальної схожості, порівняно з цим зафіксовано менше випадків схрещування ФО на основі семантичної близькості (такий тип перетворень зазвичай притаманний контамінації фразеологізмів, які мають спільний компонент).

Залучені до розгляду з різних прозових і поетичних перекладних творів Миколи Лукаша приклади контамінації дають змогу виокремити три типи цього виду структурно-семантичних перетворень (відповідно до певної специфіки в кожному з цих типів можна виокремити окремі підтипи), серед них:

1. Контамінація першої частини одного звороту та другої іншого:

а) Контамінація фразеологізмів з різною семантикою:

скуштувати вершків: (*[Нестор (до Улісса):] Хто може збити Гекторову славу, // Як не Ахілл? Воно-то ніби й жарт, // Але багато між людьми заважить, // Бо тут троянці можуть* ***скуштувати*** *// Найсмаковитіших у нас* ***вершків*** (Шекспір: 354), пор.: **скуштувати стусанів** і **збирати вершки**); **запрягти в голоблі**: (*[Пандар (до Крессіди):] Куди ж це ти? Стій! Над тобою потрібне пильне око, доки призвичаїшся. Іди, іди, не бришкай, а то* ***в голоблі запряжемо*** (Шекспір: 379), пор.: **повернути голоблі** і **запрягти в ярмо**); **замкнути губи**: (*— Що се? — закричала.— Я ж думала, що ти німий! — Паніматко, — сказав Мазетто, — я і справді німий був, тільки не зроду, а недуга* ***губи*** *мені* ***замкнула****; допіру сю ніч знов мені річ пустило, за що, як умію, дяку богові складаю* (Боккаччо: 173), пор.: **замкнути рот** і **розпустити губи**); **зривати досаду**: (*Скаржилась вона* [гасконська дама] *скрізь даремно і вирішила, нарешті, вдатися до короля, але хтось їй сказав, що то марна праця, бо король той такий нікчема та плохута, що не тільки не карає по закону, як кому кривда станеться, а з соромотною боягузністю і сам зносить будь-які образи, і всяке* ***зривало*** *на ньому свою* ***досаду****, шпетячи його та безчестячи як завгодно* (Боккаччо: 72), пор.: **зривати злість** і **зганяти досаду**); **кинути голову**: (*[Естеван:]* ***Киньте голову*** *туди. // [Менго:] Наче вішальник, отерп. // [Рехідор:] Ось і королівський герб* (Лопе де Вега: 132), пор.: **кинути оком** і **підіймати голову**); **кидати думкою**: (*Бо я вже, знаєте, туди й сюди* ***думкою кидав:*** *таки не випадає панові мойму*

[1] Усі ілюстративні приклади в статті подано без правописних змін, так, як вони засвідчені в текстах перекладів М. Лукаша.

на архієпископа йти (Сервантес: 182), пор.: **кидати очима** та **майнути думкою**); **триматися зубами**: (*От побачите, як стане ваш пан імператором (а воно з усіх прикмет до того йдеться), то* ***зубами*** *за ту владу* ***триматиметься****, а в душі жалітиме й шкодітиме за тим часом, що змарнував, не бувши володарем* (Сервантес: 527), пор.: **учепитися зубами** та **триматися свого**); **язик затинається**: (*Заговорив* [Дон Кіхот], *а голос йому тремтить,* ***язик затинається****, очі огонь крешуть* (Сервантес: 300), пор.: **дух затинається** та **язик заплітається**) ; **гнути мову**: (*... знавши від Камілли про Ансельмові чати, він* [Лотаріо] *швидко дорозумівся, куди вона свою* ***мову гнула*** (Сервантес: 226), пор.: **вести мову** та **гнути лінію**); **зламати мовчанку**: (*Тоді вирішив він* [Тедальдо], *що настав уже час ту* ***мовчанку зламати****, і, як гості їли саме овочі, устав і сказав: — Одного тільки нам бракує, щоб ся наша учта була веселою* (Боккаччо: 212), пор.: **грати в мовчанку** та **зламати слово**; **узяти очі в руки**: (*... мені навіть приємно чути, що ти* [, Санчо,] *здаєшся на мою мужність, і будь певен, що швидше тебе зрадить власний подих, ніж мій бойовий дух. А тепер слідкуй за мною слідом, як тільки можеш, та* ***візьми очі в руки*** (Сервантес: 135), пор.: **узяти ноги в руки** та **гострити очі**); **навернути на пам'ять** : (*Наш гідальго виголосив свою довжелезну, хоч і цілком зайву, промову тільки тому, що горіхи, якими його потрактували,* ***навернули*** *йому* ***на пам'ять*** *золотий вік, і йому захотілось розповісти про нього пастухам* (Сервантес: 64), пор.: **навертати на думку** та **прийти на пам'ять**).

б) Контамінація фразеологізмів з однаковою семантикою:

попускати волю: (*— Це син тесляра Буде; батько-мати — люди з достатком і* ***попускають*** *йому велику* ***волю. А*** *коли б він захотів — добре б міг учитись* (Флобер: 107), пор.: **попускати віжки** і **давати волю**); **і в гадці не класти**: (*Ансельмо вгамувався і вирішив почекати до ранку, бо він* ***і в гадці*** *собі* ***не клав****, що про Каміллу щось недобре почує, так уже на її цноту вповав і впевнявся* (Сервантес: 232), пор.: **і в думці не класти** та **і в гадці не мати**); **вилетіти димом**: (*Але Леон уже тікав; у нього було таке враження, ніби його любов, яка за ці дві години каменем заціпеніла в нерухомості церкви, тепер мала* ***вилетіти димом*** *крізь цю гінку трубу* (Флобер: 226), пор.: **піти з димом** і **вилетіти в трубу**).

2. Контамінація ФО, які мають спільний компонент (у цьому випадку спільним компонентом може бути як іменник, так і дієслово):

а) Контамінація фразеологізмів з різною семантикою:

скуштувати сухарів казенних і карбачів: (*— Атож, — відповів Хінес, — довелося вже служити Богові й королю аж чотири роки,* ***скуштував*** *я і* ***сухарів казенних****, і* ***карбачів*** (Сервантес: 128), пор.: **скуштувати карбачів** і **скуштувати сухарів казенних**); **давати лупки та щипки**: (... [мовчазні катюги] *напустились на Дон Кіхота, здерли з нього укривало і простирало і почали* ***давати лупки та щупки****, а він мусив одбиватися кулаччям* (Сервантес: 557), пор.: **давати лупки** та **давати щипки**); **затемнювати та затуркувати розум**: (*Є знов немало й таких* [людей], *що думають, ніби лопата, й заступ, і груба їжа, і тяжка праця притуплюють у хлібороба тілесні похоті і* ***затемнюють та затуркують*** *його* ***розум*** (Боккаччо: 169), пор.: **затемнювати розум** і **затуркувати розум**); **потрясти й перевернути світ**: (*[апостол Петро:] А ті, що в цирку творять гімн побожний, // Як хижий тигр терзає їм тіла, // Нове учення вже несуть — братерство // Й свободу кожному, нову ідею, // Що* ***потрясе і переверне світ*** (Мадач: 86), пор.: **потрясти світ** і **перевернути світ**); **зворушити й зм'якшити серце**: (*Якщо всі ці речі не можуть* ***зворушити й зм'якшити*** *вашого камінного* ***серця****, то, може, воно злагідніє на думку про те, що скоро ваша милость ізвідси рушить* (Сервантес: 109), пор.: **зворушити серце** та **зм'якшити серце**); **линути очима й серцем**: (... [Зораїда] *то журилась, то раділа, дивлячись по тому, що малювалось на обличчі у всіх присутніх, особливо в її гишпанця, до котрого вона щохвилі* ***очима і серцем линула*** (Сервантес: 298), пор.: **линути очима** та **линути серцем**); **пекти й лоскотати серце**: (*[Пандар (співає):] Амуре, рань // Оленя й лань, // Гостріш жала // Ота стріла //* ***Серця пече й лоскоче*** (Шекспір: 377), пор.: **пекти серце** та **лоскотати серце**); **різати вухо й серце**: (*Найдужче* ***різав вухо й серце*** *тонкий голос „піаніста", піаніст надсадно вив, заплющивши очі, задерши голову, випнувши кадик* (Горький: 719), пор.: **різати вухо** та **різати серце**).

б) Контамінація фразеологізмів з однаковою семантикою:

горе і біда притисне: (*Хоч як людину* ***притисне горе і біда****, а таки вона жива, чом би їй не послухати ради, як рада та йде од щирого серця?*(Сервантес: 172), пор.: **притис-**

не горе та **притисне біда**); **стривожити і збентежити душу**: (— *То правда, — сказав Санчо, — адже навіть ступарі валюшні гуркотом своїм* ***стривожили і збентежили душу*** *такого хороброго рицаря й шукача пригод, як ваша милость* (Сервантес: 116), пор.: **стривожити душу** та **збентежити душу**); **покласти кінець і край**: (... *та річ відома (і мені, і вам), що нема таких значних посад, які дістаються без такої чи сякої базаринки. Тим-то я й жадаю, щоб за те губернаторство ви разом із паном вашим Дон Кіхотом* ***поклали кінець і край*** *сій вікопомній пригоді* (Сервантес: 521), пор.: **покласти кінець** і **покласти край**); **укарбовуватися і впечатуватися в пам'ять**: (*Як пак він* [отець казнодій] *казав? Ага, теперішні, кае, речі, що ми очима споглядаємо, живіше, кае, і міцніше врізуються,* ***вкарбовуються і впечатуються в*** *нашу* ***пам'ять*** (Сервантес: 362), пор.: **укарбовуватися в пам'ять** і **впечатуватися в пам'ять**); **полонити серце й душу**: (*[Баштовий Лінкей (до Гелени):] Сяйво миле* ***полонило // Серце й душу****, все єство; // Нещасливця осліпило // Красотою божество* (Гете: 394), пор.: **полонити серце** та **полонити душу**); **прихиляти до себе душу й серце**: (*А тим що врода має силу і владу* ***прихиляти до себе душі й серця,*** *всі щиро прагнули услужити вродливій мавританці й запобігти її ласки* (Сервантес: 245), пор.: **прихиляти до себе душу** та **прихиляти до себе серце**); **сіяти згубу й смерть**: (*Та горе, як сама звільниться, // Клекочучи, кипляча мідь! // Потоком яросно червоним // Вона прорве затворну твердь // І вогнодишучим драконом // Навколо* ***сіє згубу й смерть*** (Шиллер: 168), пор.: **сіяти згубу** та **сіяти смерть**).

3. Лінійна контамінація фразеологізмів:

вислизати в'юном з-поміж пальців: (— *Признаюсь вам по щирості, пане Дон Кіхоте, — сказав пан Лоренсо, — я весь час чигаю, чи не дасть ваша милость десь фука, та шкода мого чигання, ви мені* ***з-поміж пальців в'юном вислизаєте*** (Сервантес: 421), пор.: **вислизати з-поміж пальців** і **витися в'юном**); **пустити в діло руки**: (*[Командор (до Лаврєнсії):] Що за мова непристойна! // Покладу я самостріла // непотрібного додолу // і* ***пущу у діло руки*** *— // годі панькатися!* (Лопе де Вега: 60), пор.: **пустити в діло** і **підняти руки**); **визівнути душу в пекло**: (*Схопили* [капітана] *та й покотили перевертом по всіх лавах, од корми до самої прови — такого гарту катюзі дали, що поки до головної щогли докотили, він уже й* ***душу в пекло визівнув*** (Сервантес: 253), пор.: **віддати душу** та **у пекло лізти**); **пірнути з головою у воду**: (*А часом — це просто своєрідна лінь, намагання уникнути відповідальності, усунутися від керівництва в вирішальний момент... Ну та що зробиш!.. Зате тобі* [, Анрі,] *випала тепер нагода* ***пірнути з головою в воду****, та ще й під бурю* (Стіль: 128), пор.: **пірнути з головою** та **як у воду пірнути**); **жаль терзає душу**: (*[Гретхен:] Чи то одна сиджу я — // Тужу, тужу, тужу я, //* ***Терзає душу жаль*** (Гете: 160), пор.: **терзати душу** та **мати жаль**); **літа прибивають к землі**: (*Та не зважаючи на ті мої добрі наміри, мушу йти туди, звідки вже навряд чи коли повернусь, бо* ***літа*** *мене* ***к землі прибивають****, ще й на пухир недугую, що просто витерпу немає* (Сервантес: 127), пор.: **літа виходять** і **прибивати до землі**); **вибити мозок з голови**: (*[Терсіт (до Ахілла):] Не стільки там того настрою, скільки розстрою... Не знаю, яка там буде музика, коли Гектор* ***виб'є*** *йому* [Аяксові] ***з голови мозок*** (Шекспір: 392), пор.: **вибити з голови** та **висушити мозок**); **зігнати оскому з зубів**: (*Стоїть вона* [жінка] *отак над кадубом, увесь отвір затуливши, навчає чоловіка, а Джаннелло, що не встиг* ***зігнати*** *зранку* ***оскоми з зубів****, вирішив хоч як-небудь своє зробити, та й підскочив до неї, як негнузданий огир до гарячої кобили в парфійських степах* (Боккаччо: 392), пор.: **зігнати оскому** та **видирати з зубів**); **накинути ласим оком**: (... [Леандра] *закохалась у нього* [Вісенте], *не інакше, як з диявольського наусту, закохалась перше, ніж він устиг* ***накинути*** *на неї* ***ласим оком*** (Сервантес: 324), пор.: **накинути оком** і **ласим оком**).

Подані вище в цьому пункті ілюстрації репрезентують модель лінійної контамінації ФО з різною семантикою. Проте в залученому до аналізу матеріалі трапляються поодинокі приклади схрещування фразеологізмів, що мають однакову семантику. Наприклад: **переселятися з сього світу на той**: (*А щоб ти* [, Елено], ***з сього світу на той переселяючись****, повірила своїм очам більше, ніж моїм речам, стрибай мерщій униз, і душа твоя, вже тепер дияволові запродана, побачить, засмутяться чи не засмутяться мої очі на вид твого карколомного падіння* (Боккаччо: 473), пор.: **іти з цього світу** та **іти на той світ**).

Фразеологічне різноманіття перекладного художнього твору завжди стає однією з підвалин неповторності стилю перекладача. Індивідуально-авторські перетворення ФО своєю незвичністю та здатністю породжувати численні асоціативні зв'язки між модифікованими зворотами та кодифікованою фраземікою виразно виокремлюються серед фразеологічного фонду Лукашевих перекладів. У контамінації фразеологізмів реалізовано найтісніший зв'язок семантики, структури й стилістики. При цьому

контаміновані одиниці здебільшого не виходять за межі оказіональності, хоч вони мають усі потенційні можливості до подолання сили тяжіння цього явища.

Література

1. Ахманова О. С. Словарь лингвистических терминов. – М.: Сов. энциклопедия, 1966. – 606 с.
2. Бабкин А. М. Русская фразеология, ее развитие, источники, лексикографическая разработка. Автореф. дис. д-ра филол. н. – Л., 1968. – 26 с.
3. Гнатюк І. С. Структурно-семантичні трансформації фразеологізмів у мові сучасної української художньої прози // Білоноженко В. М., Гнатюк І. С. Функціонування та лексикографічна розробка українських фразеологізмів. – К.: Наукова думка, 1989. – С. 107 – 141.
4. Єрмоленко С. Я. Норма мовна // Українська мова: Енциклопедія. – 2-ге вид., випр. і доп. – К.: Вид-во „Українська енциклопедія" ім. М. П. Бажана, 2004. – С. 420 – 421.
5. Мілєва І. В. Евфемізація і дисфемізація у фразеотворенні говірок сходу України. Автореф. дис. к. філол. н. – Луганськ, 2005. – 19 с.
6. Мокиенко В. М. Славянская фразеология: Учеб. пособие для филол. специальностей ун-тов. – М.: Высш. школа, 1980. – 207 с.
7. Молотков А. И. Некоторые особенности употребления фразеологизмов в современном русском языке // Нормы современного русского литературного словоупотребления – М.; Л., 1966. – С. 92 – 111.
8. Ужченко В. Д. Структурно-семантичні основи авторських видозмін у фразеології // Ужченко В. Д., Авксентьєв Л. Г. Українська фразеологія: Навч. посібник для філол. фак. ун-тів. – Х.: Основа, 1990. – С. 144 – 154.
9. Шадрин Н. А. Перевод фразеологических единиц и сопоставительная стилистика. – Саратов: Изд-во Саратов. ун-та, 1991. – 220 с.
10. Шанский Н. М. Фразеология современного русского языка. – М.: Высш. школа, 1969. – 232 с.
11. Ханпира Э. Окказиональные элементы в современной речи // Стилистические исследования. – М.: Наука, 1972. – С. 245 – 317.

Джерела

Боккаччо	Боккаччо Джованні. Декамерон / Пер. з італ. М. Лукаш. – К.: Дніпро, 1985. – 661 с.
Гете	Гете Йоганн Вольфганг. Фауст: Трагедія / Пер. з нім. М. Лукаш. – К.: Дніпро, 1981. – 540 с.
Горький	Горький Максим. Сторож / Пер. з рос. М. Лукаша // Горький М. Твори: У 16-и т. – Т. 8. – К.: Держлітвидав, 1953. – С. 696 – 721.
Лопе де Вега	Вега Карпіо Лопе Фелікс де. Овеча криниця. Собака на сіні / Пер. з ісп. М. Лукаш. – К.: Держлітвидав УРСР, 1962. – 359 с.
Мадач	Мадач Імре. Трагедія людини: Драматична поема / Пер. з угор. М. Лукаш. – К.: Дніпро, 1967. – 252 с.
Сервантес	Сервантес Сааведра Мігель де. Премудрий гідальго Дон Кіхот з Ламанчі / Пер. з ісп. М. Лукаш, А. Перепадя. – К.: Дніпро, 1995. – 703 с.
Стіль	Стіль Андре. Перший удар / Пер. з франц. М. Лукаш. – К.: Держлітвидав, 1953. – 412 с.
Флобер	Флобер Гюстав. Мадам Боварі. Побут провінції / Пер. з франц. М. Лукаш. – К.: Держлітвидав УРСР, 1961. – 237 с.
Шекспір	Шекспір Вільям. Троїл і Крессіда: Драма / Пер. з англ. М. Лукаш. – У кн.: Шекспір Вільям. Твори: В 6-ти т. – Т. 4. – К.: Дніпро, 1986 . – С. 331 – 435.
Шиллер	Шіллер Фрідріх. Лірика / Пер. з нім. М. Лукаш. – К.: Дніпро, 1967. – 195 с.

Ю. Осінчук*
Інститут української мови НАН України (Київ)
УДК 161.2.81–112:373.46+246–62

ІСТОРІЯ НАЗВ ДОБОВИХ БОГОСЛУЖІНЬ

У статті досліджується українська богослужбово-обрядова лексика в діахронному аспекті, зокрема лексико-тематична група назв добових богослужінь. Історія функціонування досліджуваних лексем простежується на конкретних прикладах церковно-богослужбових книг, літургійної літератури та лексикографічних працях лінгвістичного характеру.

* *© Ю. Осінчук, 2006*

Невід'ємним складником будь-якої християнської конфесії є богослужіння, з яким співвіднесені поняття та реалії, що використовуються в богослужбовій практиці.

Об'єктом нашої студії є історія лексико-тематичної групи назв добових богослужінь. Вона відносно невелика за кількісним складом, але часто вживана в професійному та щоденному людському мовленні.

Богослужіння, що здійснюються Святою Церквою протягом доби, називаються ***добовим колом богослужінь. У***сі служби добового кола проводяться у визначений для цього час. Добових богослужінь є дев'ять: ***вечірня***, ***повечір'я***, ***полуношниця***, ***утреня***, ***перший час***, ***третій час***, ***шостий час***, ***дев'ятий час***, ***Літургія. К***рім того, до них ще належить ***всенічна*** відправа, яка здійснюється за спеціально розробленим церковним чином напередодні великих свят.

Церковний день відповідно до Писання починається звечора, тому й добове коло розпочинаємо вечірньою службою.

Вечірня – Богослужіння в складі вечірнього богослужіння, яке починає добове коло богослужінь і яке відправляють перед повечір'ям [П.: 33].

Грецька лексема hesperinos – 'вечірня служба' послужила запозиченням для старослов'янської мови, в якій на рід цієї служби вказували лексичні форми: ***вєчєрьн"***, ***вєчєрьн""***, ***вєчєрница*** [SJStSl II: 185]. Етимологічно – це похідне утворення від назви частини доби *вєчєрѝ*, яка в свою чергу походить від праслов'янського večerъ < *veker [ЕСУМ I: 366].

Зі значеннями 'церковна служба' [Ср. I: 251]; 'вечірнє богослужіння' [СДЯ I: 404] ці лексеми функціонували в перекладних церковнослов'янських та оригінальних пам'ятках Київського періоду: *И повелѣ пѣти вечерѝнюю* [XII Усп. зб.: 11 а]. *Днемѝ есмѝ переѣздилѝ до вечерни* [Поуч. Влад. Мон.: 250]. *Ѩко на ве(ч)рниц* ѡ(т) *живыи вѝ помо(щ) по~ть(с)* [XIII СДЯ I: 404].

Поступово субстантивована форма ***вечерьн""*** у світських та богослужбових текстах припинила своє вживання. Вона була повністю витіснена стягненою формою *вєчєрн#*: *Послѣдованїє колѣнопреклонныхѝ мл, твѝ на Вєчєрни* [1646 ТПМ III: 5]. *Начало Вєчєрни творитѝс# пре(д) Вєликими Враты* [1666 Служ.: 75].

Демінутивний варіант ***вечерница*** не вийшов з ужитку в пам'ятках XV – XVII ст., але він повністю втратив богослужбово-обрядове значення, набуваючи нового семантичного статусу – 'вечірня зірка, зірниця', наприклад: *[венус] то есть де(н)ница подобна и вечернници* [СУМ 16 – I п. 17 ст. IV: 7].

Семантична структура лексеми ***вєчєрн#*** дещо ускладнилася з появою видових щодо нього складених лексем –: ***вєчєрн# мала"*** – *Сѝ вєчєра сѝвершає(т) вѝ новой Цр, кви Малиѭ Вєчєрню* [1646 ТПМ II: 75]; ***вєчєрн# вєлика"*** – *Врємєни приспѣвшиѭ Вєликі# Вєчєрн#* [1681 Служ.: 2]; ***вєчєрн# повсєднєвна"*** – *Начало Повсєднєвной Вєчєрни* [77]. Таким чином, лексема ***вєчєрн#*** опинилася в ролі домінанти гіперо-гіпонімічного ряду ***вєчєрн# – вєчєрн# мала", вєчєрн# вєлика", вєчєрн# повсєднєвна".***

Семантична спеціалізація лексеми ***вечерня*** у конфесійній сфері збережена й до сьогодні. Форма ***вечерня*** увійшла в народну мову і має сучасний варіант ***вечірня. С***УМ вважає стару форму ***вечерня*** як нормативну, а форму ***вечірня*** – діалектну [СУМ I: 344]. У сучасних церковно-релігійних текстах немає чіткого унормованого вживання досліджуваної лексеми. Так, богослужбі тексти Української Греко-Католицької Церкви (далі – УГКЦ) та Української Православної Церкви Київського Патріархату (далі – УПЦКП) друкуються українською мовою зі значними домішками церковнослов'янізмів (хоча УПЦКП більш тяжіє до використання «чистої» української мови у релігійній практиці), а Українська Автокефальна Православна Церква (далі – УАПЦ) для конфесійних потреб користується лише «чистою» українською, наприклад: *Вечірня закінчується читанням молитви св. Симона Богоприємця* [ЗБ: 577]. *У другій половині дня відправляється вечерня Великої суботи* [1998 ПЦК: 51].

Павечерниця – (повечір'я). Грецька лексема ἀποδειπνον адаптувалася в старослов'янській мові у вигляді двох лексичних запозичень ***павєчєрьн"*** та ***павєчєрьница*** – 'богослужіння після вечірні; повечір'я': *Нооучи ~ ѵтрьници, и часовѡмь, и вєч(є)рни, и повє(ч)рници* [SJStSl XXV: 1].

Українська мова XI – XIV ст. успадкувала ці дві варіантні лексеми у церковно-термінологічному значенні. Лексема ***павєчєрьн"*** функціонувала зі семантикою – 'вечірня, повечір'я': *Начаша пѣти павечерню* [Ср. II: 854].

Суфіксальна лексема ***павєчєрьница*** служила назвою двох видів окремих богослужінь: 1. Велике повечір'я – 'вечірнє богослужіння, яке здійснюється у Великий піст і в деякі великі свята': *Павечерниц# по~ть(с) велика" сй пѣни~мь* [XII СДЯ VI: 334]. 2. Мале повечір'я – 'щоденне вечірнє богослужіння, яке здійснюється після вечірні': *Паве(ч)ерница ~гда вй цр, (к)ви по~ть(с) кйждо дйлжьнй ~сть сто"ти* [334]. Цим же самим терміном у пам'ятках Київського періоду називалося вечірнє богослужіння – ***вечірня велика*** та ***вечірня мала*** ('церковна служба'): *Вй ст, оу и великою соу(б)...павечерниц# нє по~тьс#* [XII Ср. II: 854]. Крім обрядової семантики за цим терміном було закріплене ще одне значення – 'вечеринка': *Бестудная словеса и плясаніе, еже вй пиру и на свадбахй и вй павечерницахй* [там само].

Подальший розвиток термінів ***павєчєрн#*** та ***павєчєрница*** проходив без змін у структурі слова та обрядовій семантиці. У період староукраїнської доби досліджувані лексеми зі специфічно функціональним значенням – 'вечірнє богослужіння' активно функціонують у пам'ятках писемності: *Павечерницѝ жє, и кй снѝ идѝщихй Мл, твы сйверша#* [1646 ТПМ І: 225]. *Начало павєчєрн_ Вєликі#* [1690 Пол.: 78].

Виявлені факти на матеріалах конфесійних текстів дають можливість констатувати процес архаїзації термінів ***павєчєрн#*** та ***павєчєрница*** в церковній практиці новоукраїнського періоду та заміну їх новим терміном ***повєчєрїє*** – 'вечірнє богослужіння, яке здійснюється після вечерні'[Д.: 436], наприклад: *Послѣдованїє Вєликогѡ Повєчєрї#* [1810 Служ.: 136].

Незважаючи на те, що церковнослов'янізм ***павєчєрница*** був витіснений на периферію терміновживання, він фіксується лексикографічними працями до сьогодні як архаїзм, пор.: ***павечерниця*** – 'коротка після обіда вечірня служба' [Ч. РМС: 245]; ***павечерниця*** [Ж.: 595]; ***павечерница*** – 'коротка відправа після вечерні' [ЕСУМ І: 366 – 367]; ***повечерниця*** те саме що повечер'я – 'частина вечірнього богослужіння, яке відправляють після вечірні або в сполученні з нею' [П.: 98].

Нова форма ***повєчєрїє*** поступово, в процесі проникнення української лексики в церковно-термінологічну галузь, набула українізованого варіанту на фонетичному рівні – ***повечір'я***, який не втратив свого обрядового значення до наших днів – 'моління після вечірні' [Ч. ПС: 69].

Слід додати, що термін ***повечір'я*** виступає в ролі домінанти гіперо-гіпонімічного ряду: ***повечір'я*** – ***велике повечір'я***, ***мале повечір'я***: *На Малому Повечір'ї співаються триспівці* [1998 ПЦК: 27]. *У вечері замість вечірні відправляється Велике Повечір'я* [ЗБ: 614].

Час (мн. ч***аси***) – ***година*** (мн. г***одини***). Загальне часове поняття у стсл. перекладних пам'ятках передавалося лексемою ***часй*** у контекстах, що реалізували її значення 'час, пора; мить, короткий проміжок часу; частина доби' [Ц.: 776].

У спеціальному церковно-термінологічному значенні 'молитва визначеного часу' у стсл. перекладних текстах функціонували дві лексеми – ***часй*** (мн. ч***асѵ***) та ***година*** (мн. г***одинѵ***), пор.: *И по>штє трєтии часй "кожє бѣаше им обѵчаи* [SJStSl XXXXIII: 483]. *И наоуч_ а оутрєн_и и годинамй обѣднѣи и вєчєрн_и* [SJStSl VIII: 414].

У церковній практиці XI – XIV ст. лексеми ***часй*** та ***година*** виступали зі значенням 'церковна служба' [Ср. III: 1482; Ср. I: 535], наприклад: *Часи велика(г) п#тй(к): ча(с) а,* [Остр. єв.: 290]. *Близь ст, го и~рдана по>щю трети>>* годиноу [СП: 95].

Церковнослов'янське фонетичне оформлення утримало за лексемою ***часй*** її стилістичний статус і до періоду староукраїнської мови. Зазначимо тут, що ця лексема виступала в ролі домінанти у гіперо-гіпонімічних відношеннях. Так, до гіперонімного найменування ***часй*** у пам'ятках церковного вжитку XV – XVII ст. виступали такі гіпонімні кореляти: ***часй пєрвїй***, ***часй трєтїй***, ***часй шєстый***, ***часй дев#тый***: *По ѡтрни жє вй совершенїи Пєрваго Часа* [1646 ТПМ І: 226]. *Трєтїй жє часй... вй оумилєнїц ср(д)ца сйвєрша#* [226]. *Шєстый часй совєрша#* [227]. *Дев#тый жє часй, такождє сй вс#кимй бл, гоговѣнїємй сйверша#* [227].

У цей період лексема ***година*** в церковно-обрядовому значенні звужує частотність функціонування. Богослужбові тексти староукраїнського часу майже не фіксують даної лексеми. Деякі лексикографічні джерела подають лише похідні лексеми, наприклад: ***годинникй*** – 'книжка до набоженства з церковними часами (годинами)': *Годинникй – часословй* [XVII Т.: 541].

Спостереження над функціонуванням лексичної пари ***час*** (мн. ч***аси***) – ***година*** (мн. г***одини***) в богослужбових текстових джерелах дають підстави стверджувати, що вони

представлені в сучасній українській мові без переінтеграції обрядової семантики: *Малий відпуст на Часах* [1995 Служ.: 253]. *З Великоднього дня і весь тиждень, аж до суботи, Години Церковні й Повечір'я відправляються* [1995 Молит:. 330].

У богослужбово-релігійній практиці використовуються кілька видів ***часів***, а саме: ***часи пасхальні*** та синонім до даного гіпоніма ***часи великодні***; ***часи царські*** – синонім ***часи великі***; ***часи великопісні***, ***часи повсякденні***: *Часи повсякденні*, *Часи Великі*, *Часи проміжні* [Гол.: 258]. *У Велику п'ятницю Літургії не буває, а чиняться Царські часи* [ЗБ: 618]. *До Великопісних часів буває благовіст постовий у менший дзвін* [637]. *Великодні часи на святу Пасху читаються весь Світлий тиждень замість ранішніх і вечірніх молитов* [293].

Всенічна – (***всеношна***). Особливість даного богослужіння в тому, що воно починається увечері напередодні шанованих свят і триває усю ніч до світання. Ця служба складається із вечірні, утрені та першого часу [Ч. ПС: 31].

У давній гр. мові з церковно-термінологічним значенням 'всенічна служба' виступала лексема pannychios [B.: 930].

У старослов'янських перекладних пам'ятках грецькому pannychios відповідала слов'янська лексема ***обнощьница*** – 'всенічна служба': *Обноштьниц@ твориш– вѝ намѣнєнѣи црѝкѝви* [Ц.: 395].

Пам'ятки давньоукраїнського періоду на рід цієї служби фіксують стійку термінологічнау сполуку ***всенощное бдѣніе. Л***ексема ***бѝдѣни~*** запозичена із старослов'янської мови, яка поєднувала в своїй семантичній структурі значення 'пильність, увага; неспання' [Ц.: 103]: *И сицево бѝдѣни~ и несѝпани~ по вьс# нощи* [XII Усп. зб.: 56 в].

Оскільки богослужіння здійснювалося протягом ночі, то дістало відповідну назву – ***всєнощьно~ бдѣніе*** – 'церковна служба': *Повѣле архієпископѝ быти бдѣнію всенощному* [Ср. I: 197].

Церковна спеціалізація такого богослужіння в пам'ятках раннього періоду зумовила утворенню новим термінологічним сполукам – ***всенощьно~ пѣни~***, ***нощьно~ пѣни~***, наприклад: *И сѝтвори вѝ тои клѣтѝцѣ всенощьно~ пѣни~* [XII Усп. зб.: 19 б]. И *тако мало посѝпавѝ вѝстан"ше пакы на нощьно~ пѣни~* [38 б].

У процесі історичного розвитку обрядової лексики церковна термінологічна сполука ***всєнощьно~ бдѣніе*** могла розщеплюватися, утворюючи два терміни – ***всенощноє*** та ***бдѣніє***, при чому конфесійна семантика цих термінів не втрачалася: *Почали служити всенощноє* [СУМ 16 – I п. 17 ст. V: 30]. *Бдєнїє: то єсть прє(з) цѣлую ночѝ чутьє, мл, твы прє(з) всю ночѝ в'чулости бє(з) спан# в цр, кви ѡ(т)правоваными, а вєдлу(г) чину цр, ковного называєтс# всєнощноє бдєнїє* [СУМ 16 – I п. 17 ст. II: 33].

Ще лексема ***бдѣніє*** в церковно-обрядовому значенні виступала зі семантикою 'ревна молитва у нічний час': *Вмѣсто бдєнїа, пѣ(с)ни, и мл, твы, и торжества дх, о(в)наго.* [СУМ 16 – I п. 17 ст. II: 39].

Требник П. Могили фіксує лише термінологічну сполуку ***всєнощьно~ бдѣніе***: *Потом жє ѡбычноє Всєнощноє Бдѣнїє* [1646 ТПМ II: 75]. Форма цього термінологічного сполучення функціонувала майже до кінця XIX ст., оскільки до цього часу в церковній практиці використовувалася церковнослов'янська мова: *Всєнощноє жє Бдѣнїє, и оутрєни воскр(с)ны# бываютѝ* [1806 Служ.: 2].

На кінець XIX – поч. XX ст. у церковній практиці для нічного богослужіння почала використовуватися лексема ***всенощна. У***перше форму **усеношна** зафіксовано Б. Грінченком [Гр. IV: 355].

У сучасній українській мові цій лексемі властива варіантність на фонетичному рівні: ***всенічна***, ***усенічна***, ***всеношна***, ***всенощна***, ***всенощне***, ***всенічне***, ***всеношне***: *Всеношна складається з великого повечір'я* [1998 ПЦК: 29]. *Всенічна починається Великим Повечір'ям* [1993 ПЦК: 42]. *Чин великої вечірні із всенічним* [2002 Тип.: 8].

СУМ нормативним варіантом вважає лексему ***всеношна*** (***усеношна***) [СУМ I: 764].

Зрідка у церковній літературі зустрічається застаріла архаїчна форма ***всенічне бдіння***: *Всенічне бдіння починається великим повечір'ям* [ЗБ: 626].

Процес проникнення української лексики в церковно-термінологічну галузь спричинив появу нових лексичних відповідників до церковнослов'янізмів ***всенічна*** — ***цілонічна*** [I.: 7]; ***всенічне бдіння*** – ***нічне чування***, ***всенічне пильнування***: *Цілонічне чування перед кожною неділею* [Гол.: 149]. *Нічні чування відбуваються в навечір'я неділь та більших свят* [78]. *Таріль для всенощного пильнування* [Яв.: 206].

Полуношниця. До нічних богослужб належить церковна відправа ***полуношниця***, яка дістала назву від часу її виконання – ‘церковна служба, яка здійснюється опівночі’ [Ч. ПС: 71].

Дослідник церковної термінології О. Горбач зазначає, що лексема ***полуношниця*** походить від гр. menosyktion [Горб.: 60].

У вузькоспеціальному церковно-термінологічному значенні ‘церковна служба, яка здійснюється опівночі; частина всенічної’ лексема ***полѵнощьница*** та її фонетичний варіант ***полѵночьница*** засвідчуються в писемних пам’ятках давньоруськоукраїнського періоду: *Полунощница и заутреняя сѝ псалтирею и сѝ прочимѝ послѣдованіемѝ должни суть творити* [1419 Ср. II: 1140]. *Потомѝ полуночницю пѣти* [XIV/XV Сл.РЯ XVI: 270].

У світських пам’ятках цього ж хронологічного періоду на рід такого богослужіння інколи вживалося термінологічне сполучення ***полѵнощьно~ пѣни~***, наприклад: *Ѡнѣмѝ мн"щемѝ "ко братий полѵнощьно~ пѣни~ сѝвьрьшающемѝ* [XII Усп. зб.: 46 г].

Спеціальні богослужбові тексти староукраїнського часу фіксують досліджувану лексему з тим же обрядовим значенням, яке було притаманне їй у текстах раннього періоду: *Вѝ своє врєм# прєждє свитающѵ дню, полѵнощницѵ сѝверша#* [1646 ТПМ I: 226].

Крім того, на той час у церковній практиці використовувалося два види цієї служби, на що вказували такі гіпоніми: ***полѵнощница повсєднєвна***, ***полѵнощница нєдєльна***, наприклад: *Послѣдованїє полѵнощници нєдєлны" и повсєднєвны"* [1702 Лит.: 41].

Лексикографічні джерела пізнішого періоду розвитку української мови фіксують модифікований варіант – ***полуничница*** ‘всенічна служба’ [Б-Н: 452].

У сучасній церковній практиці до церковнослов’янізма ***полуношниця*** вживаються українські лексичні відповідники: ***опівнічниця*** рідко ***північна***, ***опівнічна відправа***, ***опівнічна***: *Малий відпуст на протязі цілого тижня на Повєчір’ї, Опівнічній і на Часах* [1995 Служ.: 253]. *Вечірня, Повечір’я і Опівнічна відправа*...[399]. *Ранішнє Богослужіння – Північна відправа* [399].

Сучасна богослужбова та релігійна література фіксує зрідка церковнослов’янізм ***полуношниця*** та його фонетичний варіант ***полуночниця***: *Полуношниця повсякденна* [2000 Ч.: 5]. *Воскресна Полуночниця* [Ф. Ю: 86].

Утреня. Частина ранкового богослужіння, яке відправляють після опівнічниці перед першим часом або в сполученні з ними [П.: 134].

Дослідник церковної лексики О. Горбач лексему ***утреня*** термінологічно тлумачить як еквівалент грецького церковного терміна heôthinê proseuchê ‘рання молитва’ [Горб.: 60].

М. Федорів ***утреню*** тлумачить від гр. *ортрос* (ορθρος) – ‘рання служба’ [Ф. М.: 42]. Етимологічно слово ***утреня*** утворене суфіксальним способом від праслов’янського етимона *утро*: *утреня* < *утро* < ст.ук. о***утро*** < стсл. о***утро*** < псл. *ustro, історія якого своїм початком сягає в праіндоєвропейську добу [Ог. II: 88].

Грецькій лексемі ορθρος на старослов’янському мовному ґрунті відповідали лексеми: ***оутросѝ***, ***оутрьн""***, ***оутрьница***, ***заоутрьн"***: *Єва(нѝ)г(єли~) ѡт(ѝ) лоук(ґ) на ѵтросѝ. На ѵтрьни с(в#)таго п#тѝка. Нач#токѝ оутрьници* [Ц.: 753]. *Оутрѵ жє бывшю вѝзвониша заоутрєню* [SJStSl XII: 662].

Через перекладені священно- та церковно-богослужбові книги, які використовувалися у церковному посередництві давньоукраїнського періоду, лексеми ***оутрьн"***, ***оутрьн""***, ***оутрьница*** та ***заоутрьн"*** із обрядовим значенням – ‘ранішнє богослужіння’ увійшли в українську мову XI – XIV ст.: *На оутрьни про(к) гла(с) д,* [Остр. єв.: 283]. *На заоутрьнии вѝ вѝскрѣсѝнахѝ* [10]. *Коньчавѝшю ~моу оутрьнюю начатѝ молитис"* [XII Усп. Зб.: 11 б]]. *Оутрьницю пѣти* [Ср. III: 1319].

Старослов’янська форма ***оутросѝ*** у пам’ятках давньоукраїнського періоду нами не виявлена.

І. Срезневський подає ще один термін ***завтрен"*** – ‘заутрення, церковна служба’: *Первое изучитѝ ему вечерню..., а завтреня также* [Ср. I: 903].

Церковно-обрядове значення лексеми ***оутрьн"*** розвивалося і на подальших етапах формування обрядової термінології. У староукраїнських богослужбових текстах аналізована лексема вживалася у двох фонетичних варіантах – ***оутрн#*** та ***оутри#***: *Оутрию жє соверша#* [1646 ТПМ I: 226]. *На простой жє Оутрни* [1681 Служ.: 66].

На позначення назви церковної служби із тією ж самою семою, що закріплена за лексемою ***оутрн#***, у пам’ятках досліджуваного періоду фіксується форма ***завтрен#***:

Застали тамѝ священника… у олтаря, завтреню отправуючого [СУМ 16 – І п. 17 ст. IX: 238].

У «Словнику української мови» П. Білецького-Носенка зафіксовано форму ***утредня***, який вказує на ранішню службу [Б–Н: 368].

Б. Грінченко вперше подає без модифікацій обрядовий термін ***утреня*** – ‘утреня, заутреня’ [Гр. IV: 366].

Спостереження над функціонуванням досліджуваної лексеми в лексикографічних та текстових джерелах XX – поч. XXI ст. дають підстави стверджувати, що лексема ***утреня*** виступає в ролі домінанти гіперо-гіпонімічного ряду: ***утреня*** – ***утреня велика***, ***утреня великодня***, ***утреня єрусалимська***, ***утреня з поклонами***, ***утреня повсякденна***, ***утреня страсна*** тощо. Наприклад: *Ця частина є властива великій Утрені* [Гол.: 200]. *Чин повсякденної утрені* [1992 Тип. І: 41]. *Відправляється утреня страсна з читанням 12 євангелій* [1998 ПЦК: 51]. *Починається пасхальна утреня* [52]. *Правиться Єрусалимська Утрення* [1993 ПЦК: 59]. *Відправляється в наших церквах Утреня з поклонами* [УХС: 172].

Процес проникнення української лексики в церковно-термінологічну галузь посприяв появі нових українських відповідників – ***раня*** та ***ранішня***: *Наступне богослужіння пасхальної ранньої складається переважно зі співу канону* [ЗБ: 621]. *Решта все далі як в чині Ранішньої* [2002 Служ.: 30].

Літургія – обідня. Центральне богослужіння добового кола, під час якого відбувається Таїнство Євхаристії [П.: 70].

На позначення цього богослужіння в грецькій мові функціонувала лексема λειτουργία. За структурою грецька лексема λειτουργία – це складне слово, утворене від основи λειτ – пов’язаної з λᾱός – ‘народ, люди, натовп, колектив’, і – ουργία, пов’язаної з εργον – ‘робота’, що буквально означало ‘громадська робота, громадська служба’ [ЕСУМ III: 272]. Первісно так називалися всі громадські богослужіння, пізніше цим словом стало позначатися найважливіше Богослужіння – спомин про Таємну вечерю Ісуса Христа.

Грецький термін λειτουργία – ‘богослужба’, що виступає віддієслівним утворенням від гр. λειτούργεῖν ‘виконувати богослужбу [ESJStSl VII: 428 – 429] запозичила старослов’янська мова у формі ***литоурги"***: *Ищи вь ч(ь)тє(ни~) на литурги ѡт(ѝ) полꙋ* [SJStSl XIV: 124].

Давньоукраїнським відповідником цього старослов’янського слова в його церковно-термінологічному значенні виступала лексема ***литоурги"*** – ‘християнське церковне богослужіння з причащанням; обідня’ [СДЯ IV: 406]: *И ѡ(т)пѣвши литургию ѡбѣдашти брать" на сѝвокупе кождо с бо"рг своими* [ПВЛ: 182]. *Вѝ ср(д)ѣ. сгропоу(с). нє(д). вє(ч). на литургии ишедѝ ди"(к) сѝ кадильницею* [XII Служ. Варл.: 20].

Паралельно із терміном ***литоурги"*** у цей хронологічний період пам’ятки фіксують термін-дублет ***обѣдьн"*** – ‘основне богослужіння в християнській церкві, яке здійснюється в першій половині дня, до обіду; літургія’ [СДЯ V: 576]. Етимологічно стсл. о***бѣдьн""*** суфіксальний похідний (суф. **-ьнѝ** / **-ьнь**) від псл. *obědú [ЕСУМ IV: 133].

Дана лексема активно функціонувала у пам’ятках давньоруськоукраїнського часу: *А иже се на выходѝ вышедѝ по(п) на ѡбѣдни и на ве(ч)рнии* [1280 Ср. І: 456].

Таким чином, гр. λειτουργία передане грецизмом ***литоурги"*** і продубльоване слов’янською лексемою ***обѣдьн"*** саме в церковно-термінологічному значенні. Гр. λειτουργία ‘богослужба’ слов’янські перекладачі передали власномовною лексемою ***обѣдьн"***, отже, тут, на їхню думку, існувала повна семантична відповідність між грецьким та автохтонним словом.

Дублетна пара ***литоурги"*** – ***обѣдьн"*** функціонувала протягом усієї історії розвитку церковно-обрядової термінології: *Подобаетѝ вѣдати, "кѡ аще бидетѝ лѵтиргїа* [1646 ТПМ І: 69]. *В нашой церкви: Обѣдня. Вечерня. Повечерница* [1608 СУМ 16 – І п. 17 ст. IV: 7].

Пам’ятки XV – XVIII ст. до даних термінів-дублетів фіксують синонім ***служба божая*** – *Слижби Бж, ию на погребѣ соборне служити* [1646 ТПМ І: 751].

Порівняймо функціонування зазначеної пари та її синоніма в сучасних богослужбових та літургічних текстах: *Євхаристична Літургія розпочинається приношенням дарів* [1995 ХМ: 24]. *Служба Божа в Страсний Четвер* [1993 ПЦК: 58]. *Найважливіше богослужіння – літургія, чи обідня* [ЗБ: 25].

Слід зазначити, що існують три чини ***Літургії***, а саме: ***Літургія Василія Великого*** (*Літургія св. Василія Великого починається з великої вечерні* [1998 ПЦК: 28]), ***Літургія***

Івана Золотоустого (*На Літургії св. Івана Золотоустого* *антифони звичайні* [1993 ПЦК: 43]) та ***Літургія Ранішеосвячених Дарів. О***стання – специфічна богослужба, яка відправляється у середи та п'ятниці Великого посту та в понеділок, вівторок і середу Страсного тижня, на якій не здійснюють Таїнства Євхаристії, а вірних причащають Напередосвяченими Святими Дарами [П.: 71].

У сучасній церковній практиці використовується кілька варіантів вживання цієї лексеми, наприклад: ***Літургія Напередосвячених Дарів***, ***Літургія Попередньосвячених Дарів*** та ***Літургія Ранішеосвячених Дарів***, наприклад: *До особливостей Літургії Напередосвячених Дарів належить співання стихів псальма 141-го* [УХС: 56]. *Божественна Літургія Ранішеосвячених Дарів* [1995 Служ.: 189]. *Далі звичайний порядок Літургії Передосвячених Дарів* [1998 ПЦК: 42].

Отже, лексико-тематична група назв добових богослужінь відносно невелика за кількісним складом, але часто вживана в професійному та щоденному людському мовленні. Формування її було складним багатоплановим мовно-культурним та історичним процесом.

Джерела

Б-Н Білецький-Носенко П. Словник української мови. / Підгот. до вид. В. Німчука. – К., 1966. – 423 с.
В. Вейсман А. Греческо-русский словарь. – Спб., 1897. – 1368 с.
Гол. Головацький Р. Пояснення Богослужень. – Рим, 1979. – 339 с.
Горб. Горбач О. З історії української церковно-музичної термінології. / Олекса Горбач. Термінографічна серія. За ред. Б. Рицар, Р. Микульчук. – Львів, 2004. – С. 58 – 78.
Гр. Словарь української мови / За ред. Б. Грінченка. – К., 1907. Т. 1 – 4.
Д. Дьяченко Т. Полный церковнославянский словарь. – М., 2001. – 1120 с.
ЕСУМ Етимологічний словник української мови. В 7 т. – К., 1982. – 2003.
Ж. Желехівський Ж., Недільський С. Малорусько-німецький словник. – Л., 1886. – 1122 с.
ЗБ Слобідський С. Закон Божий. – К., 2003. – 642 с.
І. Іларіон. Церковний словничок. – Холм, 1942. – 25 с.
Лит. 1702 Литургиарион. – Л., 1702.
Молит. 1975 Молитвослов. – Рим, 1975.
Молит. 1995 Молитвослов. – К., 1995.
Ог. Огієнко І. Етимологічно – семантичний словник української мови. В 4 т. / За ред. Юрія Мулика-Луцика. – Вінніпег, 1979. Т. 1 – 4.
Остр. єв. Остромирово евангелие 1056 – 1057 г. // Изд. А. Востоков. – Спб., 1843. – 623 с.
П. Пуряєва Н. Словник церковно-обрядової термінології. – Львів, 2001. – 156 с.
ПВЛ Повесть временных лет по Лаврентьевскому списку. // Полное собрание русских летописей. – Л., 1962. – Т. 1. – С. 1 – 286.
Пол. 1690 Полуустав. Содержай в себе дневную и ношную службу по преданию и чину церковному. – К., 1682.
Поуч Влад Мон. Поучение Владимира Мономаха до 1125 г. // ПВЛ. 1962. С. 240 – 262.
ПЦК 1993 Православний церковний календар // УАПЦ. – Львів., 1993. – 90 с.
ПЦК 1998 Православний церковний календар // УПЦКП. – К., 1998. – 90 с.
СДЯ Словарь древнерусского языка (XI – XIV вв.) / Гл. ред. И. Аванесов. – М., 1989. Т. 1 – 6.
Сл.РЯ Словарь русского языка XI – XVII вв. – М.: Наука, 1975. Вып. 1 – 26.
Служ. Варл. XII Служебник Варлаама ХутинськогоXII в. / Горский А. Описание славянских рукописей Московской синодальной библиотеки. – М., 1869. С. 5 – 11.
Служ. 1666 Служебник. – Львів, 1666.
Служ. 1681 Служебник. – Львів, 1681.
Служ. 1806 Служебник. – К., 1806.
Служ 1810 Служебник. – К., 1810.
Служ 1860 Служебник. – К., 1860.
Служ.1995 Служебник. – К., 1995.
Служ. 2002 Служебник. – Львів, 2002.
СП Синайський патерик / Изд. подгот. В. Голышенко, В. Дубровина. – М.: Наука, 1967. – 401 с.
Ср. Срезневский И. Словарь древнерусского языка. – М., 1989. Т. 1 – 3.
СУМ Словник української мови. В 11 т. – К.: Наук. думка, 1970 – 1980. Т. 1 – 11.
СУМ 16 – І п. 17 ст. Словник української мови 16 – І п. 17 ст.: У 28-ми вип. / НАН України Інститут українознавства ім. І. Крип'якевича. – Л., 1994 – 2003.
Т. Тимченко Є. Історичний словник українського язика. – К., 1930 – 1932. Т. 1.
Тип. 1992 Типик Української Католицької Церкви / Підгот. до видання о. Ісидор Дольницький. – Рим, 1992. – 619 с.

Тип. 2002	Типик Української Католицької Церкви / укладений о. Ісидором Дольницьким. – Львів, 2002. – 581 с.
ТПМ 1646	Могила Петро. Евхологион, албо молитвослов, или требник. – К., 1646. – Ч. 1 – 2.
Усп. зб.	Успенский сборник XII – XIII вв. / Изд. Подгот. О. Князевская. Под ре С. Коткова. – М., 1971. – 752 с.
УХС	Іванків Є. Український християнський схід. – Чикаго, 1992. – 245 с.
Ф. М.	Федорів М. Обрядові співи української церкви Галицької землі. – Філядельфія, 1983. – 290 с.
Ф. Ю.	Федорів Ю. Пояснення церковних Богослужень і Св. Тайн. – Львів, 2004. – 159 с.
ХМ 1995	Християнський молитовник. – Красилів, 1995.
Ц.	Старославянский словарь. Под. ред. Р. Цейтлин, Р. Вечерки и Е. Благовой. – М., 1990. – 842 с.
Ч. 2000	Часослов. // Видавничий відділ УПЦ Київського Патріархату. – К., 2000.
Ч. РМС	Чопей Л. Русько- Мадярський словарь. – Будапешт, 1883. – 492 с.
Ч. ПС	Чопенко В. Православний словник. – К., 1995. – 110 с.
Яв.	Яворницький Д. Історія запорізьких козаків. – Львів, 1990. – Т. 1. – 319 с.
ESJStSl	Etymologicky slovnik jazyka staroslovenskeho. – Praha, 1997.
SJStSl	Slovník jazyka staroslověnského. – Praha : Academia, 1966.

М. Щипська *
Інститут української мови НАН України (Київ)
УДК 161.2.81’374

ОСОБЛИВОСТІ ВИКОРИСТАННЯ ФРАЗЕОЛОГІЗМІВ У МОВІ СУЧАСНОЇ ПОЕЗІЇ

У статті проаналізовано деякі особливості використання фразеологізмів у поезіях сучасних авторів. Зокрема, увагу приділено семантичним трансформаціям фразеологізмів, граматичним та лексичним замінам фразеологічних одиниць[1].

У мові сучасної української поезії широко використовуються фразеологічні одиниці як у загальноприйнятій формі та значенні, так і різні способи їх структурних та семантичних трансформацій. Під трансформацією ФО розуміємо „стилістичний засіб, який полягає в цілеспрямованому оновленні семантики, модифікації структури традиційних фразеологічних одиниць” [1: 8].

Готовий фразеологізм використати простіше, аніж пристосувати його до стилю викладу, надати йому індивідуальної неповторності [6]. Це вимагає від автора великої майстерності, тонкого відчуття, глибокого знання мови. Саме тому у своїй роботі ми приділили увагу окремим різновидам трансформації фразеологізмів, що зустрічаються у поезіях сучасних авторів.

Зокрема серед власне семантичних трансформацій ФО виділяємо ті, котрі розширюють свою семантику шляхом розширення сполучних можливостей*: І було нам звелено:/* ***закопайте живцем у землю****/ ваші ліси і оселі, / закопайте чим глибше в непам’ять/ ваше зів’яле життя —/ і зійдіть з цього краю!* (Кордун: 44—45); *Ти ще* ***не був*** *з епохою* ***на „ти”****, / Іще вона в тобі вогнем мовчала./ І тільки щем святої висоти —/ яка у снах молилась і літала* (ЛУ, 29.09.05: 5). ФО **закопати живцем у землю** *кого* „у важких муках згубити, довести до загибелі, нещадно покарати кого-небудь”(СФУМ: 245), ***бути на ти*** *з ким* „добре розумітися на чомусь, досконало знати якусь справу” (СФУМ: 48) традиційно вживаються для характеристики людини, її вчинків. Заміна іменника із семантикою особи в об’єктному оточенні фразеологізмів іменниками із семантикою неістоти сприяє збільшенню абстрактності ФО і, отже, призводить до розширення їх сполучних можливостей і власне семантики.

Трапляються поодинокі випадки подвійної актуалізації ФО, тобто поєднання в одному стійкому виразі як буквального, так і переносного значення: *Міркує мовчки дід, як то сказать хлоп’яті/ сказати правду, та не напрямки.../ Іржуть на луках жеребці завзяті —/ ото її* [вороно́ї] *гривасті парубки. /Вона із ними там гуляла довго —/ лошатко привела від люботи.../ Кусає ворона́ за те гнідого, / що тільки* ***воза*** *треба з ним* ***тягти*** (Забаштанський: 78). По-особливому організуючи контекст, поет ставить традиційну ФО ***тягти воза*** „покірно миритися, погоджуватися з виконанням важкої або невдячної

* © *М. Щипська, 2006*
[1] Надалі — ФО.

роботи, одноманітних обов'язків" (СФУМ: 729) у такі мовленнєві умови, в яких реалізуються різні сторони її семантики. Вузький контекст викликає в читача відчуття семантичної залежності ФО від її генетичного джерела — йдеться про коней, які справді тягнуть воза. Проте макроконтекст утримує в свідомості справжню причину поведінки вороної, на що вказує семантика ФО.

Одним із поширених способів структурно-семантичних трансформацій ФО в сучасній українській поезії є оказіональні зміни компонентів, серед яких виділяємо граматичну та лексичну заміну.

Розділяємо думку В. М. Мокієнка та В. Д. Ужченка, котрі застерігають від перебільшення фактора авторського фразеотворення, авторської трансформації ФО та недооцінки варіювання як характерної властивості фразеологізму [2: 19—23; 5: 147]. Наголошувала на важливості розрізнення узуальної варіантності — нормативної (мовної), що фіксується словниками, і оказіональної — одиничної, індивідуальної, що є фактом мовлення і Л. Г. Скрипник [4: 128].

Граматичні зміни ФО зумовлені здатністю фразеологізму вступати в граматичні зв'язки з іншими мовними одиницями тексту, тобто виступати не самочинно, ізольовано, а як структурні елементи контексту.

Граматично змінені ФО, зафіксовані у мові сучасної поезії, різняться:

—заміною неозначеної форми дієслівного компонента на особову: *А в Мазепи вік прожитий, почуття гірке, полинне:/ скільки рідний край нещасний розтинали сили злі!/ І прикинув він, що Карло України не поглине, / а Петруха-розпинатель вже* ***стира з лиця землі*** (Білоус: 36). Традиційна ФО ***стерти з лиця землі*** *кого, що* „знищити, розгромити" (СФУМ: 267) у контексті поезії, вступаючи у граматичні зв'язки з іншими членами речення, змінює у дієслівному компоненті неозначену форму дієслова на особову. При цьому зрушень у семантиці фразеологізму та загалом висловлювання не відбувається.

Вводяться ФО в контекст і шляхом граматичного оформлення їх у вигляді дієприслівникового звороту: *Та по любощах накинув свиту, / Глянув ще в ті зорі чарівні —/ „****Звівши*** *мужа шлюбного* ***зі світу****, / А чи ж буде вірною мені?"*(Забаштанський: 80); *Що за скрегіт вночі?/ Углядаюся в темінь погрозну:/ Двоє чадо з надгробка здирають/ дідівську зорю, / І здають дефіцитну —/ за мірою циників — бронзу, / І на виручку колються, /* ***сходячи в землю сиру...***(ЛУ, 20.10.05: 5). Пристосовані до контексту оказіональні ФО ***звівши зі світу, сходячи в землю*** повністю передають значення кодифікованих ФО ***зводити зі світу*** *кого* „доводити кого-небудь до загибелі, до смерті" (СФУМ: 259), ***сходити в землю*** „умерти" (СФУМ: 264);

—заміною форми числа фразеологізму: *Ну, доки ж будем світ смішити, /* ***тернові*** *множити* ***вінки****?/ Та одне одного страшити/ за перестрашені віки* (Забаштанський: 40); *Які вкладав він[В. Семенюта] почуття, / які тони і гами!/ Заповнював усе життя/ в мистецтві* ***білі плями*** (Білоус: 104). Введення в тканину поетичного твору ФО ***терновий вінок*** „муки, страждання" (СФУМ 110), ***біла пляма*** „недосліджене, маловивчене питання" (СФУМ: 522), заміна однини множиною уможливлює існування рими та підсилює значення фразеологізму. Необхідно зауважити, що в ілюстраціях до ФО ***біла пляма*** у СФУМ фіксується форма множини, але чомусь не виноситься в реєстр.

У сучасних поетичних текстах поширена лексична заміна компонентів ФО, яка у більшості випадків супроводжується формальними змінами (їх до уваги не братимемо), що відображають граматичні зв'язки фразеологізму з іншими членами речення. Досліджений матеріал засвідчує цілеспрямовану субституцію одного, декількох та усіх компонентів ФО.

При цілеспрямованій субституції одного компонента ФО замінники і замінювані компоненти виражаються:

а) дієсловом: *У чім стояв, без жодних статків/* ***пройшов над прірвою*** *в житті./ Але залишив для нащадків/ сторінки творів золоті* (Білоус: 97). Широкий контекст умотивовує заміну компонента ***опинятися***, вираженого дієсловом, кодифікованої ФО ***опинятися над прірвою*** „потрапити в безвихідне становище, бути на краю загибелі" (СФУМ: 468) на слово вільного вжитку ***пройшов. С***убституент-дієслово актуалізує й конкретизує загальне значення ФО щодо життєвого шляху відомого українського письменника Івана Багряного;

Ця музика ще зрідка дістає, / ***з колін підводить****, вмерти не дає, / по бездоріжжю водить за собою./ І степ — лиш мій. І небо — лиш моє./ Їх музику душа, як*

брагу, п'є, / і — п'яна — пада в хащі травостою (Базилевський: 66). Оказіональна ФО ***з колін підводить*** із значенням „підтримувати, давати сили, надію" утворена шляхом інверсії та лексичної заміни компонентів ФО ***вставати з колін*** „звільнятися від гніту, поневолення" (СФУМ: 132). Субституція компонентів загальномовної ФО вмотивовується контекстом;

Хто владар сьогодні? хто зомбі?/ чом голос Архангела втих?/ Старці — ніби атомні бомби, / ***німіє душа*** *коло них* (Поклад: 30). Субституція компонента ***холоне*** загальномовної ФО ***холоне душа*** *чия, у кого і без додатка* „хто-небудь дуже непокоїться, переживає" (СФУМ: 639) словом вільного вжитку ***німіє*** використовується поетесою з метою посилення значення фразеологізму. Дієслово-субститут ***німіє*** значно експресивніше за своїм лексичним змістом від традиційного компонента ***холоне***;

Хапаю *натхнення* ***за поли****:/ „Пожди — це ж іще не тюрма!.."/ Розгойдує серпень дзвінкі виднокола, / на крила мене підійма* (ЛУ, 27.10.05: 5). Ефект посилення значення традиційної ФО ***смикати за поли*** *кого, ірон.* „настирливо, невідступно просити, вимагати у кого-небудь щось" (СФУМ: 670) досягається поетесою за допомогою заміни компонента ***смикати***, вираженого дієсловом, на більш влучний субститут-дієслово ***хапаю. Р***азом з тим відбувається розширення сполучуваності — ФО ***смикати за поли*** *кого* традиційно вживається з іменником на позначення особи в об'єктному оточенні, а заміна його іменником із абстрактною семантикою призводить до створення метафоричного образу за допомогою ФО;

б) іменником: *Та забув козак за жінку, /* ***Кинув лихом об біду****, / Та й потяг козак до шинку/ Голопузу галайду* (Забаштанський: 21). Творчий почерк майстра поетичного слова В. Забаштанського проглядає у влучній заміна традиційного іменникового компонента ***землю*** ФО ***ударити лихом об землю*** „забувати біду, горе; не журитися, не втрачати надії на краще" (СФУМ: 733) словом вільного вжитку ***біду***, яке перебуває у синонімічних відношеннях з іншим компонентом фразеологізму — ***лихо***, що сприяє римуванню в поезії та призводить до бажаного стилістичного ефекту: трансформована ФО влучно передає стан, у якому перебуває козак.

Рідше трапляється у поетичних творах сучасних авторів цілеспрямована субституція декількох ФО: *Катерина, баба зла, / дочка сучина, / на козацький край була/ вкрай розлючена./ Щоб* ***змести зі світу*** *пріч / нашу націю, / обернула славну Січ/ на руйнацію* (Білоус: 38). Кількісно нерівноцінна заміна трьох компонентів ***з лиця землі*** кодифікованої ФО ***змести з лиця землі*** *кого, що* „знищити, розгромити" (СФУМ: 267) двома компонентами ***зі світу*** з морфологічного погляду є рівноцінною, проте оказіональна ФО більш точно передає історичну трагедію руйнування Запорозької Січі;

З висоти моєї темноти/ Видно шлях твій, нене, з божих кар.../ Далі йду тим шляхом до мети/ І ***несу покути свій тягар*** (Забаштанський: 45). Вважаємо, що індивідуально-авторська ФО ***несу покути свій тягар*** утворена шляхом заміни компонента ***хрест*** загальномовної ФО ***нести свій хрест*** „терпляче зносити труднощі, все, що судилось і стало неминучим у чиємусь житті; виконувати свій обов'язок до кінця" (СФУМ: 433). Лексична заміна конкретизує висловлювання, наголошуючи на покірності та терпеливості ліричного героя.

Зустрічається у дослідженому матеріалі й повна лексична заміна компонентів, що спричинює появу нової ФО, утвореної за моделлю загальномовної. Тобто бажаний стилістичний ефект досягається у тому випадку, коли читач впізнає традиційну ФО або за формою, або за значенням.

А Самотнім Жінкам Великого Міста/ завжди тоскно у ньому, і тісно, і пісно, / і нудьга їх тяжка, і свобода їх пізня, / знічев'я терза їх яка-небудь пісня, / й холодильник порожній гуркоче ночами, /щоб скоріше дуріли — без сну і без тями, / ***без царя в голові і у серці — без клямки...*** (Голота: 37). Структурна модель та фразеологічне цілісне значення індивідуально-авторського новотвору ***у серці без клямки*** співвідноситься із загальномовною ФО ***без царя в голові*** „нерозумно, безрозсудно" (СФУМ: 755), розвиваючи новий відтінок у значенні „нестримно, невдержимо, непогамовно".

І рвонеться зойк несамовитий/ в коливної ситої юрби.../ Це пора востаннє полюбити/ і явити лицарку з раби./ І ***зітерти грим, зірвати маску****/ і почути, як болить межа./ Б'ють, віщують зміни тулумбаси, / ера спить на кінчику ножа* (Поклад: 3); *На Парнасі не можна за світ не радіти, / Там трава голуба в золотючій росі./ І дівчата усі — Афродіти, / І, їй-богу, незаймані всі./* ***Там всі кози у золоті, в сріблі — козли,*** */ Ще, гляди, рекетню підохотим, / Кожен з нас там ставав Дон Кіхотом, / Крім отих,*

що в нардепи сповзли (Забаштанський: 4). Структурна модель та фразеологічне цілісне значення індивідуально-авторського новотвору ***зітерти грим*** співвідноситься із загальномовною ФО ***зривати маску*** *з кого—чого і без додатка* „викривати кого-небудь, показуючи його справжню сутність, істинні наміри і т. ін.” (СФУМ: 272), відповідно ***в сріблі — козли*** співвідноситься з традиційною ФО ***усі кози в золоті*** *кому* „все здається, сприймається кращим, ніж є насправді” (СФУМ: 303). Розпізнати індивідуально-авторську ФО допомагає спеціально організований контекст, де поряд з оказіональною розташована загальномовна ФО, що увиразнює висловлювання.

Трапляються в поетичних текстах сучасних українських авторів і ФО, утворені за моделями традиційних, які вже набули поширення в загальномовній практиці, проте чомусь не зафіксовані „Словником фразеологізмів української мови”: *Ніколаєнко важить щоденно, / Скільки ректорам* ***ставити клізм****./ Що будує він в школах/ натхненно — / Хіба ющенків соціалізм?!* (ЛУ, 06.10.05: 5). На нашу думку, вираз ***ставити клізму*** відповідає визначенню ФО із семантикою „дуже сварити або гостро критикувати кого-небудь”;

Ще африканські риси й бакенбарди/ в салонних скалозубів на вустах./ Ще дотепи гримучі, як петарди./ Ще в захваті від чар кавалергарда/ ***левиці світські*** *й* ***молодь золота*** (Базилевський: 74). Ми вважаємо, що сполучення ***світська левиця*** потрібно кваліфікувати як ФО зі значенням „вишукана особа, що належить до привілейованого прошарку населення (про жінку)”, а вираз ***золота молодь*** відповідає визначенню „юнаки і юнки, що живуть у розкоші”.

Таким чином, граматичні зміни ФО, зафіксовані у мові сучасної української поезії, не несуть суттєвих змін щодо семантики фразеологізму. Насамперед вони свідчать про наявність у ФО парадигматичних форм [3: 94—103] для зв’язку фразеологізму з мовним оточенням. Разом з тим граматичні зміни служать для конкретизації висловлювання, виражають особливість авторської манери фразеовживання. Вони не суперечать мовній нормі, хоча питання їх кодифікації в науковій літературі остаточно не розв’язане. Лексичні заміни компонентного складу ФО змінюють якісну і кількісну структуру фразеологізму. Субституція компонентів умотивовується контекстом і здійснюється цілеспрямовано, з певною стилістичною метою: актуалізації, оновлення ФО, посилення виразності; конкретизації, деталізації фразеологізму та всього висловлювання. Лексична заміна компонентів ФО спричинює появу індивідуально-авторських новотворів, котрі з часом, можливо, увійдуть до фразеологічних словників.

Література

1. Білоноженко В. М., Гнатюк І. С. Функціонування та лексикографічна розробка українських фразеологізмів. — К.: Наук. думка, 1989.
2. Мокиенко В. М. Славянская фразеология. — М.: Выс. шк., 1989.
3. Молотков А. И. Основы фразеологии русского языка. — Л.: Наука, 1977.
4. Скрипник Л. Г. Фразеологія української мови — К.: Наук. думка, 1973.
5. Ужченко В. Д., Авксентьєв Л. Г. Українська фразеологія. — Х.: Основа, 1990.
6. Шанский Н. М. Фразеология современного русского языка. — М.: Выс. шк., 1969.

Джерела

Базилевський Базилевський В. О. Украдене небо: Поезії. — К.: Укр. письменник, 1999.

Білоус Білоус Д. Г. Ромен хіба не родич Рима?: Поезії. — К.: Укр. письменник, 1999.

Голота Голота Л. В. На чоловічий голос: Поезії. — К.: Укр. письменник. — 1999.

Забаштанський Забаштанський В. О. Найкревніша рідня — К.: Укр. письменник, 1999.

Кордун Кордун В. М. Зимовий стук дятла: Поезії. — К.: Укр. письменник, 1999.

ЛУ Літературна Україна.

Поклад Поклад Н. І. Молоде сонце: Вірші. — К.: Укр. письменник, 1999.

СФУМ Словник фразеологізмів української мови / уклад. В. М. Білоноженко та ін. — К.: Наук. думка, 2003.

О. Тищенко, к. філол. н.*
Інститут української мови НАН України (Київ)

* © *О. Тищенко, 2006*

УДК 161.2.81'322.221.24

КОНЦЕПЦІЯ І ПРИНЦИПИ УКЛАДАННЯ "СЛОВНИКА УКРАЇНСЬКОЇ ЖЕСТОВОЇ МОВИ ГЛУХИХ"

Актуальність лексикографічної параметризації жестової мови глухих, загальні засади створення „Словника[1] української жестової мови глухих" (далі – Словник), принципи добору фактичного матеріалу до Словника на основі корпусу жестової мови глухих ми розробили раніше [див. 10; 17]. У цій статті пропонуємо детальніше окреслені концепцію та принципи укладання Словника на основі сучасного вітчизняного й світового лексикографічного досвіду в галузі вербальної мови [9], мови тіла (кінестетики) [6] та жестової мови глухих [5; 2; 4; 7; 8; 19; 13].

Тип і призначення Словника. Словник планується як жестово-вербальна лексикографічна праця тлумачно-перекладного типу. У ньому в якомога найповнішій формі буде подано жестовий (лексичний та фразеологічний) склад сучасної жестової мови глухих.

Характер Словника:

– за принципом добору матеріалу та його інтерпретації: реєстраційно-нормативний. Так, дескриптивно-прескриптивний принцип подання жестів передбачає а) розміщення їх за специфічним „алфавітом", параметрами жестів – конфігурацією, локалізацією, характером руху, немануальними компонентами (візуальним, мімічним, рухом голови, торсу тощо); б) найповнішу їхню нормативно-кодифікаційну характеристику відповідно до сучасного стану української жестової мови глухих, рекомендації до вживання тощо.

– за принципом подання жестового мовного матеріалу – це Словник активного типу, оскільки планується репрезентувати жести в найтиповіших жестових контекстах, де виявлятимуться денотативні, сигніфікативні, конотативні, прагматичні, внутрішньомовні значення основних номінативних одиниць жестової мови – жестів, їхні синтагматичні зв'язки. Це сприятиме активному засвоєнню мови та володіння нею користувачами – глухими та з вадами слуху, сурдоперекладачами, а також тими, хто спілкується з такими людьми (батьками, учителями тощо).

– за принципом побудови словникової статті: Словник багатоцільового призначення. Цього плануємо досягти шляхом подання якомога ширшого обсягу інформації про словникові одиниці, що охоплюватиме їхню граматичну, стилістичну (зокрема експресивну), семантичну, етимологічну, тематичну, жестотвірну, синтагматичну, варіантнісну, „вимовну" та інші характеристики. За можливості будуть подані відомості про першу лексикографічну фіксацію жесту та її особливості. Словникова практика виявляє необхідність також уведення нових ремарок, які диференціюватимуть жести калькованого жестового мовлення чи природної жестової мови, штучні чи інсценувальні, жести, спільні для чуючих та глухих, і жести притаманні винятково системі жестової мови.

Порівняно з наявними жестовими лексикографічними працями Словник ширше охоплюватиме жестовий склад сучасної природної жестової мови глухих за рахунок добору матеріалу з живої мови комунікантів, різних за віком, статтю, соціальними, віковими та іншими параметрами, а також з різних областей України (актуалізовуватимуться діалектні – територіальні та соціальні – відмінності). Основою реєстру та ілюстративної частини будуть матеріали корпусу жестової мови, принципи добору й параметризації якого ми розробили на основі досягнень сучасної корпусної вербальної лінгвістики (див. статтю у цьому випуску).

Також ми пропонуємо нові для жестової лексикографічної практики параметри: по-перше, це тлумачення жесту або його значення (за необхідності, коли недостатньо навести вербальний відповідник), по-друге, це найповніше ремаркування відповідно до зазначених вище аспектів, по-третє, ми пропонуємо новий спосіб ілюстрування жестів (див. далі). Можливо, будуть зазначені частотні параметри, конотаційні характеристики тощо. Обов'язковою є жестова транскрипція, яка стає невід'ємною частиною словникових статей у зарубіжних словниках жестових мов (характеристику наявних словників жестової мови глухих – як вітчизняних, так і зарубіжних – див. у наступному випуску збірника). Доцільним буде визначення необхідних і достатніх ідентифікаційних ознак жесту, а у зв'язку з цим – наведення варіантів жесту.

Вважаємо за необхідне тлумачити лексичне або граматичне значення певних параметрів жесту, напр. характеру руху (за аналогією до морфем – продуктивних словотворчих одиниць). Так, повторюваність, циклічність, швидкість, інтенсивність тощо можуть виконувати смисло-

[1] Можливо, таку працю слід назвати жестівником.

або формотворчі функції, як-от: двократне повторення жесту ВИ[1] в напрямі об'єкта мови надає мовній одиниці а) нового лексичного значення належності, утворюється інший жест – ВАШ, б) або граматичного значення давального відмінка, що у вербальному еквіваленті значить ВАМ.

Структура словника. Як уже було зазначено, жест як елемент комунікативної системи глухих має такі характеристики: конфігурація, побудова, локалізація, характер руху, немануальний елемент. Традиційно (хоча й не безперечно) ці характеристики називають фонемами [7]. Упорядкований набір таких фонем щодо кожного жесту становить його нотацію (або транскрипцію). З огляду на це пропонуємо шість блоків корпусу словника. Перші п'ять блоків – це п'ять своєрідних алфавітів, де зображено, описано й символічно позначено варіанти кожної з названих вище фонем. Так, конфігурація представлена сорока вісьмома варіантами, локалізація – від п'яти до дев'яти (за різними даними), напрям руху – кількістю до десяти варіантів тощо.

Зміст цих блоків використовуватимемо для упорядкування й транскрибування жестів у **реєстрі** основного – шостого блоку, де репрезентовано власне словникові статті. Обов'язковим буде список скорочень, словесний покажчик, список джерел.

Реєстр Словника становитимуть загальновживані жестові одиниці на означення конкретних і абстрактних понять, зокрема лексико-тематичні групи руху, мови, мислення, емоційної сфери, суспільної, економічної, наукової, побутової діяльності тощо: назви осіб, зокрема за внутрішніми та зовнішніми ознаками, за родом діяльності, належності до соціальних, етнічних тощо груп; назви рослин, тварин; назви матеріалів; харчових продуктів, страв; хронотопи; назви свят; терміни; назви зовнішніх чуттів; назви емоцій, станів; назви ознак, дій, процесів, назви числових величин та ін.

До реєстру ввійдуть жести, сфера вжитку яких звужена у зв'язку із соціальними, територіальними, професійними чинниками; жестам, застарілим за внутрішньою (ДИЛІЖАНС – ВАГОН) чи зовнішньою формами (БАТЬКО – *імітування поцілунку руки*) з відповідними позначеннями. Особливу увагу приділятимемо жестам, які не мають слова-еквівалента у вербальній мові, напр. жестам – назвам понять ***жестова мова, жестове ім'я,*** жесту, що означає ***відмінність інтелектуального рівня глухої та чуючої людини*** тощо.

На відміну від наявних словників жестів, де з ономастичної групи подано обмежене коло географічних назв, Словник репрезентуватиме не продактильовані, а жестові імена й деякі прізвища людей (найуживаніші), топоніми, гідроніми, астрономічні назви, зокрема назви важливих для українців соціальних, політичних, історичних реалій.

Виокремлюватимуться жести-наслідування, імітації, які мають дискретний характер. Окремо подаватимуться продуктивні елементи – основи для творення інших жестів (напр., *імітування сліз, які течуть по щоках* лежить в основі жестів на означення понять ***плакати, згорьований, дуже сумний***, ці жести різняться немануальним компонентом, інтенсивністю руху тощо).

Структура тлумачної статті Словника. Ключовою характеристикою розташування жестів у реєстрі є конфігурація. Так, у цьому блоці Словника кількість розділів зумовлена кількістю конфігурацій, усередині кожного з них жести буде впорядковано за побудовою, локалізацією, характером руху, немануальним компонентом.

Реєстрова одиниця – жест (малюнок або фото). Далі наводиться транскрипція, вербальний опис виконання жесту, а також ремарки (граматичні, стилістичні тощо). Плануємо також характеризувати жести за їхніми онтологічними параметрами, в основі яких – класифікації Р.Краєвського [12] та Г.Зайцевої [7]. Так, передбачаємо такі групи жестів: виражальні рухи (міміка: *міміка здивування*, жест: ЗАПРОШЕННЯ); жестові знаки (безпосередньої символіки – іконічні: ТАРІЛКА, опосередкованої символіки – асоціативні, штучні: ПОТРІБНО); жести

[1] Тут і далі приймаємо такий спосіб запису: **жести** умовно означаємо вербальними еквівалентами, надрукованими великими літерами (СТІЛ, СТРАШНО, ПРИХОДЬТЕ ДО НАС У ГОСТІ), **вербальний опис виконання жесту** – курсивом (ГОРДИЙ – *рукою в конфігурації „великий палець" провести по груді знизу вверх*) **дактильовані** слова та їхня частини – малими буквами через рисочку (О-л-е-н-а, з-СУХО), **слова- (словосполучення-, речення-) відповідники** беремо в лапки, означувані поняття виділяємо напівжирним курсивом (ЛИЖІ – жест на означення ***особи, дії, предмета, явища, пов'язаних з лижами,*** відповідає таким одиницям, як „лижі", „лижник", „ходити на лижах" тощо). Умовність позначення жесту словом, напр.: ВЕСЕЛИЙ, пояснюємо тим, що вербальний еквівалент часто звужує виражальні можливості жесту, тому доцільніше подавати групу слів: „веселий", „весело", розширюючи його узус.

природної жестової мови, напр., ДИТИНА; жести калькованого жестового мовлення: напр., згаданий жест ДИТИНА, але зі значенням ***лялька;*** спільні жести тощо.

Семантичну характеристику становитимуть: а) значення жесту, виражене тлумаченням, відсиланням до певних реєстрових одиниць (синонімів, антонімів, твірних одиниць та ін.), наведенням вербальних еквівалентів – слів або фразеологізмів; б) особливості синтагматичної (смислової та граматичної) сполучуваності; в) ілюстративний матеріал; г) довідковий відділ.

Омонімічні жести у Словнику подаватимуться окремо під відповідною нумерацією. Варіанти одного жесту (зі зміною і без зміни внутрішньої форми) – гніздовим розташуванням.

Правила виконання жесту (правила „вимови", „орфокінетика"). Традиційним у словниках жестів є словесний опис його виконання. Це буде відбито також у транскрипції (хоча вона, вважаємо, все одно не передає повною мірою специфіки того чи іншого жесту, однак дає змогу точніше, ніж словами, й економніше, ніж малюнком, передати графічно жестову фразу). При цьому зазначатимуться ідентифікаційні (необхідні й достатні) ознаки (конфігурація, локалізація, характер руху, немануальний компонент), варіанти виконання. Напр., Р.Краєвський так описує жест зі значенням ***болото:*** *пальці обох кистей в положенні дактильного знака „а"* (стиснуті в кулак – Т.О.), *кисті перед грудьми, кінчики вказівних пальців спрямовані вниз. Права кисть вище, ліва трохи нижче. Робимо поперемінно два-три рухи обома руками вгору і вниз.. Жестовий знак інсценувальний (коли витягуємо праву ногу з болота, грузне ліва). Іноді зустрічається інший „почерк". Такі самі рухи робимо, коли кисті перебувають у положенні дактильного знака „в"* (долоня розкрита, всі пальці випрямлені й зімкнені – О.Т.), *що по суті одне й те саме* [12]. Так, автор у статті до реєстрового жесту БОЛОТО об'єднав словесний опис виконання, мотивацію (звідки можна виокремити ідентифікаційні ознаки), варіанти виконання, які не виявляють відтінків лексичного чи граматичного значень.

Тлумачення жестів – це основний спосіб їхньої семантичної характеристики. Для цього вважаємо за необхідне звертатися до досвіду мовних тлумачних словників, описово подаючи те чи інше поняття, означене жестом. Пояснення мають бути адекватними, стислими й достатніми, однотипними.

Відсутність ґрунтовних лінгвістичних досліджень природної жестової мови ускладнює цей аспект роботи, до сьогодні жести перекладали словом-еквівалентом (у кращому разі – кількома словами або словосполученням). Проте специфіка жестової мови (синкретичність багатьох жестів, які в різних жестових контекстах означають особу, явище стан, функцію тощо, напр.: жест зі значенням ***вступити, вступ, вступання, надходження;*** зі значенням ***початок, починати; всередині, зміст*** тощо) вимагає іншого способу семантичної характеристики, а саме – пояснення. Залежно від ситуації пояснюватиметься а) жест як такий (*жест на означення* ***предмета, дії, ознаки, процесу... пов'язаних із упорядкуванням певним чином інформації***) із наступною конкретизацією: „список", „меню", „складений у вигляді графіку" тощо; б) лексичне (жестове) значення реєстрової одиниці (у цьому разі доцільно вжити опис явища, предмета або його вербальну назву: „стіл", „стілець", „ваза").

Окремо в статті (чи окремою словниковою статтею, якщо це самостійний жест) будуть розроблені ті вживання жесту, які базуються на вербальних фразеологізмах або близькі їм за значенням, напр., жест, аналогічний вислову „на серці коти шкребуть". Увагу буде приділено й окремим значенням та відтінкам значень жестів, а також випадкам вживання жестів у переносному значенні (із зауваженням: оказіональне чи загальновживане), із певною конотацією та ін.

Тлумачення жестів сприятиме унаочненню **словотвірної (жестотвірної) мотивації.** У Словнику реалізовуватиметься зв'язок мотивувального жесту й мотивованого через залучення першого в тлумаченні другого або через дані довідкової частини.

Ілюстрації. Особливості функціонування реєстрової одиниці в мові, її значеннєві відтінки пропонуємо ілюструвати не традиційно, словесними фразами (або не лише ними), а жестовими фразами, відбитими серією фото чи малюнків. Таким чином ми не звужуватимемо жестову мову до ознак калькованого жестового мовлення як мануальної форми вербальної мови, оскільки склад фрази природної жестової мови зазвичай не відповідає словесному, а ілюстрований жест – певному слову (ідеться передусім про мовні одиниці – назви процесів, абстрактних понять, станів тощо). До того ж семантика синкретичних жестів відбивається лише в жестовому контексті. Так, ілюстрація може бути додатковим поясненням значення реєстрової одиниці.

Усі жести, використані в ілюстрації, мають бути у складі реєстру.

Довідкова частина зазначатиме особливості стилістичного вживання жесту, інформуватиме про його лексикографічні фіксації, етимологію, мотивацію або дериваційні можливості. Напр., жест зі значенням ***лічити, лічба, рахунок*** є інценувальним, уживаним як глухими, так і чуючими: він виконується як *почергове загинання пальців*. Так само жест зі значенням ***прошу, будь ласка*** – це *складені мовби під час молитви долоні обох рук.* Жест зі значенням ***благати*** теж інсценувальний, має прозору мотивацію, проте є одиницею винятково мови глухих: *вказівним і середнім пальцем правої руки імітують „падання на коліна" на розкриту долоню лівої.* Жест – еквівалент слова „епідемія" утворений складанням двох інших жестових одиниць: ЗАРАЗА+ПОШИРЮВАТИСЯ. Жест зі значенням ***різний, багато*** зазвичай використовують для утворення узагальненої назви класу однорідних предметів, явищ: жест СТІЛ, ЛІЖКО+РІЗНИЙ має значення ***меблі.***

Цінною також є відомості про запозичені, інтернаціональні жести, жести, спільні для української та російської мов, а також їхній зв'язок з міжнародною мовою глухих та взвємовпливи.

Таким чином, запропоновані принципи добору, аналізу та інтерпретації жестового матеріалу найповніше на сьогодні репрезентуватимуть особливості мови глухих у Словнику. Звичайно, деякі з положень будуть уточнюватися, коригуватися відповідно до результатів аналізу дібраного відеоматеріалу, тим самим поглиблюватиметься лінгвістична теорія української жестової мови глухих.

Література

1. Жест. Відео. – ТОВ „Республіканський навчально-науковий центр ДІНІТ"– CD. – Ч.1, 2.
2. Зайцева Г.Л. Жестовая речь. Дактилологія. – М., 2000. – С. 37–38.
3. Український тематичний відеословник жестової мови: Навчально-методичний посібник. Продюсер випуску Н.Іванюшева – у 3 CD. – К.: Київська організація УТОГ, Культурний центр УТОГ, Українсько-Канадський Альянс для глухих і слабочуючих, 2003.
4. Иванюшева Н.В. Мимико-жестикуляционная речь. – К.: Реклама, 1968. – 290с.
5. Краєвський Р. Г. Мова жестів глухих. – Київ: Радянська школа, 1964. – 219с.
6. Крейдлин Г. Е. Невербальная семиотика: Язык тела и естественный язык. – М.: Новое литературное обозрение, 2002. – 592 с.
7. Лексикон жестов. – Донецк, Центр для глухих „Видеть значит слышать», 2004.
8. Словник жестової мови: церковна лексика / К. Дятлова, Д. Заварицький. К.: «Видавництво «Глас», 2006. – CD, відеокасета.
9. Тараненко О. О. Новий словник української мови (концепція і принципи укладання словника). – Київ–Кам'янець-Подільський, 1996. – 171 с.
10. Тищенко О.М. Засади створення корпусу української жестової мови глухих. Лексикографічний бюлетень. – Київ, 2006. – Вип. 13. – С. 43–47.
11. Тищенко О. М., Кульбіда С. В. Українська жестова мова глухих як об'єкт лексикографічної параметризації / Лексикографічний бюлетень. – Київ, 2005. – Вип. 12. – С. 37–43.
12. Фрадкина Р. Н. Говорящие руки. Тематический словарь жестового языка глухих России. – М., 2001. – 597с.
13. Svenskt teckenspråkslexikon. – Sveriges Dövas Riksförbund, 1997.

І. Прядко*
Інститут української мови НАН України (Київ)
УДК 161.2.81'373.7

ФРАЗЕОЛОГІЧНІ СИНОНІМИ З НУМЕРАТИВНИМ КОМПОНЕНТОМ В УКРАЇНСЬКІЙ МОВІ

У статті розглянуто найбільш загальні особливості синонімічних фразеологізмів з нумеративним компонентом у складі. Наголошено, що взаємозаміна у контекстах синонімічних ФО можлива лише у випадку їх тотожності. Залежно від позначуваного числа нумеративний компонент переважно впливає на відтінки фразеологічних значень.

Фразеологізми як мовні одиниці становлять певну систему в мові й перебувають у системних зв'язках між собою – синонімічних, антонімічних та полісемічних. Питанням системних відношень у фразеології української мови присвячені праці Л. Г. Скрипник, Л. С. Паламарчука, М. І. Сидоренка, В. Д. Ужченка, Л. Г. Авксентьєва, А. М. Матвієнка, Н. М. Неровні, Л. Г. Боярової та інших.

Дослідження фразеологічних системних зв'язків ведуться у порівнянні з синонімією, антонімією та полісемією лексичною. Це поглиблює знання про особливості семантики

* © *І. Прядко, 2006*

фразеологізмів, дає можливість визначити їх своєрідність як мовних одиниць та закономірності семантичних зв'язків.

Явище синонімії дуже поширене у фразеологічній системі української мови. Це пояснюється багатством виражальних можливостей фразеологізмів у різнобічному відображенні явищ навколишньої дійсності. Серед розмаїття властивостей та рис фразеологічних одиниць їхня здатність вступати у синонімічні відношення є однією із визначальних [2: 6].

Проблема синонімії фразеологічних одиниць, на думку багатьох учених, є найбільш дискусійною в сучасній лінгвістиці. Дослідники зазначають, що фразеологічна синонімія – значно складніше явище, ніж лексична. Це виключає взаємозаміну синонімічних фразеологізмів у контекстах, що буде розглядатися нижче. Важливими проблемами є також розрізнення фразеологічної синонімії та варіантності, встановлення поняття фразеологічної синонімії, а отже, і визначення її критеріїв.

Серед низки диференційних ознак фразеологічних синонімів мовознавці виділяють тотожність або близькість значення, співвіднесеність з тією самою частиною мови, однакову синтаксичну сполучуваність, наявність відтінків у семантиці, різне стилістичне забарвлення. За визначенням Л. Г. Скрипник, "фразеологічними синонімами вважаються фразеологічні одиниці, які позначають той самий предмет дійсності, виражають те саме поняття, відтіняючи різні строни його, і при різній внутрішній формі і неоднаковому лексичному складі мають однотипне категоріальне значення, однакову семантичну сполучуваність із словами оточення" [4: 212].

Широко представлені фразеологічні синоніми і серед фразеологізмів з нумеративним компонентом, які є об'єктом нашого дослідження. Приєднуємося до думки тих фразеологів, які вважають синонімічними фразеологізми близькі за семантикою і здатні виконувати однакову синтаксичну роль, але відмінні лексичним складом та можливими відтінками значень [1: 72]. Розглянемо найбільш загальні особливості синонімічних фразеологізмів з нумеративним компонентом, зафіксованих у Фразеологічному словнику української мови (уклад.: В. Білоноженко, В. Винник та ін.) [9], Словнику фразеологізмів української мови (уклад.: В. Білоноженко, І. Гнатюк та ін.) [6], Фразеологічному словнику української мови Г. Удовиченка [8], а також у Словнику української мови в 11 томах [5].

Кожен синонімічний фразеологізм по-своєму розкриває певне поняття. Так, вирази *згинатись у три дуги (погибелі)* ("плазувати, запобігати перед кимсь, виявляючи покірність" [6: 151]) та *ставати на двох лапках* ("бути готовим виконати будь-які бажання, чиї-небудь прохання; прислужувати, підкорятись чиїйсь волі, владі; служити комусь" [6: 687]) означають приниження, запопадливе схиляння перед ким-небудь, вияв покірності. Проте помітні відтінки у значеннях, що пов'язані із різними обставинами виникнення фразеологізмів. Перший виник у ситуації, коли людина схилялася вниз, кланялась, виявляючи покірність, другий – коли вона хоч, навпаки, буквально не піднімалась угору, але своєю поведінкою та рухами виявляла запопадливість у інший спосіб. Маємо образ неоднакового ступеня абстрактності – від меншого до більшого, що свідчать ілюстрації: "*Слуга .. вийшов до послів і, зігнувшись у три погибелі, мовчки показав обома руками на двері" (Р. Іваничук); "– Цуцик ти, цуцик, – думає. – На двох лапках стаєш. До чого це воно? Хоч і начальство ж отець Полієвкт, а він же чоловік, як і ти" (А. Тесленко).*

Людське мислення в різних ситуаціях виділяє різні сторони одного явища, дії тощо і по-різному виражає пов'язану з цим думку. Так з'являються різні вирази, що, об'єднуючись спільним значенням, різняться певними відтінками. Саме тому фразеологізми *в десятому коліні* та *десята (сьома) вода на киселі*, означаючи дуже далеку рідню, не є тотожними, бо перший іноді вживається і на позначення нащадків*: "Приходить у гості до Скоробагатьків якась далека їх родичка, того ступеня родичання, що називається десятою водою на киселі "(О. Донченко);"Я готовий познакомитись [познайомитись] з вашими родичами, хоч би й в десятому коліні" (І. Нечуй-Левицький);"– А як розколупають, що ми зовсім не з десятого коліна нащадки?" (Григорій Тютюнник).*

Значення виразу *тисяча і одна* "багато" є дуже загальним, що ілюструє, наприклад, такий контекст: *"Ми озброєні знаннями, досвідом, терплячістю, зрештою, своєю небайдужістю до його [вихованця] долі… В нас на нього інструкцій тисяча і одна…(О. Гончар).* В той час як у виразі *у дві пригорщі не захопиш* первісно йдеться про багато конкретних речей, і лише згодом він набуває узагальнення – "дуже багато" [5, т.III:

391], ”дуже багато чого-небудь” [6: 255; 9, т.І: 321]: *”[Карпо:] Цокотуха! Достометна [достеменна] сорока-білобока: наскрекотала, наскрекотала, що й в дві пригорщі не захопиш (М. Кропивницький).*

Як свідчать лексикографічні джерела, вибірково, поверхово – *п’яте через десяте* – можна читати, розповідати, знати, снідати, знайомитись з кимсь тощо [6: 589; 8, т. ІІ: 178; 9,т. ІІ: 724], натомість *у двох словах* (”дуже коротко, стисло”[5, т.ІХ: 370], ”дуже стисло” [6: 665; 9, т. ІІ: 828], а отже, так само вибірково, поверхово) – лише розповідати про щось: *”Поснідавши п’яте через десяте, Микола одягнувся, вхопив ранець з книжками і вибіг з хати”(В. Гжицький)*; *”Так ото слухай, скінчу вже в двох словах”(М. Коцюбинський).*

Багато справ починається *з нуля*, тобто ”з нічого, без попередньої основи; від самого початку” [6: 442], а *з першого кілка*, або ”з самого початку, з нічого” [6: 295] – переважно будівництво.

Наведені приклади свідчать також і про те, що з двох закладених у синонімічних (як однозначних, так і багатозначних) фразеологізмах понять більш загальне може вживатись щодо конкретного, а не навпаки.

Фразеологізм *в сто очей дивитися* має більшу експресивність і виражає більший ступінь вияву ознаки уважності та пильності порівняно з *дивитися обома*, що ілюструє вплив нумеративного компонента на значення фразеологічного виразу.

Семантична домінанта ”швидко” об’єднує ФО *у двадцять чотири години; в три скоки; у три мига; алюр три хрести; одна нога тут, а друга там*, проте за ”ступенем” швидкості та об’єктної спрямованості їх можна розмістити у такий ряд: *алюр три хрести (”дуже швидко! негайно!”* [6: 20])*; в три скоки* (”дуже швидко, моментально, умить” [6: 658])*; у три* мига (”дуже швидко, миттю” [9, т. І: 485]) – ”швидко, миттю”; *одна нога тут, а друга там* (”уживається для вираження наказу, прохання або обіцянки дуже швидко сходити, збігати і т. ін. куди-небудь [6: 438]”) –” швидко, не забарюючись”; *у двадцять чотири години* (”за короткий час, негайно” [6: 153] – ”швидко” (проте у порівнянні 24 години – досить тривалий час).

В обох синонімічних парах фразеологізмів *все одно; один чорт* (об’єднані спільним поняттям ”однаково”) та *ні одна душа; ні один собака* (спільне поняття ”ніхто”) другий зворот має відверто грубий, зневажливий відтінок, вирізняється емоційно-стильовою навантаженістю.

Нумеративний компонент у складі ФО надав можливість відтворення яскравіших, експресивніших образів, а отже, стали можливими і такі фразеологічні синоніми: із загальним значенням ”жадібно, багато їсти” – *у три горла; за трьох; за трьох дурних*; із загальним значенням розумової обмеженості людини, її нетямущості, некмітливості – *голова – два вуха; у голові сім літ не метено; нема десятої (сьомої, дев’ятої та ін.) клепки в голові; абсолютний нуль; нуль без палички*; із загальним значенням ”бути щасливим, у піднесеному настрої” – *на сьомому (п’ятому, десятому) небі; як на сто коней висадив* тощо.

Спільна ідея ”бути немічним, хворим, в стані повного виснаження, на грані смерті” об’єднує звороти *стояти однією ногою в могилі; душа на одній нитці держиться; тільки одна душа залишилася; одна тінь залишилася. І*х синонімічність зрозуміла й поза контекстом.

Прозорими є ситуації виникнення синонімічних фразеологізмів *співати однієї й тієї* (”весь час повторювати те саме, ту саму думку” [9, т. ІІ: 849]) й *закручувати одну і ту ж пластинку* (”говорити весь час те саме” [9, т. І: 309]); *як пяте колесо до воза, як собаці другий хвіст, як собаці п’ята нога* (”зовсім не потрібний” [6: 304, 674, 438]); *міряти однією міркою(на одну мірку, на один аршин)* (”оцінювати, характеризувати і т. ін. кого-, що-небудь однаково, без урахування індивідуальних особливостей; вважати однаковими” [6: 394]) і *стригти під одну гребінку* (”однаково оцінювати всіх” [6: 694]) та багатьох інших.

Різна образність синонімів викликає різні асоціації, по-своєму відображає те саме явище дійсності.

У визначенні синонімічних відносин між фразеологізмами мовознавці виділяють можливість їх взаємозаміни у контекстах [1: 73; 4: 226]. Це твердження, на нашу думку, вимагає деяких уточнень. Така взаємозаміна можлива лише у випадках тотожності фразеологізмів. Наявні семантичні відтінки перешкоджатимуть взаємозаміні, тим більше у тому самому лексичному оточенні.

Так фразеологічні одиниці *нуль без палички* й *абсолютний нуль* не мають суттєвої відмінності у семантиці і позначають нікчемну, непотрібну людину, яка нічого не варта. У реченні ” – *Я – ніщо. Розумієте, абсолютний нуль. Навіть не перекотиполе. Те, котячись по ріллі, бодай розсіває своє насіння і вірить, що проросте десятками нових пагінців. А я – вакуум, порожнеча” (Л. Дмитерко)* при спробі заміни фразеологічної одиниці на синонімічну *нуль без палички* виявляється ледь помітний відтінок у значенні – пом’якшується інтенсивність вияву ознаки вартості людини, додається незначний відтінок зневажливості у її визначенні.

Навіть і тотожні фразеологізми не завжди бездоганно ”ввійдуть” у речення, не порушивши стилістичного забарвлення (*один на один* і *на чотири ока* – семантична домінанта ”удвох, без сторонніх”), ритмомелодики (*ні в дві ні в три; ні в дві ні в п’ять; ні в п’ять ні в дев’ять; ні в п’ять ні в десять* – спільне поняття ”стати розгубившись, зніяковівши”) чи інших нюансів: *до сьомих віників* та *до другого пришестя* (загальне значення ”довго, тривалий час пам’ятати покарання”); *не рік не два; не перший рік* (спільне значення ”протягом тривалого часу”); *першої марки; перший сорт* (спільний образ ”дуже добрий, найкращий”); *у першу чергу; першим ділом* (семантична домінанта ”насамперед, передусім, спочатку”); *всипати по перше (шосте) число; дати під сьоме ребро; здирати сім шкур* (найзагальніше значення ”суворо покарати кого-небудь, переважно б’ючи”).

Отже, можливість повноцінної взаємозаміни у контекстах не можна вважати одним із визначних показників синонімічності фразеологізмів. Існує багато обставин, які роблять цю взаємозаміну неможливою.

Як показують наведені приклади, синонімічні ряди фразеологізмів з нумеративним компонентом можна побудувати як зі структурно одномодельних, так і з різномодельних зворотів, хоча в такому разі вони іноді і не є найближчими синонімами. Специфіка досліджуваної групи фразеологізмів обмежує нас і в цьому плані. Серед фразеологізмів з компонентом один виявився синонімічний ряд зворотів з дуже близькими моделями (відмінкова та прийменниково-відмінкова форми іменника з числівником) : *одним духом; одним подихом; одним махом; одним ударом; за одним махом; в один миг; в одну мить; за одним присідом; в один дух; за (в) один раз; за одним разом; з одного маху; в одну хвилину*, об’єднаних загальним значеням ”відразу, тут же, дуже швидко, миттю”: *”Гармаші, вправні, дебелі хлопці, одним духом викинули німецьку зенітку з кругляка й поставили на її місце свою” (О. Гончар); ”В один дух прибіг”(Приказка); ”В одну хвилину скочила дитина в куток; закапали від болю й страху сльози” (М. Коцюбинський); ”Хіба в земстві не ті люди, що й були? Хіба вони в один раз перемінилися?” (П. Мирний); ”Коли б мені літ двадцять назад, то оцей стакан за одним разом до дна осушив” (П. Мирний).*

Із фразеологічною синонімією тісно пов’язане явище варіантності. Це суміжні явища фразеології. Вирішальна роль у їх розмежуванні належить внутрішній формі, або образній основі [2: 19], фразеологічної одиниці. Стійкість внутрішньої форми завжди виступає показником варіантності фразеологізмів, а її порушення, тобто утворення фразеологічної одиниці з іншою внутрішньою формою, є свідченням синонімічності. Цей факт є надійною її ознакою. Варіанти ж – ”форми вживання одного й того ж фразеологізму” [3: 1], ”фразеологічні одиниці з фонетичними, морфологічними або словотвірними змінами одного й того ж компонента” [2: 18].

Чотири фразеологічних одиниці з нумеративними компонентами *ні в дві ні в три; ні в дві ні в п’ять* (”перебувати в нерішучості, розгубитися, знітитися; зовсім несподівано” [6: 189]*; ні в п’ять ні в дев’ять; ні в п’ять ні в десять*(”знітившись, розгубитися” [6: 589] об’єднані загальним значенням розгубленості і не зафіксовані в інших формах. Вони мають сталу будову, що зумовлена, крім інших чинників, ще й ритмомелодикою (нема фразеологізмів *ні в дві ні в сім* чи *ні в три ні в п’ять* тощо). Створене цими зворотами спільне поняття, яке не змінюється під впливом контексту і не набуває й найменшого відтінку у значенні, може дати підставу вважати їх лексичними варіантами у фразеологізмах: *”Чувшись вільним, лис за курку Та й махнув, куди видать, Чоловік став наче дурень, Жінка теж ні в дві ні в п’ять”(І. Франко);”– А ти чого тут хочеш, заволоко? – Леон став ні в дві ні в три на таке привітання” (І. Франко); ”Доктор стояв, як то кажуть, ні в п’ять і ні в дев’ять” (І. Франко);”Пани регочуть, мужики поставали, ні в п’ять, ні в десять. Вони поспускали голови і не знають, що діяти (В. Стефаник).* Але спостереження за компонентним складом дозволяє нам виділити в цій групі два фразеологізми з варійованими компонентами: *ні в дві ні в три (ні в п’ять)* та *ні в п’ять ні*

в дев'ять (ні в десять), які, в свою чергу, перебувають між собою у синонімічних відношеннях.

Однаковий образ, а саме "дуже товста людина" створює ряд фразеологічних одиниць з компонентами-числівниками: *два на два; три на три; три на чотири; шість на шість; сім на вісім; вісім на сім; вісім на вісім; дев'ять на дванадцять. В*они є абсолютними синонімами, оскільки відтворюють спільний образ, а також здатні до варіювання: *три на три (на чотири); вісім на сім (на вісім); сім (вісім) на вісім.*

Серед ФО з нумеративним компонентом синоніми представлені помірною кількістю. Синонімічні ряди нечисленні, переважно складаються з двох-трьох одиниць. Більшість синонімів за значенням є близькими, тотожних порівняно менше. Нумеративний компонент, безперечно, впливає на семантику фразеологізмів, а отже, і на утворення синонімічних одиниць. У деяких випадках він здебільшого виступає засобом емоційного забарвлення виразу, суттєво ж впливає на значення фразеологізмів числівник один і зумовлює велику кількість синонімічних рядів. Залежно від позначуваного числа, нумеративний компонент у більшості випадків має вплив на відтінки фразеологічних значень.

Література

1. Авксентьєв Л. Г. Сучасна українська мова: Фразеологія. – Харків: Вища шк., 1988.
2. Боярова Л. Г. Синонімія дієслівних фразеологічних одиниць сучасної української літературної мови. – Автореферат дис. … канд.філол. наук. – К., 1981.
3. Неровня Н. М. Варіантність і синонімія в українській фразеології.– Автореферат дис. … канд. філол. наук. – К., 1990.
4. Скрипник Л. Г. Фразеологія української мови. – К.: Наук. думка, 1973.
5. Словник української мови: В 11 т. – К.: Наук. думка, 1970–1980.
6. Словник фразеологізмів української мови, – К.: Наук. думка, 2003.
7. Ужченко В. Д., Авксентьєв Л. Г. Українська фразеологія, Харків, 1990.
8. Фразеологічний словник української мови: У 2 т. – К.: Вища школа, 1984.
9. Фразеологічний словник української мови: У 2 т. – К.: Наук.думка, 1993.

РЕЦЕНЗІЇ

А. Поликарпов, д. филол. н.*
МГУ им. М. В. Ломоносова (Россия, Москва)

ОБ ОДНОЙ РЕЦЕНЗИИ

С чувством недоумения и горечи прочитал в журнале «Мовознавство» (№5 за 2005 г., ст. 80-83) рецензию В. Широкова на монографию О. Демской-Кульчицкой «Основи національного корпусу української мови» (Київ, 2005. – 219 ст.). Рецензия обращает на себя внимание явно неакадемическим стилем и, что главное, наличием в ней многих терминов и сентенций, странных, а то и некорректных с позиций языковеда. Позволим себе привести несколько цитат.

«Насправді лінгвістичний корпус є далекосяжним комп'ютерним узагальненням та розвитком такого об'єкта як звичайна лексична картотека» (ст. 80).

В этом «определении корпуса» чуть ли не каждое слово вызывает вопросы. Что такое «**лингвистический** корпус»? Лингвистика – это наука о языке, лингвистический – относящийся к лингвистике: лингвистическая теория, лингвистическая энциклопедия, лингвистическое образование, лингвистический конгресс и под. В таком случае, лингвистический корпус – это некий корпус, касающийся науки о языке, но не языка. Следует сказать, что этой ошибки в употреблении термина «лингвистический» квалифицированный лингвист не допустит.

Что такое «комп'ютерне узагальнення лексичної картотеки»? «Словник української мови» в 11-и томах дает такое определение существительному «узагальнення»: «загальний висновок, загальні положення, що ґрунтуються на порівнянні окремих предметів, фактів, явищ і виявленні їх спільних рис» (СУМ, т. X, ст. 403). Исходя из этого определения, можно с полным правом утверждать, что приведенное высказывание уважаемого рецензента никакого научного, лингвистического смысла не имеет. Так написать может только человек, весьма далекий от корпусной лингвистики и вообще от языкознания, возможно, не видевший ни одного текстового корпуса. Но ведь рецензент возглавляет «Український мовно-інформаційний фонд», в котором, как он утверждает, создан «лингвистический корпус», на базе которого осуществлен ряд исследований самого разного характера (см. библиографию в конце рецензии – ст. 83). Получается, что уважаемый рецензент сам себе противоречит. Если созданный в этом учреждении корпус соответствует приведенному выше определению, то при чем здесь исследования по грамматике? Напрашивается вывод, что эти исследования не имеют отношения к корпусу, хотя и выполнены в данном учреждении.

Действительно, среди сфер возможных применений текстового корпуса есть и лексикография, создание автоматической лексической картотеки, но это – лишь одно из многих применений корпуса, как хорошо и убедительно показано в рецензируемой В. Широковым монографии (ст. 39 и дальше).

Еще одна цитата:

«Наукова проблема у цій галузі лежить у площині побудови формалізованих моделей лінгвістичної компетенції (принаймні, певних її ділянок), апробації та верифікації цих моделей та створення на їхній основі ефективно діючого програмного забезпечення» (ст. 80). Витиевато и туманно. Конечно, программное обеспечение текстового корпуса – важная задача. Но до ее выполнения должно быть решено много лингвистических научных проблем, касающихся отбора базовых единиц хранения будущего корпуса, обеспечения его репрезентативности, структуризации текстовой базы, разработки лингвистических алгоритмов, обеспечивающих использование корпуса в различных исследованиях, создание базы характеристических данных для осуществления этих алгоритмов при решении различных лингвистических задач и т.п. Все эти проблемы детально рассматриваются в обсуждаемой монографии, но не являются, судя по всему, важными для рецензента В. Широкова. Возможно, именно их он объединил под термином «лингвистическая компетенция»? Если расшифровывать, выявить действительное содержание этого зашифрованного В. Широковым обозначения, то окажется, что это самое важное знание в корпусной лингвистике – знание о принципах и способах создания и анализа корпусов текстов. Ведущим в этом знании являются

* © *А.Поликарпов, 2006*

фундаментальные лингвистические представления о системе языка и формах ее реализации в текстах, а не само по себе программное обеспечение для организации и анализа корпуса. Базовая корпусная информация, заключающаяся в содержании и структуре текстов корпуса, их разметке жанрово-стилистической, хронологической, источниковой, авторской и иной отнесенностью, закладывается еще при создании корпуса и в определенном смысле является далее неизменной (пока не меняется сам состав берущихся для корпуса текстов). Также закладывается в самый начальный период и далее последовательно реализуется лингвистами, работающими с данным корпусом, идеология корпуса – концепция уровневого членения каждого текста корпуса, сегментации его единиц, глубина категоризации этих единиц и т.п. Идеология корпуса и базовая корпусная информация существует независимо и до разработки любого соответствующего программного обеспечения. Программные комплексы разрабатываются для обслуживания задач использования и анализа этой базовой лингвистической информации, а не наоборот. В противоположность В. Широкову мы считаем, что создание «эффективного программного обеспечения» не может быть целью корпусной лингвистики. Ее целью является решение различных лингвистических задач на контролируемом текстовом материале, представительно отображающем часть языковой практики данного общества. Считать иначе – значит заниматься непонятным программистским нарциссизмом, не имеющим ничего общего с реальной корпусной работой.

И далее:

«Ми б не стали робити цих зауважень, якби авторка з такою наполегливістю не педалювала цифри, якими вона фіксує як абсолютні, так і відносні пропорції окремих елементів свого гіпотетичного корпусу – без обґрунтування і без смислу. Бачимо, що і з філологією у пані Орисі не дуже добре» (ст. 82). Не удивительно ли, что об уровне филологической подготовки Демской-Кульчицкой рецензент судит по упорству, с которым она «педалирует цифры»? Между тем, эти цифры свидетельствуют о том, насколько серьезно автор монографии рассматривает структуру и представительность будущего корпуса. И смысл в этих ее выкладках безусловно есть, просто рецензент, видимо, не имея необходимых базовых знаний по общей и корпусной лингвистике, его не улавливает. Иное дело, что в монографии О. Демской-Кульчицкой следовало бы детальнее обосновать соотношение представления текстов разных функциональных стилей, возможно, некоторые виды текстов, исчисляющиеся всего тысячей словоупотреблении, в корпус не включать или, наоборот, серьезнее обосновать их включение; есть и опечатки в таблице 14 (ст. 104), на которые необходимо обратить внимание. Т.е., я считаю, что хотя и в монографии О. Демской-Кульчицкой «Основи національного корпусу української мови» можно найти недостатки, это явно не означает, что она из одних недостатков состоит. Более того, я считаю монографию О. Демской-Кульчицкой очень важным достижением украинской корпусной лингвистики, т.к. в ней представлены принципы работы по созданию национального корпуса, без обсуждения которых корпусная работа подобного масштаба просто невозможна.

В связи с этим еще одна цитата из рецензии В. Широкова:

«Цілі та завдання, які ставить собі пані Демська-Кульчицька, є надуманими і не мають нічого спільного зі справжніми проблемами сучасної корпусної лінгвістики, насамперед, з проблемою, про яку ми говорили вище: побудовою ефективних засобів автоматичного та автоматизованого маркування лінгвістичних структур у текстах, наявних у корпусі» (стор.81).

Рецензент явно не понимает, что такое создание концепции некоторого проекта. До разработки «ефективних засобів автоматичного та автоматизованого маркування лінгвістичних структур у текстах" при планировании работ по созданию серьезного корпуса надо, как минимум, разработать (а) перечень задач, которые будут решаться на основе использования корпуса, (б) концепцию отбора текстов в корпус (определить оптимальное соотношение в нем объема текстов различных функциональных сфер и типов коммуникации), (в) концепцию основных используемых единиц анализа текстов, (г) концепцию характеризации единиц корпуса различными классами помет и т.п. Подчеркнем, что сама концепция создания, структуризации и разметки корпуса базируется на представлении его разработчиков о том, какие задачи в будущем будут решаться на его основе. В итоге, "способы автоматического и автоматизированного маркирования единиц и структур текстов" могут быть очень разными, вытекающими

именно из той концепции, которая принята авторами-разработчиками корпуса. Рассуждать о таком маркировании без наличия концепции, значит обнаруживать ее отсутствие. Если же что-то делается без общего плана (концепции), то, естественно, это чревато для такой работы теми или иными тупиками. Святая вера некоторых технарей при лингвистике, что достаточно сгрести в кучу какой-то большой объем текстов, нагнать побольше программистов, и дело сделано – находится в вопиющем противоречии с действительностью. Моментами подобного грубого и малоперспективного подхода к проблеме составления национального корпуса украинского языка характеризуется и рецензия пана Широкова.

Например, это видно и по такой фразе этой рецензии:

«Це, взагалі, унікальна ситуація у мовознавчій науці – описувати властивості об'єкта, якого не існує в природі» (ст. 82). Разве рецензенту не известно, что концепция любого проекта создается до реализации самого проекта, а не тогда, когда проект уже осуществлен? Разве рецензент никогда не видел ни одного проспекта будущего большого издания, будь то грамматика, словарь или атлас? Концепция и состоит в том, что в ней как можно подробнее описываются принципы и возможные свойства еще не существующего объекта, с тем, чтобы первоначально проверить эти принципы на восприятии коллег и далее продуктивно использовать таким образом проверенный план при реализации проекта. Иное, т.е. работа без плана, используя лексику рецензента – просто шарлатанство.

Дело в том, что у неофитов от лингвистики часто путаются такие понятия как «корпус текстов» и «коллекция текстов». Текстовых коллекций, т.е. простых собраний (библиотек) текстовых файлов сейчас пруд пруди в Интернете. Пользоваться ими и их анализировать в некоторых отношениях может любой, имеющий доступ в Интернет. Корпус же текстов – это более организованная и структурированная на основе некоторых принципов, а также многоаспектно размеченная совокупность текстов. Он дает дополнительные возможности лингвистам для разнообразного анализа текстов и их единиц.

Видел ли кто-нибудь продукт Украинского языково-информационного фонда под названием «Украинский национальный лингвистический корпус»? Каковы принципы его создания и основные структурные характеристики? Каковы результаты анализа? По крайней мере, опубликованная в 2005-ом году (на самом деле, в 2006-ом) коллективная монография под общей редакцией В. Широкова «Корпусная лингвистика» (К.: Довіра, 2005) не содержит ясно сформулированных принципов и конструкторских решений для создания такого корпуса. Так, например, в разделе 4 «Інформаційні та лінгвістичні засади побудови українського національного лінгвістичного корпусу» (ст. 103-193) речь идет о лингвистических проблемах и лингвистических технологиях информационного общества и общества знаний; о принципах формирования и основных направлениях использования и развития УНКЛН (Украинского национального лингвистического корпуса); о грамматической подсистеме и принципах ее моделирования; о парадигматизации и лемматизации; о средствах семантической разметки. Все эти разделы с точки зрения корпусной лингвистики не являются проблемами концептуализации корпуса. Не менее далеки от концепции, а также и от корпусной лингвистики формулировки типа: «зрозуміло, що кожному суспільно-виробничому укладу відповідає характерний для цього укладу комплекс технологічних інструментів. Зокрема, для інформаційного суспільства це були інформаційно-комунікаційні технології. Для суспільства знань це, очевидно, мусять стати технології оперування знаннями причому на всьому її життєвому циклі...» (ст. 106); «Інформація є об'єктивною характеристикою об'єктивних процесів. На нашу думку, існує глибока аналогія між визначенням понять інформації, з одного боку та енергії, яка також є певною об'єктивною характеристикою, – з іншого.» (ст. 107); «Однією з домінант новітньої історії, як відомо, є змагання за джерела енергії. Особливого значення та гостроти ця проблема набула з розвитком технотронної ери, коли виявилося, що джерела енергії не є невичерпними й у 70-х роках у світі спалахнула енергетична криза» (ст. 113) и далее «Склалося враження, що відбувається збільшення виробництва при зменшенні вкладених в нього ресурсів» (ст. 113); «УНЛК належить до типу мовно-інформаційних систем, складна структура та широкий набір функцій якої спонукають до якомога глибшого осмислення, насамперед, його лінгвістичних принципів» (ст. 120). Но самыми парадоксальными в разделе, посвященном концепции создания текстового корпуса национального языка, являются даже не размышления о

глобальном мировом кризисе и его преодолении, а таблицы, присутствующие в нем: табл. 4.1 «Зіставлення властивостей енергії та інформації» (ст. 107), табл. 4.4 «Динаміка енергоспоживання в розвинених країнах у 1970-1982 рр. (в млн. тонн нафтового еквівалента)» (ст. 114).

Все это не только не является концепцией текстового корпуса, но вводит в заблуждение читателя, поскольку происходит профанация корпусной лингвистики и подмена вымыслами В. Широкова об энергии реальных проблем и задач корпусного направления в лингвистике.

Что же касается утверждения в рецензии В. Широкова о надуманности целей, которые поставила перед собой Демская-Кульчицкая в своей монографии, то с этим никак нельзя согласиться. В монографии представлена стройная, логически непротиворечивая концепция структуры будущего корпуса, сфер его использования, его предметней области, очерчены первостепенные задачи, которые необходимо решить на первом этапе его создания и последующие, касающиеся формирования реализационной стратегии (как создания корпуса, так и его использования), разработана детальная схема маркирования слов текста (определения их грамматических характеристик). Иными словами, в монографии определены лингвистические предпосылки и задачи создания корпуса, частично создана база лингвистической характеризации для будущего маркирования единиц текста. Большего от концепции вряд ли нужно требовать. Что же касается построения автоматических или автоматизированных способов самого маркирования, процесса его осуществления, то это – задача следующего этапа – этапа выработки различных способов эксплуатации и развития созданного корпуса. При этом способы маркирования могут меняться (в т.ч. и автоматизироваться во все большей степени, т.к. может накапливаться и формализоваться информация для построения алгоритмов анализа текстов на основе статистической обработки ранее проведенной и верифицированной разметки представительных обучающих выборок), однако сама концепция характеристической базы разметки остается той же самой, если она с самого начала имела твердые лингвистические основания.

Рецензент В. Широков считает недостатком монографии О. Демской-Кульчицкой то, что в ней много места уделено обзору чужих работ. Я же думаю, что это не недостаток, а достоинство работы. А как можно создавать концепцию, не ознакомившись с уже существующим опытом коллег по созданию других текстовых корпусов? Только на основе учета опыта других и можно не только избежать некоторых часто возникающих ошибок, но и создать такой корпус, который оказался бы уникальным, выполняющим свою, особую роль. В данном случае – это роль наиболее представительного корпуса текстов современного украинского языка, который на основе первичной разметки основными морфологическими и лексико-семантическими категориями и последующего многостороннего его анализа-развития содержал бы многоуровневую и точную информацию о различных морфологических, лексических, словообразовательных, синтаксических и иных единицах текстов, употребленных в различных охарактеризованных жанрово-стилевых сферах их бытования. Именно задачу подготовки развернутого плана работ по созданию подобного корпуса и поставила перед собой О. Демская-Кульчицкая в своей монографии. И неплохо справилась с нею. Конечно, есть в ее монографии некоторые спорные места, а также отдельные решения, требующие более детального обоснования (в т.ч., иногда немотивированное употребление англоязычных терминов, вместо существующих украинских), но эти аспекты не настолько значимы, чтобы можно было сделать вывод: «ніякої концепції національного корпусу української мови у авторки немає». Просто, рецензент В. Широков, не будучи лингвистом, не увидел тех лингвистических проблем, которые подняты и решены в монографии (и без решения которых вести работы по созданию большого корпуса любого языка просто невозможно). Так бывает, когда представитель технических или физико-математических наук возомнит себя специалистом в такой непростой предметной науке, как языкознание, и пытается заменить реальные проблемы, существующие в ней, и методы, достаточно отработанные и очевидные в ней, своим ограниченным, а значит, и непрофессиональным видением этих проблем и методов.

Возможно, такое критическое рвение рецензента вызвано тем, что он усмотрел конкуренцию создаваемого с участием О. Демской-Кульчицкой корпуса тому корпусу, который создан под его руководством? Видимо, поэтому он и пытается доказать, что никакого иного корпуса, чем тот, который у него, создавать не нужно. Эта позиция не

только не свидетельствует о широком научном кругозоре рецензента, но и, возможно, говорит о его мафиозно-монополистических замашках. Будущий Национальный корпус украинского языка, концепция которого представлена в монографии О. Демской-Кульчицкой, коренным образом будет отличаться от корпуса, создаваемого в Украинском языково-информационном фонде, в частности, если учесть то некорректное определение корпуса, которое дано В. Широковым в его рецензии.

Наличие в любой национальной лингвистической традиции не одного, а многих текстовых корпусов – не недостаток, с которым надо бороться, а нормальное, даже очень полезное для развития этой области состояние. Пытаться подавлять коллег-конкурентов может лишь тот, кто возомнил себя либо носителем абсолютной истины в данной области, либо тот, кто считает, что главное в науке – только застолбить какой-либо предметный участок, заявить на него свои, неизвестно откуда взявшиеся, эксклюзивные права на него, чтобы туда никому другому доступа не было. Эта позиция, естественно, далека от норм научного общежития. Все-таки, во все времена нормой было искать истину, а не огораживаться, критиковать коллег (если есть предмет для критики), корректно конкурировать с ними, а не пытаться их уничтожить. Силовые методы приходят в науку из других областей и долго в ней не задерживаются.

Не будем далеко ходить за примером – в России уже существует или создается несколько языковых корпусов современного русского языка: Национальный корпус русского языка, Полистилевой корпус современного русского языка, Большой корпус русского языка, Корпус текстов А. С. Пушкина и др. Аналогично в США, Англии, Германии и других странах, в том числе и в других славянских, создано по несколько корпусов различного целевого назначения. Каждому ученому, если это действительно серьезный ученый, понятно, что ни одна научная проблема не может быть раз и навсегда решена каким-то одним ученым или даже коллективом. А уж на корпусном материале столько разных проблем может быть поставлено и решено! Почему же на Украине кто-то должен иметь монопольное право на истину, а также на приватизацию для кого-то лакомых участков территории науки? Наука – самая свободная и открытая сфера деятельности человека. Поэтому "стоять насмерть" на застолбленном участке и тормозить работы других коллективов не пристало ученому. Да и бесперспективно такое огораживание и стояние.

Разумеется, степень эффективности каждого из создаваемых Корпусов покажет время, хотя многое, как показывает уже имеющаяся практика, вытекает из концепции того или иного корпуса (если она, конечно же есть).

В заключение своей рецензии В. Широков упрекает члена-корреспондента НАН Украины И. Р. Выхованца и Е. Г. Городенскую за то, что они не выступили против рекомендации к печати монографии О. М. Демской-Кульчицкой: "Невже і ви знайшли у пані Орисі щось нове в українській морфології?" (ст. 83). Этот выпад рецензента свидетельствует о его полном непонимании задач как общей, так и корпусной лингвистики, а именно в той части, что разметка того или иного корпуса морфологическими пометами – это не открытие нового в морфологии, а реализация наиболее точных схем морфологической квалификации словоформ и лексем, выработанных в лингвистике к данному моменту в данной лингвистической традиции. Только на основе такой первичной разметки и возможен дальнейший углубленный лингвистический анализ морфологических проблем, например, того, какие семантические классы лексем, в каких формах, в каких синтаксических условиях (в сочетании с лексемами каких семантических и морфологических классов, в каких синтаксических конструкциях), в каких типах текстов и с какой частотой реализуют те или иные свои морфологические категории. Т.е. только на этой основе и возможно то искомое ПОЛУЧЕНИЕ НОВОГО ЗНАНИЯ, не известного ранее, которое и является целью науки. Корпусная лингвистика дает возможность для получения такого нового знания, т.к. в структуре корпусного материала синтезируется, с одной стороны, информация о естественных единицах и их конфигурациях в тексте с категориальной информацией, накопленной об этих единицах в лингвистике и закрепленной в корпусе в ходе первичной разметки текста, с другой. Автоматизированное обследование коррелятивных связей между этими и иными динамически добавляемыми типами информации приводит к выявлению скрытых структур, открытию нового знания, а точнее, подтверждению или опровержению той или иной гипотезы о структурах текстов определенного содержательного и стилевого диапазона, которая вытекает из той или иной концепции.

Если В. Широков с такой постановкой вопроса не согласен, то это означает либо то, что у него имеется своя, правда, никому не известная НОВАЯ концепция украинской морфологии (что маловероятно, если судить по его работам), либо он не понимает того, о чем он пишет. Второе – более вероятно.

И последнее. В. Широков в конце рецензии пишет, что она была обсуждена и одобрена на заседании ученого совета организации «Український мовно-інформаційний фонд». Хочется спросить: неужели среди членов ученого совета этой организации не нашлось специалиста, который указал бы рецензенту на ошибки в его тексте? И в редколлегии «Мовознавства» почему-то никто не заметил несуразности текста, отнюдь не поднимающего авторитет журнала? Но, как оказывается, В. Широков является заместителем редактора этого журнала. Конечно, внутреннее дело академического института, кого кем избирать или назначать в редакцию профильного для Института языкознания НАНУ органа, но закономерно возникает вопрос: неужели на Украине такой дефицит языковедов, что заместителем редактора ведущего журнала пришлось сделать доктора технических наук, явно не разбирающегося в лингвистической проблематике, что хорошо видно по написанной им рецензии, а также по его трудам? А куда смотрел редактор этою журнала? Академик В. М. Русановский достаточно серьезный ученый, он же не мог не увидеть, что рецензия В. Широкова далека от науки, от профессионализма, от нормальных принципов научной дискуссии. Почему же он одобрил ее публикацию? Чем можно объяснить такую нетребовательность?

20. 04. 2006.

РІЗНЕ

ЛЕКСИКОГРАФІЧНИЙ БЮЛЕТЕНЬ

Шановні колеги!

Відділ лексикології та комп'ютерної лексикографії Інституту української мови НАНУ запрошує Вас до участі в збірнику наукових праць **"Лексикографічний бюлетень",** де передбачені рубрики:

1. **Лексикографія, лексикологія: теорія та практика.**
2. **Комп'ютерна лінгвістика.**
3. **Персоналії**.
4. **Рецензії, анотації.**
5. **Хроніки**.
6. **Різне**.

Вимоги до оформлення матеріалів:

– обсяг до 0, 5 др. арк.;
– поля: праве – 1, 5 см; ліве, верхнє та нижнє – по 2, 5 см.;
– гарнітура Times New Roman; розмір 14; інтервал звичайний;
– абзацний відступ 0, 5 см, міжрядковий інтервал 1, 5;
– ініціали в межах статті подавати ***уніфіковано*** через знак "нерозривний проміжок" (Ctrl+Shift+проміжок): М. Лісовий або М. П. Лісовий;
– назву друкувати по центру з великої літери;
– відомості про автора та **УДК** розміщувати праворуч над назвою:

Ольга Петренко, к. філол. н.
Інститут української мови (Київ)
УДК 161.2.81'374.46
Один з аспектів лексикографічної фіксації термінів

і також на окремому аркуші, зазначивши ***адресу*** й ***контактний телефон***;
– після назви статті обов'язкове ***резюме*** українською мовою (5–10 рядків);
– список використаної літератури подавати ***нумеровано*** в кінці статті ***за алфавітом***, посилання в тексті оформляти у ***квадратних дужках***, напр.:

..національної культури й є новим етапом у розвитку її традиції [1: 23; 2].

Література

1. Русская рок-поэзия: текст и контекст // Сб. науч. трудов. – Тверь, 1999.– Вып. 2.
2. Старченко О. М. Про сучасну..;

– список джерел, скорочень подавати в алфавітному порядку, після списку літератури, причому **між скороченням і повною назвою джерела –** знак **табуляції,** напр.:

СДЯ Словарь древнерусского языка (XI – XIV вв.) / Гл. ред. И. Аванесов. – М., 1989. Т. 1 – 6.
Сл.РЯ Словарь русского языка XI – XVII вв. – М.: Наука, 1975. Вып. 1 – 26.
Служ. Варл. XII Служебник Варлаама ХутинськогоXII в. / Горский А. Описание славянских рукописей Московской синодальной библиотеки. – М.,

– посилання на джерела робити у круглих дужках, напр.:

.../ Далі йду тим шляхом до мети/ І ***несу покути свій тягар*** (Забаштанський: 45).

– назву, абзаци, відомості про автора та інше не форматувати за допомогою знака проміжку.
– матеріали подавати **на дискеті 3, 5** у вигляді файлу у форматі Word 97 (або пізніша версія) та в друкованому варіанті (у двох примірниках);
– **спеціальні шрифти** (у разі їх використання) надсилати разом із файлом статті.

Бюлетень визнано **ВАК як фахове видання** за спеціальністю 10.02.01 – українська мова (див. "Бюлетень Вищої атестаційної комісії України", 2005, вип. 9: 21).

Вартість однієї сторінки – 6 грн.

Матеріали до **14 випуску** просимо надсилати за адресою:

О. Тищенко, відділ лексикології та комп'ютерної лексикографії, Інститут української мови НАНУ, вул. М. Грушевського, 4, м. Київ, 01001, Україна (тел.: 8-097-339-836-7) ***до 1 вересня 2006 року.***

Редакційна колегія

Наукове видання

Лексикографічний бюлетень

Л 43 Лексикографічний бюлетень: Збірник наукових праць / Відпов. ред. д. філол. н., проф. В. Німчук.– К., 2006. – Вип. 13. – 208 с.

ISBN 966-02-3450-3 (серія)
ISBN 966-02-4002-3 (вип. 13)

Свідоцтво про державну реєстрацію
друкованого засобу масової інформації КВ №10023 від 04.07.2005
Оригінал-макет О. Тищенко
Підписано до друку 22.11.05
Гарнітура Таймс. Наклад 500 прим. Умов. друк. арк. 18
Надруковано в типографії СПД О. Сидоренка
вул. Деревообробна, 4, Київ.