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 вималювалася 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. Е**кспериментальний 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. І**ндуктивний 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**, суряднiсть, 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 понятєвого 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*