Є.А. Карпіловська

ВСТУП ДО ПРИКЛАДНОЇ ЛІНГВІСТИКИ:
КОМП'ЮТЕРНА ЛІНГВІСТИКА

Підручник

Донецьк
Юго-Восток
2006

УДК 81’33(075)
ББК Ш1р30-253.1я73+Ш111я73
К21

Рецензенти:

Бацевич Ф.С. — доктор філологічних наук, професор, зав. кафедри загального мовознавства (Львівський національний університет імені Івана Франка)

Клименко Н.Ф. — доктор філологічних наук, професор, провідний науковий співробітник відділу структурно-математичної лінгвістики (Інститут мовознавства ім. О.О. Потебні НАН України)

Відповідальний редактор:

Загнітко А.П. — доктор філологічних наук, професор

Затверджено Міністерством освіти і науки України як підручник для студентів філологічних спеціальностей вищих навчальних закладів (лист № 14/18.2-2729 від 05.12.2005 р.)

Карпіловська Є.А.

К21 Вступ до прикладної лінгвістики: комп’ютерна лінгвістика: Підручник.— Донецьк: ТОВ «Юго-Восток, Лтд», 2006.— 188 с.

ISBN 966-374-078-7

Пропонований підручник подає основи нового напрямку теоретичних і прикладних досліджень сучасного мовознавства — комп’ютерної лінгвістики. Читачі у першій частині підручника — «Основи комп’ютерної лінгвістики» — познайомляться з об’єктом, предметом і методами цієї лінгвістичної дисципліни, засадничими поняттями її термінологічного апарату. Друга частина — «Лінгвістичні комп’ютерні інтелектуальні системи» — містить опис основних типів комп’ютерних систем з лінгвістичним забезпеченням.

Підручник адресований студентам-філологам, які здобувають спеціальність «Прикладна лінгвістика», викладачам вузів, аспірантам, усім, хто цікавиться проблемами комп’ютерного опрацювання мовної інформації.

УДК 81’33(075)
ББК Ш1р30-253.1я73+Ш111я73

ISBN 966-374-078-7

ЗМІСТ

ПЕРЕДМОВА: ДЛЯ КОГО І ЯК НАПИСАНО ЦЮ КНИЖКУ

Комп'ютер дедалі більше стає невід'ємним складником життя нашого суспільства. Перефразовуючи слова, сказані М.С.Грушевським з приводу значення латини та польської мови для освіченого українця XVI–XVII століть, можна твердити, що без комп'ютера ми скоро взагалі не зможемо "порушитися в практичнім обіході"[1], а в галузі оброблення інформації, передусім мовної, і поготів. Комп'ютер, як раніше телефон, радіоприймач, магнітофон, диктофон і телевізор, проникає в усі сфери життя сучасної людини, стає незамінним помічником в її інтелектуальній, пізнавальній, професійній діяльності як потужний засіб генерування, опрацювання та передавання інформації, різноманітних знань, а також як зручний і оперативний засіб обміну інформацією або знаннями. Роль комп'ютера як інструмента пізнання дійсності зростає передусім у виробничій та науковій сферах суспільного життя. Оскільки робота з комп'ютером становить один з різновидів комунікації, спілкування, то лінгвістам відведено особливе місце в постійному вдосконаленні як самого комп'ютера, його програмного забезпечення, так і в створенні нових засобів опрацювання інформації, нових інформаційних технологій. Крім того, використання комп'ютера для розв'язання власне лінгвістичних завдань виводить на нові обрії розвитку і саму лінгвістику, докорінно змінює технологію дослідницької праці, відкриває перед лінгвістами нові можливості у вивченні будови та функціонування мови. Комп'ютер не тільки звільняє користувача від виконання технічної роботи, пов'язаної з пошуком, добором та впорядкуванням інформації, але й становить для лінгвіста-дослідника новий тип адресата його роботи, а отже, відкриває можливості для якісно нового вивчення мови. Комп'ютеру – технічному пристрою, який не має людської пам'яті, знань, асоціацій, інтуїції, людської здатності мислити – мовний матеріал треба описати і подати для сприйняття в інший спосіб, в іншій формі. А такий інший погляд на об'єкт дослідження, інший спосіб його опису завжди відкриває його нові аспекти, нові грані, насичує дослідження новими спостереженнями та висновками.

Цю книжку й адресовано студентам-філологам, які своїм фахом обрали комп'ютерну лінгвістику, а отже, тим читачам, які не лише цікавляться лінгвістичними проблемами комп'ютеризації суспільства, а

[1] **Грушевський М.С.** Культурно-національний рух на Україні в XVI-XVII віці // Грушевський М.С. Духовна Україна. – К., 1994. – С.148.

прагнуть займатися цими проблемами професійно, прагнуть стати розробниками комп'ютерних засобів вивчення мови або ставлять собі на меті оволодіти такими засобами для розв'язання теоретичних і практичних завдань сучасного українського мовознавства. У колі своїх майбутніх співбесідників бачимо передусім філологів-дослідників української мови, для якої комп'ютеризація відкриває заманливі перспективи оперативного та успішного розв'язання вкрай актуальних проблем розбудови та внормування її лексикону й граматичного ладу, накопичення та опрацювання різноманітної інформації про стан і закономірності її функціонування в різних соціальних, регіональних, стильових, комунікативних виявах, а завдяки цьому й вироблення гнучких і ефективних рекомендацій стосовно мовного планування, вірогідних прогнозів про можливі тенденції і напрямки розвитку української мови як мови держави, мови, що покликана забезпечити всі потреби життя сучасного українського суспільства.

Ми ставили собі на меті познайомити читачів з предметом і методами аналізу мовного матеріалу, які вирізняють сучасну комп'ютерну лінгвістику з-поміж інших лінгвістичних дисциплін, ввести їх у проблематику цієї мовознавчої дисципліни, подати зразки опрацювання конкретних дослідницьких завдань за допомогою комп'ютера. Дві частини книжки – "Основи комп'ютерної лінгвістики" та "Лінгвістичні інтелектуальні комп'ютерні системи" – і представляють предмет, завдання, понятєвий та процедурний апарат комп'ютерної лінгвістики як, з одного боку, галузі сучасного мовознавства і, з другого, – як галузі сучасної інформатики, як складника досліджень, спрямованих на створення автоматизованих інтелектуальних систем, або систем зі штучним інтелектом. Наскільки це було можливим, обговорюючи ту чи іншу конкретну проблему, ми спиралися на дослідження, виконані вченими України або здійснені на українському мовному матеріалі. Ми прагнули показати читачам досягнення української комп'ютерної лінгвістики і завдяки цьому висвітлити стан розроблення в ній тієї чи іншої проблеми, окреслити ті завдання, які ще чекають свого розв'язання. Однак ми не забували про настанову Великого Кобзаря – не лише свого не цуратися, але і чужого научатися. Саме тому читачам подано відомості й про визначні здобутки зарубіжної комп'ютерної лінгвістики, важливі й цікаві для обговорюваної проблематики.

До кожного розділу додано словнички основних термінів, оскільки рівень володіння термінологічним апаратом певної наукової дисципліни, як відомо, засвідчує рівень обізнаності з її предметом. У списку літератури в кінці книжки вміщено праці, знайомство з якими дозволить читачам глибше проникнути в суть обговорюваних проблем, знайти додаткову поживу для своїх роздумів і пошуків у розв'язанні мовознавчих завдань за допомогою комп'ютера, а то й просто одержати відповіді на

свої питання. Тож сподіваємося, що наша книжка виконає роль дороговказу у таємничий, але принадний світ комп'ютерної лінгвістики, в якому допитливі і вдумливі читачі вже самостійно шукатимуть свої власні шляхи.

Автор складає сердечну подяку колегам, які взяли на себе труд прочитати рукопис книжки і своїми зауваганами та порадами допомогли поліпшити виклад порушених у ній проблем: завідувачу кафедри української мови Донецького національного університету доктору філологічних наук професору Анатолію Панасовичу Загніткові, доцентам цієї кафедри кандидатам філологічних наук Любові Дмитрівні Фроляк та Михайлові Олексійовичу Вінтоніву; завідувачу кафедри загального мовознавства Львівського національного університету імені Івана Франка доктору філологічних наук професору Флорію Сергійовичу Бацевичу, провідному науковому співробітнику відділу структурно-математичної лінгвістики Інституту мовознавства ім.О.О.Потебні НАН України доктору філологічних наук професору Ніні Федорівні Клименко; завідувачці лабораторії комп'ютерної лінгвістики кафедри сучасної української мови Київського національного університету імені Тараса Шевченка доценту кандидату філологічних наук Наталії Петрівні Дарчук та фундатору цієї лабораторії і її першій завідувачці доценту кандидату філологічних наук Людмилі Антонівні Алексієнко; студентам-україністам Київського та Донецького національних університетів, які здобували спеціальність "Прикладна лінгвістика" і в постійному спілкуванні з якими народжувалася ця книжка. Особливу подяку автор висловлює деканату філологічного факультету Донецького національного університету і особисто декану доктору філологічних наук професору Євгенові Степановичу Отіну за створення надзвичайно сприятливої атмосфери, в якій і стали можливими написання та видання цього підручника.

РОЗДІЛ І. ОСНОВИ КОМП'ЮТЕРНОЇ ЛІНГВІСТИКИ

§1. Комп'ютерна лінгвістика (КЛ) – новий етап розвитку теоретичної та прикладної лінгвістики

- Об'єкт, завдання і методи дослідження КЛ
- КЛ у відношеннях з іншими науковими дисциплінами
- Інформація та її різновиди
- Формалізація інформації внутрішня і зовнішня
- Лінгвістичне забезпечення комп'ютера (lingware)

Поява нового напрямку в науці – явище не випадкове. Його спричинюють, як і кожне явище, чинники зовнішні (позамовні, позалінгвістичні) і внутрішні (суто мовні, суто лінгвістичні), об'єктивні й суб'єктивні. Отже, ***комп'ютерна лінгвістика (КЛ)*** – явище, причини виникнення якого слід шукати в стані розвитку як сучасного суспільства, так і самої мови та науки, що її вивчає, – лінгвістики. За точку відліку у наших роздумах візьмемо визначення комп'ютерної лінгвістики, загальноприйняті у фаховій літературі:

- **"Термін "комп'ютерна лінгвістика" задає загальну орієнтацію на використання комп'ютерів для розв'язання різноманітних наукових та практичних завдань, пов'язаних з мовою, аж ніяк не обмежуючи способи розв'язання цих завдань"[2].**
- **"Під терміном "комп'ютерна лінгвістика (computational linguistics) звичайно розуміють широку царину використання комп'ютерних інструментів – програм, комп'ютерних технологій організації та оброблення даних – для моделювання функціонування мови в тих чи інших умовах, ситуаціях, проблемних галузях, а також сферу застосування комп'ютерних моделей мови не лише в лінгвістиці, а й у суміжних з нею дисциплінах"**[3].

З поданих визначень спробуємо окреслити об'єкт, предмет КЛ, методи дослідження мови, тобто складники будь-якої самостійної наукової дисципліни. Не забуваймо при цьому знамениту настанову Ф.де Сосюра, що погляд на об'єкт формує сам об'єкт. Аспект, в якому КЛ вивчає мову, і застосовуваний нею інструмент такого вивчення – комп'ютер – і вирізняють її з-поміж інших лінгвістичних дисциплін, формують особливий погляд на мову як об'єкт дослідження. За змістом та обсягом опрацьовуваних завдань компьютерна лінгвістика стає аналогом лінгвістики традиційної, або некомп'ютерної, оскільки

---

[2] **Городецкий Б.Ю.** Компьютерная лингвистика: моделирование языкового общения // Новое в зарубежной лингвистике. Компьютерная лингвистика. - М., 1989. - Вып. XXIV. - С.10.

[3] **Баранов А.Н.** Введение в прикладную лингвистику. – М., 2001. - С.13.

об'єкт дослідження КЛ становить мова в усіх трьох способах свого існування:

- ***мовна система*** – сукупність певних одиниць з властивими їм формальними, змістовими та функціональними властивостями;
- ***мовлення*** – різноманітні продукти реалізації мовної системи в певних умовах та ситуаціях комунікації;
- ***мовна діяльність*** – процес використання мовної системи й створення продуктів такого застосування мови в тих чи інших умовах комунікації.

Як бачимо, об'єктом розв'язання конкретних завдань у межах КЛ може бути кожен зі складників відомої тріади Ф.де Сосюра : – **langue** "мова" – **parole** "мовлення" – **langage** "мовна діяльність".

Від традиційної, некомп'ютерної, лінгвістики КЛ відрізняє особливий спосіб вивчення об'єкта за допомогою комп'ютера. Він зумовлює необхідність розроблення методів та прийомів аналізу й синтезу мовних фактів, відмінних від традиційних. Недарма у вищеподаному визначенні КЛ з праці А.М.Баранова з'явилося слово "моделювання". Комп'ютер, і цього не слід забувати, – потужний, швидкий, зручний, але ... технічний засіб опрацювання інформації взагалі і мовної інформації, зокрема. Підкреслимо, технічний засіб, а отже, він позбавлений того апарату сприйняття, породження й використання мови, яким володіє людина, позбавлений передусім інтелекту людини, її знань та вмінь, що створюють у мозку людини свою особливу модель реального світу. Тут ми відразу застережемо читачів від помилок двох типів при застосуванні комп'ютера в лінгвістичних дослідженнях: як від недооцінювання його можливостей, так і від їхнього перебільшення. Ці два типи помилок у міждисциплінарних дослідженнях свого часу дуже дотепно й влучно визначив відомий російський генетик О.О.Любищев. Аналізуючи типові помилки у застосуванні математики в біології, він так і назвав дві свої статті, присвячені цій проблемі і опубліковані 1969 р.: "Помилки від браку обізнаності" та "Помилки, пов'язані з надміром ентузіазму"[4]. Використання комп'ютера в ролі дослідника мови неможливе без створення відповідної, "зрозумілої" йому моделі реального світу, моделі реальних об'єктів та процесів, придатної для опрацювання тими засобами, які є в арсеналі комп'ютера. Цим і зумовлений окремий, самостійний предмет дослідження КЛ.

Предмет дослідження КЛ – ознаки будови, змісту та функціонування одиниць мовної системи, продуктів мовлення та мовної діяльності – звукових або письмових текстів, які могли б служити для їхнього моделювання й використання в процесах комп'ютерного опрацювання мовної інформації

[4] Подаємо за працею: **Российская** научная эмиграция: Двадцать портретов / Под ред. академиков Г.М.Бонгарда-Левина и В.Е.Захарова.– М., 2001. – С.256.

Комп'ютерна лінгвістика є водночас дисципліною теоретичною, фундаментальною і практичною, прикладною. Застосування нового дослідницького інструмента – комп'ютера – дає змогу на якісно новій фактичній основі, за допомогою нових дослідницьких методів та прийомів аналізу й синтезу мовних фактів розв'язувати теоретичні завдання сучасного мовознавства, одержувати відомості про будову та функціонування мовних об'єктів, які уможливлюють зміни в самій парадигмі знання про мову, оскільки дозволяють не лише по-іншому розв'язувати старі, а й ставити нові лінгвістичні завдання. Використання спеціального технічного засобу і підпорядкованих йому прийомів та процедур моделювання мовних об'єктів ставить КЛ в ряд інших прикладних лінгвістичних дисциплін, таких, наприклад, як лінгводидактика, логопедія, експериментальна фонетика, стенографія, нейро- та біолінгвістика. Врахування комплексного характеру предмета дослідження КЛ, специфіки використання в її дослідженнях особливого технічного засобу дає підстави сформулювати таке її визначення:

Комп'ютерна лінгвістика – самостійна лінгвістична дисципліна, яка розв'язує теоретичні й прикладні завдання мовознавства за допомогою комп'ютера

Передумовою будь-якого комп'ютерного моделювання є формалізація мовного матеріалу, тобто пошук засобів формального, унаочненого, приступного сприйняттю комп'ютера представлення досліджуваного мовного об'єкта – одиниці мови або процесу з її участю. Такими формальними ознаками будови мовних об'єктів можуть служити кількісні, сполучувальні (комбінаторні) й позиційні характеристики їхніх складників. Наприклад, морфемну будову сучасного українського слова можна змоделювати, врахувавши загальну кількість морфем в його складі чи кількість морфем окремих класів і типів (кореневих і афіксальних або флективних), позиції тих чи інших морфем у слові та способи їхнього комбінування. За такими ж формальними ознаками можна змоделювати будову мовних одиниць і на інших рівнях мовної системи. Скажімо, будову складу або словосполуки чи й цілого речення або тексту. Для побудови різноманітних комп'ютерних моделей необхідно накопичити достатню, а в ідеалі й вичерпну інформацію про такі формальні ознаки будови й функціонування різнотипних мовних одиниць. Отже, треба мати відомості – знання – про те, як побудовані мовні одиниці, що вони становлять, і відомості – знання, про те, як вони функціонують, вживаються, "поводять себе" у мовленні, в процесах використання системи мови при комунікації. Знання першого типу визначають у КЛ як ***декларативні,*** або ***статичні***, знання другого типу – як ***процедурні,*** або ***динамічні***.

Наприклад, про морфемну будову сучасного українського слова накопичені знання як декларативні, так і процедурні. Перші становлять відомості про типи і класи морфем у складі слів окремих частин мови та лексико-граматичних розрядів, їхню буквену (графічну) та фонетич-

ну (звукову) структуру, загальну кількість морфем у слові та кількість морфем окремих типів або класів, їхній розподіл у слові за позиціями відносно стрижневої морфеми українського слова – кореня, а також про схеми комбінування морфем у словах. Другі складають відомості про вживання слів з тими чи іншими морфемами в різних типах текстів, їхній розподіл за частинами мови або лексико-граматичними розрядами, за тематичними групами чи функціональними стилями мови, певними сферами життя суспільства й діяльності людини (спеціалізовані, термінологічні різновиди мови, її підмови). Так, інформація про те, що простий іменник сучасної української мови можна побудувати за 23 моделями комбінування кореня з префіксами, суфіксами та флексіями, є знанням декларативним, а повідомлення про те, що найпотужнішою з цих моделей, тобто такою, яка реалізована у переважній більшості слів сучасного українського лексикону, є модель **RSF** (корінь+суфікс+флексія), є знанням процедурним. Така модель комбінування класів морфем втілена майже у чверті всіх простих іменників, засвідчених сучасними українськими словниками з найпоказовішими за обсягом та складом реєстрами (у 7548 словах з 32075)[5]. Відомості про те, що суфікс **-іт-** у слові ***без-роб-ітн(ий)*** займає 1-у післякореневу позицію, є двофонемним, має фонемну структуру зразка **ГП** (голосний+приголосний), становить наголошений склад у цьому слові, є декларативним знанням про цей суфікс, адже вони нічого нам не повідомляють про правила вживання цієї афіксальної морфеми або її розміщення в слові, комбінування з іншими складниками слова (суфіксами, флексіями чи коренями), закономірності пристосування до них. Таку інформацію надають знання процедурні. З них довідуємося, що цей суфікс є формальним варіантом – аломорфом – суфікса **-от-** (пор. ***без-роб-ітн(ий)*** ***←роб-от(а)***) і з'являється внаслідок комбінування з суфіксом **-н-**, що закриває цей відкритий склад і спричинює чергування голосних **о/і** в суфіксі **-от-**.

КЛ у своєму поняттєвому й методико-процедурному апараті опису та моделювання мовної інформації спирається на здобутки не лише традиційної, теоретичної лінгвістики, а й творчо розвиває напрацювання лінгвістики структурної та математичної. Саме останні накопичили

[5] Зацікавлених читачів відсилаємо до праць, в яких такі декларативні й процедурні знання про морфемну будову сучасного українського слова подані докладно за матеріалами комп'ютерного морфемно-словотвірного фонду української мови, сформованого в Інституті мовознавства ім.О.О.Потебні НАН України. Обсяг генерального реєстру слів у цьому фонді становить близько 167 тис. лексем, з яких простих слів – понад 130 тис. Див.: **Клименко Н.Ф., Карпіловська Є.А.** Морфемні структури слів у сучасній українській літературній мові // Мовознавство. – 1991. - № 4. – С.10-21; **Карпіловська Є.А.** Морфемна сітка як інструмент дослідження будови слова // Українське мовознавство. – 1992. – вип.19. – С.100-110.

цінний і корисний для КЛ арсенал декларативних та процедурних знань про мовні об'єкти та їхнє функціонування, оскільки особливу увагу приділяли формі представлення знань, виробленню методик і процедур опису мовних фактів з опорою на формальні ознаки їхньої будови, змісту та функціонування в системі мови або в продуктах її реалізації у різноманітних комунікативних ситуаціях. У певному сенсі КЛ можна вважати спадкоємицею саме тих галузей прикладної лінгвістики, які у своїх дослідженнях оперували формально-логічними та математичними методами, а також тих напрацювань у моделюванні й аналізі мовних об'єктів та мовленнєвих процесів, які накопичили структурна та математична лінгвістика. Розрізнення, з одного боку, традиційної, теоретичної лінгвістики, а з іншого – лінгвістики прикладної і зумовлене тими різними методами, якими вони оперують в процесі аналізу й опису мовного матеріалу, а також самим призначенням результатів таких лінгвістичних досліджень. Румунська дослідниця Тетяна Слама-Казаку так визначила принципову відмінність між теоретичною, або "чистою" та прикладною лінгвістикою: перша займається встановленням певних загальних теоретичних принципів, що належать окремим аспектам конкретної мови або й феномену "мови" взагалі, а друга розв'язує конкретні практичні завдання[6]. Комп'ютерна лінгвістика з її спрямуванням на розв'язання як теоретичних, так і практичних завдань виявляється ширшою за прикладну лінгвістику в такому її традиційному розумінні. Вона становить нову комплексну лінгвістичну дисципліну на перетині лінгвістики теоретичної та прикладної.

З появою комп'ютера кардинально змінилося розуміння таких понять, як ***інформаційна сфера, інформаційне середовище, інформаційна технологія.*** Необхідність для нормальної життєдіяльності суспільства подальшої оптимізації процесів формування, передавання, сприйняття й зберігання інформації стимулювала становлення й бурхливий розвиток такої нової наукової дисципліни, як ***інформатика.*** В найзагальнішому трактуванні інформатику й визначають як науку про інформацію, її типи, способи і засоби породження, організації та використання. Оскільки інформація у мовній формі, тобто оформлена засобами як природних, так і штучних мов, посідає в сучасних інформаційних потоках чільне, якщо не переважне місце, лінгвістика активно залучена до створення засобів її опрацювання, передусім – за допомогою комп'ютера. З огляду на це КЛ становить також невід'ємний і важливий складник сучасної інформатики. Зв'язки КЛ з іншими лінгвістичними та нелінгвістичними науковими дисциплінами в узагальненому вигляді представимо в табл. 1 та 2.

[6] **Слама-Казаку Т.** Место прикладной лингвистики в системе наук: отношение ПЛ к "лингвистике" // Новое в зарубежной лингвистике. – М., 1983. – Вып. XII. Прикладная лингвистика. – С.25.

Таблиця 1. Комп'ютерна лінгвістика та інші лінгвістичні дисципліни

Лінгвістичні дисципліни	**Комп'ютерна лінгвістика як складник мовознавства**
Фонологія Фонетика Акцентологія Інтонологія Морфемологія та дериватологія Лексикологія та фразеологія Граматика (морфологія та синтаксис) Семасіологія	Комп'ютерна граматика Аналізатори та синтезатори усного мовлення
Лексикографія	Комп'ютерна лексикографія: комп'ютерні версії традиційних словників, автоматичні словники, словникозорієнтовані бази даних та лексикографічні процесори
Теорія та практика перекладу (перекладознавство)	Системи машинного перекладу
Лінгвістика тексту	Системи автоматичного перероблення тексту (АПТ), або опрацювання тексту (АСОТ): автоматичний морфологічний (АМА), синтаксичний (АСА) та логіко-семантичний аналіз, автоматична компресія тексту (індексування, реферування, анотування). Текстозорієнтовані бази даних: корпуси текстів та електронні картотеки (=ілюстративні корпуси, корпуси цитат), комп'ютерні словопокажчики, конкорданси, частотні словники та текстові процесори
Термінологія та термінографія	Комп'ютерні термінологічні бази даних та словники. Комп'ютерні експертні системи
Історія мови	Комп'ютерні моделі реконструкції (прогнозування) минулих станів мови (моделі комп'ютерної ретрогностики). Комп'ютерні версії історичних та етимологічних словників. Комп'ютерне дешифрування давніх писемностей
Лінгвістична стилістика Культура мови Соціолінгвістика Етнолінгвістика Лінгвокультурологія	Комп'ютерна стилеметрія, атрибуція текстів, стилістична діагностика. Системи орфографічного та орфоепічного контролю. Автоматичні редактори текстів. Комп'ютерні моделі мовної концептуалізації світу. Автоматичні лінгвокраїнознавчі, етнолінгвістичні та лінгвокультурологічні бази даних і словники

Таблиця 2. Комп'ютерна лінгвістика та нелінгвістичні дисципліни

Нелінгвістичні дисципліни	Комп'ютерна лінгвістика як складник інформатики та систем штучного інтелекту
Політологія	Політична лінгвістика, моделі впливу й оцінки в процесах мовного спілкування, моделі мовного планування та будівництва
Соціологія, культурологія, етнологія, теорія комунікації	Моделювання процесів взаємодії соціо- та лінгводинаміки, ментальних стереотипів, прототипів, етнокультурної специфіки мовної категоризації дійсності, моделі мовного спілкування за допомогою комп'ютера
Кібернетика Обчислювальна математика. Програмування	Лінгвостатистика. Квантитативна лінгвістика. Ймовірнісні моделі мови для синхро- і футурогностики. Лінгвістичне забезпечення систем АПТ, або АСОТ (лінгвістичні алгоритми та процесори). Лінгвістичне забезпечення діалогових систем Лінгвістичні проблеми побудови штучних мов з природномовною компонентою та мов програмування високого рівня
Інформатика	Інформаційно-пошукові системи. Моделі мовного кодування інформації. Стратегія створення лінгвістичних баз даних та знань. Автоматичне розпізнавання та синтез мовлення
Біологія, фізіологія, психологія, медицина	Моделі організації пам'яті людини та розумовомовленнєвих процесів. Логічні структури представлення знань Діагностика хвороб за даними порушень мовленнєвої діяльності людини

Інформація, за дефініцією тлумачного "Словника української мови" (далі – СУМ), це – "Відомості про які-небудь події, чиюсь діяльність і т.ін.; повідомлення про щось (СУМ, IV, 42). Ширше визначення інформації знаходимо в спеціальних словниках, зокрема в словниках термінів обчислювальних систем, пор.: "Інформація – це сукупність символів, або образів, що несуть змістове навантаження"[7]. Засновник української кібернетики В.М.Глушков вважав, що можна "...з одного боку, схарактеризувати інформацію як сукупність можливих відомостей, які циркулюють у природі й суспільстві, у тім числі і в створених людиною технічних системах, а з іншого боку, такий розгляд дає можливість описати її як міру неоднорідності в розподілі енергії (або речовини) в просторі та в часі (підкреслення наше – Є.К.)[8]. Уже в самому визначенні цього поняття вміщено вказівку на такі риси об'єктів, процесів, явищ довкілля, які відрізняють їх від інших, виділяють їх з-поміж інших подібних та неподібних, є показниками їхньої окремішності й неповторності, їхніми своєрідними "наличками". Наскільки важливим для інформації є

[7] **Толковый** словарь по вычислительным системам. – М., 1989. – С.233.
[8] **Глушков В.М.** Гносеологические основы математизации науки // Диалектика и логика научного познания. – М., 1966. - С. 406.

віднайдення вирізняльних ознак описуваної предметної галузі, вміння оперувати ними для вироблення правильного її сприйняття й використання для розв'язання поставлених завдань, може свідчити такий спогад одного з творців мови програмування Бейсик Джона Кемені. Його родина емігрувала до Америки з Угорщини і в час, про який він згадує, Д.Кемені ще досить посередньо володів англійською мовою. І от йому, сором'язливому 16-річному хлопцеві, довелося 1945 р. складати усний іспит в одній з нью-йоркських шкіл у р-ні Манхеттена. Він так описує цей іспит у своїх спогадах: "Мій словниковий запас був страшенно обмежений і тому в кожному запитанні я вловлював лише кілька слів. Проте оскільки це був тест, в якому для кожного запитання було запропоновано набір варіантів як відповідь, то й того, що я розумів, було цілком достатньо для побудови моделі і віднайдення правильного рішення. Я зумів відкрити цей код і одержав одну з найвищих оцінок у всьому Нью-Йорку"[9]. Отже, інформацію в найширшому розумінні цього слова можна визначити як сигнал у будь-якій формі про якийсь об'єкт або явище, сигнал, разом з тим такий, який служить виразним їхнім образом, символом, що дозволяє розпізнати такі об'єкт або явище, однозначно вирізнити їх з-поміж інших подібних і неподібних. Вочевидь, саме таке трактування найповніше відповідає значенню етимона цієї запозиченої з латини лексеми, адже **īnfōrmātio** і означає "роз'яснення, пояснення; виклад, поняття". Український учений А.О.Білецький у своїй оригінальній лінгвосеміотичній теорії запропонував детальну класифікацію типів інформації, врахувавши різноманітні чинники, умови її творення та різновиди застосування[10] (див. табл. 3):

Таблиця 3. Типи інформації за класифікацією А.О.Білецького

Ознака	Тип інформації
1. Джерело створення	антропогенна – фізіогенна
2. Форма	кодована (системна) – некодована (одинична)
3. Зв'язок із ситуацією	ситуаційна – екстраситуаційна
4. Спосіб оформлення	мовна (вербальна) – немовна (екстравербальна)
5. Сфера призначення	побутова (узуальна) – спеціальна
6. Характер змістового навантаження	логічна – естетична

Виходячи з критеріїв визначення інформації, запропонованих А.О.Білецьким, ***мовну інформацію*** можна визначити як **антропогенні** (=створені людиною) **кодовані** (=закріплені у спеціальній формі) **екстраситуаційні** (=незалежні від конкретної ситуації спілкування) **верба-**

[9] Цит. за кн. "**Язык** компьютера". – М., 1989. – С.27.

[10] **Белецкий А.А.** Семиотический аспект языковой системы // // Структурная и математическая лингвистика. - К., 1979. – вып.7. – С.11-18.

льні (=представлені у мовній, або словесній формі) повідомлення про об'єкти та явища дійсності, яким властиві також логічний, естетичний, побутовий, спеціальний та експресивний компоненти.

У теорії А.О.Білецького також вичерпно окреслений спектр функцій, виконуваних мовною системою як спеціалізованою системою обміну інформацією, або різновидом інформаційної системи. Таких функцій А.О.Білецький виділив 7:

1. ***Трансформаційна*** (від лат. **transformo** "перетворюю, змінюю форму") – перетворення інформації;
2. ***Консервативна*** (від лат. **cōnservo** "зберігаю") – зберігання інформації;
3. ***Транслятивна*** (від лат. **trānslātio** "перенос, переміщення") – передавання інформації;
4. ***Димeнсійна*** (від лат. **dīmēnsio** "вимірювання, розмір") – вимірювання інформації;
5. ***Утилітарна*** (від лат. **ūtilitās** "користь, придатність") – використання інформації;
6. ***Естетична*** (від грецьк. **αἰσθητικός** "чуттєво сприйманий") – вибір для оформлення та передавання інформації засобів, оптимальних для її сприйняття;
7. ***Метасемантична*** (від грецьк. **μετα** "зверх, понад" та **σημαντικός** "означальний") – відображення в оформленні інформації додаткового змісту, який впливає на відчуття людини, стимулює в неї певні реакції на інформацію (пор., наприклад, різні способи оформлення прохання: наказ, уклінне прохання, дружнє спонукання до дії, нейтральний заклик, побажання тощо).

Для комп'ютерного опрацювання інформація має бути представлена у придатній, доступній для "розуміння" комп'ютера формі. Відтак істотного значення набуває форма запису інформації, або, інакше, мова її запису. У КЛ розрізняють: 1) ***формалізацію внутрішню***, або ***формалізацію системи мови***, ***створення моделей мовних об'єктів*** та 2) ***формалізацію зовнішню***, або ***формалізацію моделей мовних об'єктів, створення моделей моделей***, тобто такої форми їхнього представлення, яка була б приступна комп'ютеру та іншим технічним засобам опрацювання інформації. ***Внутрішня формалізація*** полягає у структуруванні інформації, вираженої різними одиницями мови або їхніми сукупностями, виділенні визначальних ознак її будови та правил функціонування. Результатом цього різновиду формалізації мови і стають моделі певних мовних об'єктів або процесів, що з ними відбуваються. ***Зовнішня формалізація*** становить завдання прикладної і, зокрема, комп'ютерної лінгвістики, оскільки зорієнтована на представлення комп'ютеру чи іншому технічному засобу опрацювання інформації уже таких мовних моделей, побудованих за

допомогою формально-логічних або математичних методів. Її результат становлять такі моделі мовних об'єктів, які придатні для використання комп'ютера або іншого технічного засобу. Здійснюючи зовнішню формалізацію мовної інформації слід керуватися мудрими настановами О.Ф.Лосева, висловленими у його праці «Вступ до загальної теорії мовних моделей»: «Традиційне мовознавство, обтяжене накопиченими впродовж десятиліть величезними матеріалами, безперечно, потребує уточнення своїх основних категорій і часткового перегляду своїх методів…Однак на кожному кроці необхідно пам'ятати і те, що лінгвістика не є математика, а математика не є лінгвістика. Змішування цих галузей веде до нових бід науки і навіть доводить до повного розриву традиційну лінгвістику з математичною, до повної неможливості для лінгвіста скористатися з великих досягнень математичних методів»[11]. І як висновок з таких настанов – заклик О.Ф.Лосева до всіх, хто використовує різноманітні моделі мовних об'єктів, побудовані за допомогою формально-логічних і математичних методів: «Модель становить тільки форму вираження, яка може здобути свою повну наукову значущість лише з урахуванням того, про форму чого йдеться… Модель не повинна бути відірваною від мовного змісту, а, навпаки, повинна вказувати шлях для вивчення цього останнього»[12].

Інформація, записана у певній формі, перетворюється на ***дані*** або ***знання***. В основі такого формалізованого запису інформації завжди лежить певна ***інформаційна модель*** тієї предметної галузі, якої стосується така інформація. ***Предметну галузь*** становить тематично однорідна ділянка дійсності – сфера природи, виробництва, громадського або приватного життя людини. Ступінь деталізації структурування предметної галузі залежить від конкретного завдання, яке ставить перед собою дослідник.

Фахівці з інформатики, визначаючи предмет дослідження цієї наукової галузі як "технологію побудови, аналізу та використання людино-комп'ютерного (програмного) знання"[13], основне її завдання і вбачають у розробленні засобів та методів побудови, аналізу та узагальнення інформаційних моделей предметних галузей. ***Інформаційна модель*** є структурованим і представленим у формалізованому вигляді описом окремого об'єкта або й предметної галузі в цілому, який дозволяє одержати інформацію про них за допомогою комп'ютера. Лінгвістичні моделі становлять різновид інформаційних моделей. Їхня відмінність від інших моделей полягає в універсальному характері мови як спеціалі-

[11] **Лосев А.Ф.** Введение в общую теорию языковых моделей. – М.: Едиториал УРСС, 2004. – Изд. 2-е, стереотип. – С. 11.
[12] Там же. – С.15.ы
[13] **Козачков Л.С.** Прикладная логика информатики. – К., 1990. – С.6.

зованої системи обміну інформацією, що обслуговує будь-який різновид розумової діяльності людини. Пригадаємо, що всі інші знакові системи, поширені в людському суспільстві, так чи інакше ґрунтуються на системі природної мови (пор. системи математичних або хімічних символів, знаки регулювання дорожнього руху або так звану "мову" морських прапорців чи семафорну морську абетку та абетку Морзе).

Зауважимо принагідно, що будь-який мовний продукт уже становить певну інформаційну модель того об'єкта дійсності, який він називає або описує. Позначаючи особу словами ***мудрець, молодик, дівчинка, бородань, мовознавець, українець, киянин*** або ***ліберал, неформал***, ми використовуємо різні інформаційні моделі найменування особи, вироблені в лексиконі сучасної української мови. Кожна з цих лексем своєю змістовою структурою вказує на різні ознаки особи: її зовнішність, рівень розумового розвитку, стать, вік, фах, національну належність, місце проживання, політичні чи естетичні уподобання, а їхня форма (письмова й звукова) засвідчує спосіб представлення в українській мові такої інформації про особу. Отже, будь-яка формалізація мовного матеріалу полягає у витворенні уже вторинних інформаційних моделей, а саме: формалізації для комп'ютера інформації, уже сформалізованої засобами природної мови для людини. Недарма, пригадаймо, людську мову й називають вторинною сигнальною системою на відміну від органів чуття як первинної сигнальної системи.

Один об'єкт або одна предметна галузь можуть мати кілька інформаційних моделей залежно від того, яку інформацію про них ми моделюємо. Так, уже подані вище слова ***мудрець, молодик, дівчинка, бородань, мовознавець, українець, ліберал*** та ***неформал*** матимуть різні інформаційні моделі, в яких буде сформалізовано інформацію про їхню фонемну, складову, акцентну, графемну, морфемну, словотвірну та семантичну будову. Причому за одними ознаками ці слова будуть мати спільні інформаційні моделі, а за іншими – різні. Наприклад, слова ***бородань*** і ***українець*** матимуть спільні морфемну та словотвірну моделі, пор., відповідно, **RSF** (корінь+суфікс+флексія), **TO+CI$_{но}$** (твірна основа+суфікс іменника-назви особи). Проте, як можемо пересвідчитися, спільність моделей їхнього опису виявляється на певному рівні узагальнення властивих їм характеристик. Уже введення до ознак створюваної словотвірної моделі ознаки "належність твірної одиниці до певної ономасіологічної категорії і, відповідно, категоріально-розрядне значення її твірної основи" виявляє відмінність між цими іменниками, а отже, й необхідність на такому рівні їхнього опису вироблення інших інформаційних моделей, побудованих на більш детальних ознаках їхнього змісту. Так, іменники ***бородань*** і подібні до нього ***вусань, лобур, носач*** описуватимуть уточнені словотвірні моделі на зразок "твірна основа зі значенням "частина тіла людини"+суфікс іменника-назви

особи", а іменники ***українець*** і ***англієць, росіянин, француз, індус*** – моделі "твірна основа зі значенням "країна, держава" +суфікс іменника-назви особи". Оскільки похідні слова мають прозору, легко зрозумілу внутрішню форму, то такі словотвірні моделі можуть водночас виступати і як різновиди моделей семантичних.

Основне завдання при створенні інформаційної моделі якраз і полягає у відшуканні таких формальних ознак інформації про об'єкт, які є визначальними для "розпізнавання" його комп'ютером за певним типом інформації, встановлення рівня узагальнення опису такого мовного об'єкта-прототипа. Такі вихідні ознаки становлять параметри, або одиниці побудови інформаційної моделі. Сама ж побудова інформаційної моделі передбачає виявлення залежностей між цими ознаками, схем їхнього комбінування, взаємодії в об'єктах модельованої предметної галузі і, відповідно, розташування таких одиниць побудови у створеній моделі за певними правилами. Підсумком побудови моделі є вироблення її формального представлення, тобто обрання певної форми (графічної, символьної, цифрової) для її унаочнення. Як бачимо, і процес побудови інформаційної моделі передбачає виконання процедур внутрішньої та зовнішньої формалізації. У свою чергу, правила побудови інформаційної моделі служать правилами її застосування до опису й аналізу об'єктів, які вона моделює. В узагальненому вигляді етапи створення інформаційної, зокрема лінгвістичної, моделі предметної галузі можна представити в такий спосіб:

1. Виділення формальних ознак, визначальних для окремого типу інформації про модельований об'єкт.
2. Встановлення функціонального навантаження формальних ознак в об'єкті, відношень між ними.
3. З'ясування схем взаємодії або взаємовиключення формальних ознак, їхнього комбінування
4. Вироблення правил дії побудованої інформаційної моделі для одержання нової інформації про будову об'єкта або його властивості.

Вибір способу побудови лінгвістичної моделі, її унаочнення й графічного представлення, інтерпретація результатів її застосування залежать від характеру модельованого мовного матеріалу, способів його аналізу та опису, а також від характеру поставленого дослідником завдання. Коректність побудови й застосування моделі – запорука якісно виконаного лінгвістичного дослідження та вірогідності й пояснювальної сили його результатів, їхнього значення для опрацювання тієї чи іншої проблеми. Поняття "забезпечення роботи комп'ютера" передбачає на сьогодні створення трьох його необхідних складників, трьох "втілень" такого забезпечення: 1) ***інструментального,*** або ***апаратного***, так званого ***hardware*** (від англ. **hardware** буквально "важке оснащення, важкий одяг"), до складу якого входять необхідні для виконання певних

класів завдань технічні пристрої або засоби; 2) ***алгоритмічного*** та ***програмного***, так званого ***software*** (від англ. **software** буквально "м'яке оснащення, м'який одяг"), яке об'єднує процедури автоматизованого логічного опрацювання інформації, та 3) ***лінгвістичного***, або так званого ***lingware***, яке охоплює бази даних або знань та спеціальні лінгвістичні процесори (засоби лінгвістичного аналізу) для опрацювання інформації природною мовою. Чимдалі більше на наших очах для підвищення рівня розв'язуваних з допомогою комп'ютера завдань зростає значення саме останнього компонента забезпечення комп'ютера.

🕮 Терміни

- **прикладна лінгвістика (applied linguistics)** – розділ лінгвістики, предметом дослідження якого є 1) застосування методів інших наукових дисциплін для вивчення будови мовної системи та закономірностей її реалізації та 2) застосування здобутків лінгвістики у розв'язанні завдань інших галузей науки та суспільної практики
- **комп'ютерна лінгвістика (computational linguistics)** – лінгвістична дисципліна, яка розв'язує теоретичні й прикладні завдання мовознавства за допомогою комп'ютера
- **інформатика** – наука про типи інформації, способи, засоби її створення, процеси обміну інформацією у природі, техніці й суспільстві
  - **інформація** – послідовність символів, які є образами певних об'єктів та явищ
  - **предметна галузь** – певна тематично однорідна ділянка природи, суспільного життя, громадського або приватного життя людини
  - **інформаційна модель** – інформація про окремий об'єкт або предметну галузь в цілому, представлена у певному структурованому й формалізованому вигляді, придатному, зокрема, для опрацювання комп'ютером
  - **дані –** інформація про об'єкти, їхні властивості або дії, представлена у структурованому та формалізованому вигляді **знання –** дані про об'єкти, їхні властивості або дії зі вказівкою на умови та правила використання або з їхньою оцінкою
  - **забезпечення роботи комп'ютера –** сукупність технічних пристроїв, засобів, алгоритмічних та програмних процедур, баз даних або знань для виконання поставлених завдань
    - **інструментальне (=апаратне, hardware) –** сукупність технічних пристроїв та засобів
    - **алгоритмічне та програмне (=software) –** сукупність алгоритмічних та програмних процедур

- **лінгвістичне (=lingware) –** сукупність баз даних або знань та спеціальних лінгвістичних процесорів для опрацювання інформації природною мовою

§2. Інформаційні моделі лінгвістичних об'єктів

- Модель та її оригінал – натурний об'єкт (=прообраз, прототип)
- Правила побудови та функції моделей
- Типи лінгвістичних моделей
- Комп'ютерне моделювання та конструювання нових лінгвістичних об'єктів

Модель (від лат. **modulus** зменш. від **modus** "міра, спосіб") – конструкція, зразок будови певного об'єкта або ділянки дійсності, у лінгвістиці – зразок будови одиниць мови або процесів, що відбуваються за їхньою участю. Необхідність у створенні моделі, або аналога-заступника якогось лінгвістичного об'єкта (=натурного об'єкта, оригінала, прототипу моделі) виникає тоді, коли безпосередньому споглядан-ню дослідника недоступні внутрішні властивості такого об'єкта, його будова або ж умови функціонування цього об'єкта настільки багато-гранні і складні, що обмеженість людської пам'яті та сприйняття не дозволяють швидко й логічно послідовно опрацювати потрібний обсяг інформації про такий об'єкт. На допомогу досліднику в таких випадках приходить комп'ютер з його значно більшим за людський обсягом опе-ративної пам'яті, потужнішою швидкодією, тобто швидкістю здійснення певних операцій опрацювання інформації, здатністю одночасно за чи-сленними ознаками обробляти величезні масиви інформації, при цьо-му, на відміну від людини, не втомлюючись, не відволікаючись, не за-буваючи поставлене завдання і нічого в ньому не змінюючи під час його виконання. На хід виконання завдання комп'ютером вплинути мо-жуть лише суто технічні причини: відмова якогось пристрою або зник-нення струму в мережі, що зайвий раз нагадує користувачам про те, що вони мають справу з технічним засобом опрацювання інформації.

Для того, щоб модель дійсно служила аналогом натурного об'єкта, вона повинна відповідати певним вимогам до її побудови. Р.Г.Піотров-ський – автор теорії лінгвістичного автомата як універсальної комп'ю-терної моделі оброблення текстової інформації – сформулював їх так:

1. Бути спрощеним аналогом, але не копією оригіналу.
2. Не бути складнішою за оригінал. Водночас застосування моделі дозволяє одержати потрібну інформацію про оригінал швидше за інші прийоми його дослідження.
3. Побудова моделі має бути вільною від суперечностей (логічно ко-ректною), вичерпною і гранично простою. Проте між цими трьома критеріями існує пріоритетна залежність: критерій коректності важ-

ливіший за критерії вичерпності та простоти, а критерій вичерпності істотніший за критерій простоти. Якщо існує кілька моделей з однаковими показниками всіх трьох перелічених критеріїв, то перевагу віддають тій з них, яка швидше діє, тобто дозволяє дослідникові за коротший час одержати потрібну інформацію про оригінал.

4. Модель повинна мати універсальний характер, що дає змогу застосовувати її не до якогось конкретного об'єкта, а до певного класу натурних об'єктів.
5. Модель повинна мати пояснювальну силу, тобто здатність передбачувати, виявляти й пояснювати ще не реалізовані властивості оригіналу.
6. Модель мусить мати і евристичні (пошукові) властивості, тобто генерувати нові знання про оригінал[14].

Таблиця 4. Типи лінгвістичних моделей

Об'єкт моделювання	Тип моделі
Структурні властивості оригінала, його будова	Статична, структурна, класифікаторна, таксономічна
Функціонування оригінала	Динамічна, функціональна, процесуальна
Система мови	Мовна
Мовленнєва діяльність	Мовленнєва
Аналіз оригінала	Індуктивна (=аналітична)
Синтез оригінала	Дедуктивна (=синтетична (породжувальна, генерувальна)
Нове знання про оригінал	Гіпотетична (=евристична, пошукова)
Нові об'єкти із заданими властивостями	Відтворювальна (=конструювальна, імітаційна)

Лінгвістичні моделі різняться як за об'єктом моделювання, так і за способом їхнього застосування. Див. узагальнення уже створених типів лінгвістичних моделей у табл. 4:

У лінгвістиці потреба в моделюванні мовних об'єктів виникла разом з необхідністю їхнього опису та унаочнення для запису мовних повідомлень, обміну ними або для їхнього зберігання, а також з метою навчання мови. Отже, використовуючи певне письмо (графемний запис звукових повідомлень) або навчаючись мови з допомогою певних граматик, ми працюємо з лінгвістичними моделями, розрахованими на людське сприйняття мовної інформації. Такі, некомп'ютерні, моделі відповідають усім вимогам до моделей, створюваним для оброблення мовної інформації комп'ютером. Переконаємося в цьому на прикладі моделей системи приголосних у сучасній українській мові. Причому свідомо відберемо моделі, різні за способом побудови або графічного

[14] **Пиотровский Р.Г.** Лингвистический автомат и его речемыслительное обоснование. – Минск, 1999. – С.25-26.

зображення. Традиційну, загальноприйняту і всім нам знайому ще зі шкільної лави модель подає в табличному вигляді (див. табл. 5) Г.П.Півторак у своїй статті "Приголосні звуки" в енциклопедії "Українська мова" (К., 2000, с. 485). 45 приголосних звуків, реалізованих у словах сучасної української мови, описано за рядом основних і додаткових ознак. Основні ознаки: місце, спосіб творення та наявність голосу; як додаткові ознаки виступають наявність шуму та твердість/м'якість. Отже, кожен з приголосних звуків у такій моделі постає як пучок перетину основних та додаткових ознак.

Таблиця 5. Класифікація приголосних звуків сучасної української мови (за моделлю Г.П.Півторака)

Ознака голосу	Ознака способу творення	Ознака шуму	Ознака місця творення									
			губні		язикові						гортанні	
					передньо язикові		середньо язикові		задньо язикові			
			тв.	м.	тв.	м.	тв.	м.	тв.	м.	тв.	м.
шумні	зімкнені	дзвінкі	**б**	**б'**	**д**		**д'**		**ґ**	**ґ'**		
		глухі	**п**	**п'**	**т**		**т'**		**к**	**к'**		
	щілинні	дзвінкі			**ж, з**	**ж',з'**					**г**	**г'**
		глухі	**ф**	**ф'**	**ш, с**	**ш',с'**			**х**	**х'**		
	зімкнено щілинні	дзвінкі										
		глухі										
	африкати	дзвінкі			**дж**	**дж'**						
		глухі			**дз**	**дз'**						
					ч, ц	**ч',ц'**						
сонорні	зімкнені											
	щілинні		**в**	**в'**	**л**			**л', j**				
	зімкнено-щілинні		**м**	**м'**	**н, р**	**р'**		**н'**				

Запропонована модель узагальнення інформації пояснює, за якими формальними ознаками протиставляються один одному або об'єднуються приголосні звуки сучасної української мови, а разом з тим засвідчує нереалізовані можливості української фонетичної системи. Гіпотетичні (передбачувальні) властивості цієї моделі дають змогу встановити обмеження на реалізацію тих чи інших звуків, що їх накладає система сучасної української мови. Нереалізовані ресурси підсистеми приголосних звуків у поданому графічному варіанті моделі засвідчили пусті клітинки. Реальних приголосних звуків – 45, нереалізованих, але закладених в потенціалі цієї підсистеми – 75. Отже, потенціал підсистеми приголосних реалізовано у сучасній українській мові на 60%. Так, наприклад, передньоязикові щілинні приголосні реалізують опозицію і за дзвінкістю/глухістю, і за твердістю/м'якістю на відміну від губних щілинних, у яких опозиція за дзвінкістю/глухістю залишилася нереалізованою, пор.: **ж:ж'** та **ж:ш** з **ф:ф'.** Свого часу саме таблична модель упорядкування голосних звуків у межах коренів індоєвропейсь-

кої прамови дала можливість Ф.де Сосюру виявити зниклі так звані "сонантні коефіцієнти", тобто звуки, які після голосних могли виступати як нескладотвірні сонанти, а в редукованому корені, тобто в корені, що втратив голосний, ставали складотвірними елементами[15]. Згодом, у 1906-1907 рр., після відкриття неситських документів з м.Богазкьой (Туреччина) і їхнього розшифрування такі фонеми було виявлено в хетській мові, якою були писані ці пам'ятки, – одній з найдавніших індоєвропейських мов, що мала писемність[16].

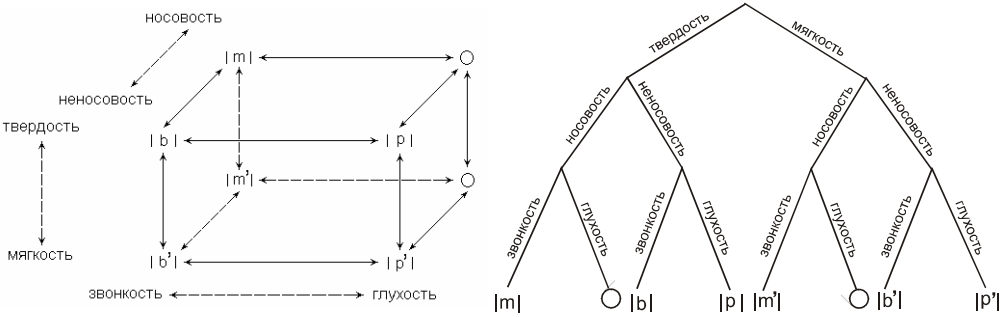


Рис. 1. Кубічна (стереометрична) та деревна моделі російських губних приголосних (за Р.Г.Піотровським)

Можливі і інші графічні способи зображення моделей організації приголосних. Наприклад, Р.Г.Піотровський запропонував для системи приголосних сучасної російської мови стереометричний і "деревний" способи зображення таких мовних моделей. Згідно з першим кожна з ознак зіставлення звуків задає координати певної площини об'ємного зображення моделі. Використовуючи стереометричний принцип зображення, за трьома ознаками носовість/неносовість (в інших термінах, сонорність/несонорність (шумність)), дзвінкість/глухість, твердість/м'якість можна побудувати кубічну модель для російських приголосних **б, б', п, п', м, м'** (див. рис. 1). За такими ж класифікаційними ознаками можна змоделювати взаємовідношення і відповідних українських приголосних (пор. вище табл. 5):

У моделі з деревним способом зображення кожна з ознак зіставлення приголосних формує вузол графа-аналога описуваної підсистеми приголосних, а конкретна опозиція, що формується на підставі такої ознаки, задає його гілки. Пусті кружки в поданих графічних зображен-

[15] Зацікавлених відсилаємо до праці **Ф. де Сосюра** "Мемуар о первоначальной системе гласных в индоевропейском языке // Соссюр Ф. де. Труды по языкознанию - М. 1977. – С.423-425.

[16] **Пиотровский Р.Г.** Зазнач. праця. – С.30-31.

нях моделей також засвідчують нереалізовані можливості підсистеми приголосних сучасних російської та української мов.

Усі подані моделі розраховані на людське сприйняття звукової інформації. Її комп'ютерне моделювання повинно бути принципово іншим, оскільки комп'ютер при "розпізнаванні" мовних одиниць спирається лише на їхню форму (при графічному зображенні інформації) або на їхні суто фізичні, акустичні, характеристики при сприйнятті інформації в усній формі. У розв'язанні цього завдання можливі два шляхи:

1. Комп'ютер забезпечують аналогом тієї чи іншої літери (звуку) з вичерпною інформацією про її сполучувальні та позиційні властивості, що в свою чергу, реалізують її змісторозрізнювальні функції.
2. Для комп'ютера створюють алгоритм встановлення тієї чи іншої літери (звуку) на основі обстеження її сполучувальних та позиційних властивостей, її функціонування в письмовому чи усному тексті.

Перший шлях дозволяє досягти поставленої мети за допомогою декларативних знань про модельований об'єкт, другий – за допомогою процедурного знання про нього.

За способом побудови виділяють індуктивні та дедуктивні моделі. Перші йдуть від аналізу конкретного мовного матеріалу до формування певної гіпотези про закономірності його організації та функціонування, другі, навпаки, будуються на основі заданої гіпотези і перевіряють її вірогідність на реальному мовному матеріалі. До індуктивних моделей належать, наприклад, ***дистрибутивна модель***, яка визначає класи мовних одиниць на основі аналізу їхнього розподілу в тексті; ***трансформаційна модель***, що встановлює змістову тотожність/нетотожність мовних одиниць завдяки аналізу їхніх можливих перетворень (трансформацій); ***статистико-комбінаторна модель М.Д.Андреєва***, з допомогою якої встановлюють морфологічні класи слів на основі аналізу частоти їхніх кінцевих буквосполук, а також ***ймовірнісні моделі***, наприклад, ***модель побудови семантичного поля прикметників А.Я.Шайкевича***[17]. Дедуктивні моделі становлять інструмент побудови одних мовних об'єктів з інших, наприклад, одиниць вищих рівнів мовної системи з одиниць її нижчих рівнів. Вони складаються з певних заданих дослідником вихідних (мінімальних) одиниць та правил їхнього перетворення. При побудові таких моделей дослідники намагаються дотримуватися правил виведення (синтезу) одиниць, за своєю чіткістю наближених до математичного числення. Крім того, дедуктивні лінгвіс-

[17] Див. докладніше про ці моделі у ст. **В.С.Перебийніс** "Модель" в енциклопедії "Українська мова" (К., 2000, С. 339-340), а також: **Андреев Н.Д.** Статистико-комбинаторные методы в теоретическом и прикладном языковедении. – Ленинград, 1967 та **Шайкевич А.Я.** Дистрибутивно-статистический анализ текстов. – Ленинград, 1982.

тичні моделі наближені до високоформалізованих логіко-математичних моделей за рядом вимог до визначення як вихідних їхніх одиниць, так і правил їхнього синтезу, а саме:

1. **Простота**: мінімальна кількість вихідних одиниць та правил їхнього синтезу.
2. **Несуперечність**: правила такої моделі не повинні суперечити одне одному.
3. **Вичерпність, або ефективність**: породжувати якомога більше одиниць, що мають аналоги в мові, а отже, відповідати нормам реальної мовної системи, мати якомога нижчий ступінь "шуму" – породження одиниць, не відповідних реальним словам мови.
4. **Однорідність**: правила повинні ґрунтуватися на однакових критеріях (формальних, семантичних, функціональних, кількісних)[18].

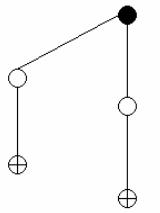
Прикладом ефективної дедуктивної лінгвістичної моделі може служити ***аплікативна породжувальна модель С.К.Шаумяна***[19]. Ця модель має два генератори (породжувачі) мовних одиниць: ***генератор слів***, що моделює процеси морфологічного словотворення, та ***генератор класів слів***, що моделює процеси породження синтаксичних структур і речень. В основі моделі лежать мінімальні одиниці (класи непохідних основ слів або ядерних (далі неподільних) конструкцій), між якими згідно з правилами моделі існують два типи відношень – приєднання чи заміщення. Словотвірний формант, з допомогою якого здійснюється творення нового слова або нової конструкції, в моделі С.К.Шаумяна називається ***релятор***. Аплікативна породжувальна модель має символьний та графічний запис. У символьному записі **0** позначає вихідну одиницю породження (основу або корінь слова), а **R** – певний релятор (слово- або формотворчий засіб: префікс, суфікс, флексію в ролі суфікса). Релятор вказує на ті перетворення, які відбулися з аморфною, неоформленою і несамостійною в мові вихідною одиницею породження. Він позначає ту функцію, яку виконує така одиниця у висловленні. Отже, аморфна вихідна одиниця виконує роль аргумента функції (її об'єкта), а релятор виражає певну функцію такого об'єкта. Процес генерування і полягає у приєднанні певних реляторів до вихідних одиниць породження на окремих тактах роботи того чи іншого генератора. Звідси й назва такої породжувальної моделі – аплікативна, або приєднувальна (від лат. **applico** "приєднувати, прикладати").

[18] Див. **Перебийніс В.С.** Зазнач. праця. – С.340.

[19] Читачів, зацікавлених докладніше познайомитися з принципами побудови та застосування цієї моделі, відсилаємо до праці: **Шаумян С.К., Соболева П.А.** Основания порождающей грамматики русского языка: Введение в генотипические структуры. - М., 1968.

Н.Ф.Клименко застосувала цю модель у дослідженні системи афіксального словотворення сучасної української мови[20]. Об'єктом моделювання вона обрала слова чотирьох повнозначних частин мови: іменники, дієслова, прикметники та прислівники. Мінімальними одиницями в її варіанті аплікативної породжувальної моделі служили корені таких слів (**0**), реляторами були формальні компоненти (флексії чи афікси), з допомогою яких утворюються зазначені слова, відповідно, $\mathbf{R_1}$ – афікс дієслова (причому, для вивчення розмаїття словотвірних структур дієслівних форм до уваги бралися змінні дієслівні форми), $\mathbf{R_2}$ – афікс іменника, $\mathbf{R_3}$ – афікс прикметника і $\mathbf{R_4}$ – афікс прислівника. За правилами моделі на кожному такті, або кроці її роботи до вихідної одиниці або комбінації одиниць додавався лише один релятор. Так, на першому кроці відбувається оформлення кореня у вихідне слово – до нього додається флексія. Для іменника це будуть одиниці із символьним записом $\mathbf{R_20}$ (пор. з реальними словами ***дум-(а), мор(е), сад-0***). За правилами запису математичних перетворень релятор – показник функції пишеться зліва від її аргумента, а отже, запис $\mathbf{R_20}$ слід читати так: над вихідним аморфним об'єктом **0** (у поданих словах – коренем або твірною основою) здійснено функцію перетворення її на іменник, що й позначає релятор $\mathbf{R_2}$. Другий такт роботи моделі дає вже дві структури $\mathbf{R^2_20}$ та $\mathbf{R_2R_10}$ (пор. з реальними словами ***дум-оньк-(а), сад-ок-0*** або ***верт-ун-0, сиді-нн-(я)***).

Творення іменників на 2-му такті роботи моделі можна зобразити як



Для графічного зображення кожного з частиномовних класів породжуваних структур визначено напрямок відповідної гілки графа: | вказує на напрямок розгортання графа, який моделює творення іменників; / – дієслів; \ – прикметників та _ прислівників.

Спираючись на правила дії такої моделі, можна окреслити в цілому процес морфологічного творення слів певної частини мови, дослідити реалізовані й нереалізовані ресурси мовної системи, виявивши при цьому внутрішньо- і міжчастиномовні переходи, можливості кількісного розгортання структури слів, комбінації реляторів. Так, наприклад, виявилося, що слова з дієслівними коренями можуть мати більшу кіль-

[20] Див. докладніше про теоретичне обґрунтування, хід виконання і одержані результати цього дослідження у праці: **Клименко Н.Ф.** Система афіксального словотворення сучасної української мови. - К., 1973.

кість реляторів, ніж слова з іменними коренями. Найдовший (11-морфемний) іменник сучасної української мови ***не-о-по-да-т-к-ов-ува-н-ість-0*** має саме дієслівний корінь. Дієслівний корінь має також іменник найвищого в словотвірній системі української мови – 8-го ступеня похідності – ***неплатоспроможність***, пор. ***могти́ → про-могти́ → промогти́-ся → с-промогти́ся → спромо́ж-н(ий) → плат-о-спромо́жний → не-платоспромо́жний → неплатоспромо́жн-ість***.

Нереалізовані можливості в конструюванні морфемної структури слів сучасної української мови дала можливість виявити породжувальна модель, яка ґрунтується на принципі аплікації (додавання на кожному кроці роботи моделі лише одного складника), а для свого графічного зображення використовує принцип сітки – графа з гранями, орієнтованими за напрямком розгортання вихідної (мінімальної) структури слова. Моделювання будови слів з допомогою таких сіток-орієнтованих графів запропонував на початку 50-х років минулого століття німецький учений П.Менцерат для унаочнення процесів можливого ускладнення фонемної та графемної структури слів сучасної німецької мови[21]. Морфемна сітка становить модель механізму творення морфемних структур простих (=з одним коренем) слів окремих частин мови, представлену як послідовне розгортання на лінійній осі певної мінімальної за довжиною, вихідної морфемної структури слова вліво (збільшення її докореневої, або префіксальної, частини) та вправо (збільшення її післякореневої, або суфіксальної, частини). На першому етапі побудови сітки від вихідної одиниці на гранях графа на відповідних тактах роботи моделі позначаються реальні морфемні структури слів аналізованої частини мови, тобто такі, які засвідчені в реальних словах мови. На другому етапі всі грані графа з'єднуються між собою у замкнену сітку. Нереалізовані, гіпотетичні структури, які не відповідають конкретним словам мови, у такій сітці позначено знаком ***** і подано в сірих комірках (див. рис. 2). Так, виявилося, що мінімальною для іменників української мови є структура **R**, властива незмінюваним запозиченим словам на зразок ***бра, табло, кашне***. За рахунок суфіксів та флексій структури іменників можуть розгортатися до 6-ого такту ро-

[21] Зацікавлені читачі зможуть докладніше познайомитися з принципом побудови таких моделей та з результатами досліджень, виконаних за їхньою допомогою на матеріалі різних мов, у працях: **Menzerath P.** Die Architektonik des deutschen Wortschatzes // Phonetische Studien. - 1954. - № 3; **Клименко Н.Ф., Карпіловська Є.А.** Морфемні структури слів у сучасній українській літературній мові // Мовознавство. – 1991. - № 4. – С.10-21; **Карпіловська Є.А.** Морфемна сітка як інструмент дослідження будови слова // Українське мовознавство. – 1992. – вип.19. – С.100-110; **Клименко Н.Ф., Карпіловська Є.А.** Морфеміка слов'янських мов як об'єкт типологічного вивчення // Мовознавство. – 1998. - № 2-3. – С.117-135.

боти моделі (пор. ***пис-а-н-к-ар-к(а), сл-а-н-ц-ювa-нн(я))***, за рахунок префіксів – до 3-го такту її роботи (пор. ***спів-до-по-відь-0, мета-с-полук(а)).*** Кількісні показники коло реалізованих структур засвідчують їхню пояснювальну силу (кількість описуваних ними слів), а також одночасно й породжувальну потужність таких моделей морфемної будови слова. Відношення реалізованих моделей будови слів до загальної їхньої кількості у сітці дозволяє обчислити міру реалізації словопороджувального механізму іменників у сучасній українській мові (23:32). Вона досить висока і становить 71,8%. Інформація про морфемну будову слова, узагальнена в описаній моделі, дає змогу визначити й оптимальну (найуживанішу) кількість складників у морфемній структурі простого українського слова. Як засвідчує подана нижче морфемна сітка, майже 56% всіх проаналізованих простих іменників (17875 слів з 32075 проаналізованих) містять у своєму складі 3-4 морфеми. Взагалі ж, як підтвердив аналіз за допомогою морфемних сіток будови, крім простих іменників, також дієслів, прикметників та прислівників, оптимальними для простих слів є структури саме з 3-5 морфем.

*3PR	*2PR	*PR	R (280)
3PRF (2)	2PRF (89)	PRF (1650)	RF(5732)
3PRSF (10)	2PRSF (155)	PRSF (2807)	RSF (7548)
3PR2SF (7)	2PR2SF (440)	R2SF (4532)	R2SF (5681)
3PR3SF (3)	2PR3SF (119)	PR3SF (947)	R3SF (1693)
*3PR4SF	2PR4SF (6)	PR4SF (28)	R4SF (332)
*3PR5SF	*2PR5SF	PR5SF (3)	R5SF (10)
3PR6SF (1)	*2PR6SF	*PR6SF	*R6SF

Рис. 2. Морфемна сітка простих іменників української мови

Породжувальні моделі поклали початок новому напрямку у сучасних лінгвістичних дослідженнях – конструюванню нових лінгвістичних об'єктів, тобто таких об'єктів, які насамперед уможливлюють одержання якісно нової інформації про будову та функціонування мовної системи, виявляють її нереалізований потенціал і регулятори реалізації словопороджувальних можливостей. Докладніше про здобутки лінгвістичного конструювання ми поговоримо далі, у §5 цього розділу, присвяченому сучасній комп'ютерній лексикографії. Тут, перш ніж перейти до розгляду комп'ютерних породжувальних моделей, ми обмежимося

лише поданням визначення нового лінгвістичного об'єкта, яке запропонував Ю.М.Караулов. "Створити, побудувати якусь "річ", – пише він –, означає не тільки вміти пояснити ті властивості мови, які в ній використано і на яких вона ґрунтується, не тільки пояснити певні закономірності мовної структури, а й з'ясувати нові властивості побудованого об'єкта, що так чи інакше характеризують досліджувану мову, а отже – розширити наші знання про людську мову взагалі. Таким чином, новим лінгвістичним об'єктом будемо називати таке представлення фактів, мовних даних, яке генерує нову інформацію про мову"[22]. І далі: "Головний принцип лінгвістичного конструювання – "як зробити" той чи інший об'єкт, – лише на перший погляд може уявлятися суто технічним за своїм змістом. Насправді такий підхід містить багато нових можливостей, нових виходів на фундаментальні проблеми науки про мову"[23]. Прикладом такої моделі конструювання нового лінгвістичного об'єкта може служити розроблена нами і реалізована на комп'ютері модель морфемного породження простих дієслів[24]. Як відомо, дієслівне словотворення в сучасній українській мові має розгалужену систему внутрішньочастиномовних переходів, внаслідок чого структури слів ніби виводяться одна з одної за певними правилами розширення вихідної, мінімальної за кількістю складників, структури. Об'єктом аналізу для відбору елементів моделі та формулювання правил її роботи послужили дієслова тематичної групи звучання, вилучені методом суцільної вибірки з 11-томного тлумачного "Словника української мови". Їхні морфемні структури було проаналізовано за такими ознаками: 1) класи морфем у їхньому складі; 2) позиції, які здатні займати у структурі слова афіксальні морфеми, 3) обов'язковість/необов'язковість появи в дієслові певної афіксальної морфеми в певній позиції; 4) розподіл афіксальних морфем за позиціями. У символьному записі моделі стрижнева морфема слова – корінь – вміщувалася в центрі; стрілками вправо та вліво від кореня позначалися переходи, відповідно, на післякореневі, суфіксальні й постфіксальні, та докореневі, префіксальні, афіксальні позиції. Морфеми, здатні вживатися в одній слівній позиції, відокремлювалися одна від одної в записі знаком диз'юнкції **|**, тобто знаком логічної операції "або-або" – вибору з сукупності елементів лише одного можливого**.** Морфеми в позиціях, необов'язкових для побудови структури модельованих дієслів, в записі моделі було взято в квадратні дужки. Якщо морфема в структурах могла вживатися у кількох позиціях, то в записі моделі вона подавалася в максимально віддаленій від ко-

[22] **Караулов Ю.Н.** Лингвистическое конструирование и тезаурус литературного языка. – М., 1981. – С.16.
[23] **Караулов Ю.Н.** Зазнач. праця. – С.17.
[24] Зацікавлених читачів відсилаємо до праці: **Карпіловська Є.А.** Конструювання складних словотворчих одиниць. – К., 1990.

реня позиції. Наприклад, в дієсловах з коренем **хлюп-** морфема **-а-** могла виступати в 1-ій (***хлюп-а-ти***), 2-ій (***хлюп-от-а-ти***) та 3-ій (***хлюп-ос-т-а-ти***) післякореневій позиції. У записі моделі для породження всіх перелічених дієслівних структур морфему **-а-** вміщено у 3-ій післякореневій позиції і в такий спосіб уможливлено вклинювання між коренем і цією морфемою інших суфіксальних морфем: **-ос-** і **-т-**. Наприклад, на підставі аналізу морфемного складу 23 дієслів з коренем **дзвон-** була побудована модель їхнього морфемного породження з таким символьним записом:

[по]←[ви|від|до|з|за|об|пере|по|про|роз]←дзвон→и|ува→ти→[ся]

Вихідними для роботи запропонованої породжувальної моделі є структури, що складаються з кореня і морфем в обов'язкових афіксальних позиціях. Так, для поданої вище моделі вихідними виявилися дві структури ***дзвон-и-ти*** та ***дзвон-ува-ти***. На кожному такті роботи моделі до структури, згенерованої на попередньому такті, додається лише одна морфема з тих, що засвідчені в необов'язкових позиціях. Зручність запропонованого символьного запису полягає в тому, що за його допомогою, навіть не вдаючися до комп'ютерного перебору можливих комбінацій, дослідник може визначити загальну кількість передбачених моделлю дієслівних структур. Зробити це можна, підрахувавши кількість породжуваних за правилами моделі дієслівних структур у кожній позиції. Так, у 2-ій докореневій позиції засвідчений лише один префікс **по-** в необов'язковій позиції, отже, можливі 2 дієслівні структури – з цією морфемою і без неї. У 1-ій докореневій необов'язковій позиції вміщено 10 префіксів, а отже, можливі 11 структур – з кожним з цих префіксів і без них усіх. 1-а післякоренева позиція обов'язкова, в ній функціонують 2 морфеми і можливі 2 структури – з кожною з них. 2-а післякоренева позиція також обов'язкова і в ній засвідчено 1 морфему **-ти**, а отже, можлива лише 1 структура. 3-я післякоренева позиція містить 1 морфему **-ся** в необов'язковій позиції, відповідно, можливі 2 структури – з цією морфемою і без неї. Перемноживши всі встановлені за позиціями структури, одержимо показник загального потенціалу породження дієслів за цією моделлю: 2*11*2*1*2=88. Отже, СУМ засвідчив у своєму реєстрі лише 26% (23:88) дієслів з коренем **дзвон-**, можливих за цією породжувальною моделлю. Згенеровані комп'ютером 65 відсутніх у СУМі дієслів аналізуємо далі з погляду їхньої припустимості, або відповідності нормам побудови морфемної структури дієслів групи звучання і дієслів сучасної української мови в цілому. На етапі генерування слів комп'ютером було запроваджено єдине обмеження для уникнення явного "шуму" (завідомо неправильно побудованих структур): відкидалися структури із зяянням голосних у післякореневій частині на зразок ***хлюпнуіти*** або ***хлюпуваати***. У докореневій частині цього обмеження немає, оскільки тут діють інші закони сполучення

морфем, ніж у післякореневій частині слів, і зяяння голосних на стиках префікса і кореня припустиме, пор. ***до-опрацювати*** або ***перо-брати***. У докореневій частині дієслів на відміну від їхньої післякореневої частини також спостерігаємо комбінацію однакових за формою морфем, пор. ***по-пошуміти, пере-перейменовувати***. Натомість такі поєднання властиві післякореневим частинам слів інших частин мови, зокрема прикметникам, пор. ***друж-ин-ин, княг-ин-ин*** або формам дієслів, пор. форми пасивних дієприкметників в орудному відмінку однини: ***припуст-им-им, терп-им-им***.

Для встановлення нормативності згенерованих комп'ютером, але не засвідчених СУМом дієслів було створено спеціальну систему фільтрів. Саме ця система і служила засобом перевірки ефективності, пояснювальної і передбачувальної сили побудованої дедуктивної моделі, оскільки вона ґрунтувалася на інформації, вміщеній у структурах реальних слів мови. Систему таких фільтрів, а отже, й процедуру визначення за їхньою допомогою ступеня припустимості генерованих комп'ютером структур можна будувати з різною мірою жорсткості критеріїв. У нашому дослідженні такою мірою жорсткості служили критерії: 1) структура, властива дієсловам аналізованої тематичної групи звучання та 2) структура, властива дієсловам як граматичному класу слів у цілому. Перший фільтр становив реєстр попарних сполук до- і післякореневих морфем між собою. Наприклад, у докореневій частині було виявлено ряд сполук, не засвідчених у дієсловах звучання, а отже, не включених до таблиць попарної сполучуваності таких морфем: **по-від, по-до, по-з, по-про** тощо. Проте, як показує аналіз дієслів інших тематичних груп, такі комбінації префіксів цілком припустимі в сучасному українському дієслівному словотворенні, пор. з їхньою реалізацією у структурах дієслів інших тематичних груп: ***повідвозити, подоварювати, позбирати, попрогравати***. Отже, для тематичної групи звучання структури дієслів на зразок ***повідзвонити, попродзвонювати*** слід визнати штучними, нереальними, проте в межах дієслів як граматичного класу слів такі структури є цілком припустимими, побудованими правильно з погляду і форми, і змісту.

Другий фільтр становить реєстр усіх засвідчених у реальних словах поєднань афіксальних морфем до- і післякореневої частини, афіксальних оточень, або каркасів кореня. Наприклад, в обстежених дієсловах звучання не засвідчено комбінацій афіксальних морфем на зразок **по+ви[#]и+ти+ся, по+ви[#]ува+ти, [#]ува+ти, [#]ува+ти+ся**[25]**.** На цій підставі слід визнати нереальними для дієслів з коренем **дзвон-**

[25] Знак **#** у таких записах вказує на змінну частину моделі морфемної будови дієслова, тобто на можливість функціонування в позиції між наявними афіксальними морфемами кореня або інших афіксальних морфем у до- і післякореневій позиціях.

структури ***повидзвонитися, повидзвонювати, дзвонювати, дзвонюватися.*** Проте знову ж таки для граматичного класу дієслів взагалі вони припустимі, пор. ***повигонитися, повибілювати, регулювати, меблюватися***. Оскільки в тематичній групі дієслів звучання не трапилося жодної лексеми, де б суфікс **-ювa-** функціонував без префікса, то згенеровані комп'ютером структури на зразок ***дзвонювати*** або ***дзвонюватися*** слід визнати лише напівправильними, тобто правильними з погляду форми (пор. вище з ***регулювати, меблюватися***), проте не осмисленими, неправильними з погляду змісту.

Зовсім неприпустимими, неправильними з погляду як форми, так і змісту, словами-"монстрами" виявилися структури на зразок ***дзюркотти, дзижчти***, ***муркотти***, оскільки дієслівні суфікси на позначення інтенсивної або безперервної дії обов'язково поєднані з тематичними голосними **-а-** чи **-і-**, пор. ***дзюрк-от-а-ти, дзюрк-от-і-ти, дзиж-ч-а-ти, мурк-от-і-ти***. Таким чином, встановлення функціональних властивостей складників морфемної структури дієслів, визначення закономірностей дії селективного (відбіркового) механізму сучасної української мови при побудові слів цього граматичного класу дало змогу створити модель для їхнього комп'ютерного синтезу з досить великою гіпотетичною та пояснювальною силою, а також підвищити надійність одержуваних з допомогою моделі даних на основі розроблених процедур їхньої верифікації (перевірки) завдяки системі фільтрів відповідності їхніх форми та семантики нормам граматики сучасної української мови.

У табл.4 "Типи лінгвістичних моделей" подано ще одну опозицію моделей за способом аналізу мовних об'єктів і самим спрямуванням такого аналізу – моделі ***статичні***, або ***структурні, класифікаторні, таксономічні*** та ***динамічні,*** або ***функціональні, процесуальні.*** Саме визначення цих моделей вказує на ознаку, покладену в основу їхнього протиставлення – стан, в якому перебуває модельований мовний об'єкт. Статичні моделі відображають його статику, або стан спокою, рівноваги і спрямовані на унаочнення структури, будови, устрою такого об'єкта, класифікацію його складників за певними їхніми ознаками або властивостями. На цій підставі статичні моделі називають ще структурними, класифікаторними, таксономічними (від грецьк. τάξις "клас, розряд; порядок"). До таких моделей можна віднести, наприклад, подані вище моделі будови підсистем приголосних в українській та російській мовах. Динамічні моделі відображають рух, динаміку мовного об'єкта, ті процеси, які з ним відбуваються, і тому їх ще називають функціональними, процесуальними. Якщо статичні моделі унаочнюють будову якогось окремого об'єкта або їхньої певним чином впорядкованої сукупності, то динамічні моделі унаявнюють процеси виведення одних мовних об'єктів з інших, їхні взаємоперетворення. До таких на-

лежать представлені в цьому параграфі породжувальні моделі. Суттєва відмінність у будові цих моделей полягає в тому, що для інтерпретації статичних моделей цілком достатньо унаочнення в графічному вигляді ознак, за якими вони побудовані. Для використання динамічних моделей обов'язковими є також правила взаємодії таких ознак, їхнього перетворення для одержання аналога об'єкта в іншому стані. Інакше кажучи, для інтерпретації динамічної моделі крім будови її оригінала в початковому, вихідному стані потрібні правила переходу до його будови в усіх наступних, похідних станах у процесі функціонування такого мовного об'єкта-оригінала.

🕮 Терміни

- **модель –** конструкція, структура, формально-логічна побудова, що служить аналогом реального об'єкта (=натурного об'єкта, оригінала, прототипа, прообразу)
  - **дедуктивна модель (= модель синтезу, породжувальна модель) –** модель, створена на основі гіпотетично заданих вихідних елементів та правил синтезу з них інших, похідних, об'єктів
  - **індуктивна модель (= модель аналізу) –** модель, що створена на основі результатів аналізу реального об'єкта
  - **динамічна модель (= функціональна, процесуальна модель)** – модель, яка відображає динаміку оригінала-мовного об'єкта або об'єктів, його функціонування, процеси, що з ним відбуваються
  - **статична модель (= структурна, класифікаторна, таксономічна модель)** – модель, яка відображає будову, устрій оригінала-мовного об'єкта чи об'єктів в статиці, або стані спокою, класифікує його за певними ознаками
  - **оригінал моделі (=натурний об'єкт, прототип, прообраз)** – реальний об'єкт – у лінгвістиці – клас мовних одиниць або текст як організована сукупність таких класів одиниць, для вивчення якого створюють модель;
  - **гіпотетична функція моделі** – здатність моделі служити засобом передбачення нових об'єктів або нових властивостей уже наявних об'єктів
  - **пояснювальна функція моделі** – здатність моделі відображати інформацію про реальні об'єкти та їхні властивості, пояснювати закономірності їхньої будови та функціонування
  - **конструювання лінгвістичних об'єктів** – процес створення за певними дедуктивними моделями нових лінгвістичних об'єктів, що дають змогу одержати якісно нову інформацію про будову та функціонування мовних одиниць та їхніх сукупностей

§3. Бази даних і бази знань (=інтелектуальні бази даних)

- Етапи проектування бази даних
- Словнико- та текстозорієнтовані лінгвістичні бази даних
- Способи організації лінгвістичних баз даних
- Комп'ютерна копія та комп'ютерна версія словника-оригінала
- Граматика комп'ютерної версії традиційного словника
- Стратегія проектування бази даних морфемно-словотвірного фонду української мови як словнико-зорієнтованої бази

Лінгвіст, який у своїй роботі з мовним матеріалом використовує комп'ютер, працює у так званому комп'ютерному середовищі. Змодельовані для комп'ютера мовні об'єкти, декларативні та процедурні знання про них становлять фактографічне підґрунтя для розв'язання певних завдань КЛ. Розрізняють два види таких фактичних основ діяльності лінгвіста-комп'ютерника залежно від типу відомостей про модельовані об'єкти. Інформацію про мовні об'єкти незалежно від умов їхньої реалізації, ситуацій і особливостей використання, їхнього зв'язку з іншими мовними об'єктами прийнято називати ***даними (data)***, а масив (сукупність) такої інформації – ***базою даних (database)***. Відомості про можливості й способи застосування мовних об'єктів у різних ситуаціях спілкування, у різних продуктах мовної діяльності, судження про такі об'єкти, їхню оцінку, а отже, відомості, на основі яких можна робити певні умовиводи про мовні об'єкти, називають ***знаннями (knowledges)***, а сукупність таких відомостей – ***базою знань (knowledge base)***, або ***інтелектуальною базою даних (intelligent database)***. Таким чином, база даних відносно бази знань виступає як вихідний продукт до похідного, оскільки остання містить інформацію про відношення між даними та їхні ціннісні характеристики. Скажімо, ми одержуємо на екрані комп'ютера повідомлення, оформлене англійською мовою. Проте доки ми не вміємо це повідомлення зрозуміти (прочитати й перекласти українською мовою), такі дані не становлять наше знання. Проектування бази даних передбачає два етапи:

1. **Інфологічний** – етап відбору інформації та її структурування, або етап внутрішньої формалізації інформації, етап моделювання змісту інформації.
2. **Датологічний** – етап оформлення інформації відповідною мовою представлення, придатною для комп'ютерного опрацювання, або етап зовнішньої формалізації інформації, етап моделювання її форми, перетворення інформації на дані

Завдання інфологічного етапу проектування бази даних полягає у відборі об'єктів опису, типів інформації про їхню будову та функціону-

вання. Це етап вивчення та опису певної предметної галузі, її внутрішньої формалізації. Результатом роботи лінгвіста на цьому етапі є концептуальна інформаційна модель такої предметної галузі. Завданням другого, датологічного, етапу є вироблення способів представлення об'єктів та інформації про них у пам'яті комп'ютера, спеціальних маркерів-сигналізаторів для безпомилкового "розпізнавання" комп'ютером того чи іншого типу інформації, правил взаємодії типів інформації та одержання з бази даних відомостей у потрібному вигляді або обсязі, тобто це етап зовнішньої формалізації інформації про мовні об'єкти. Датологічний етап, у свою чергу, містить дві стадії організації даних: ***логічну*** – пов'язання концептуальної інформаційної моделі, одержаної на інфологічному етапі, з операційною системою певного типу комп'ютера, з тими СКБД (системами керування базами даних), якими він по+слугується, вибір форми організації даних, прийнятої і прийнятної для роботи таких СКБД, та ***фізичний*** – вибір раціональної структури організації бази даних у пам'яті комп'ютера, методів роботи з нею, виходячи з тих можливостей, які надають технічні показники комп'ютера (обсяг його оперативної та дискової пам'яті, швидкодія), його апаратне та програмне забезпечення, можливості СКБД. Стратегія створення показових і надійних лінгвістичних баз даних спрямована на їхнє багаторазове та різноаспектне використання, тобто оброблення за різними параметрами та типами інформації про подані в таких базах мовні об'єкти. В узагальненому вигляді процедуру проектування бази даних зображено на рис. 3:

І етап	**ІНФОЛОГІЧНИЙ**
добір об'єктів опису	
визначення їхніх властивостей та відношень між ними	
здійснення внутрішньої формалізації об'єктів, їхніх властивостей та відношень між ними	
створення концептуальної інфологічної моделі предметної галузі	
ІІ етап	**ДАТОЛОГІЧНИЙ**
	Логічна стадія
	здійснення зовнішньої формалізації інфологічної моделі
	вибір СКБД і припасування інфологічної моделі до її можливостей
	Фізична стадія
	вибір раціональної структури організації бази даних у пам'яті комп'ютера, методів роботи з нею з огляду на можливості СКБД

Рис. 3. Етапи проектування бази даних

За способом організації інформації лінгвістичні бази даних, як і бази даних з інших галузей знання, можна поділити на два основні типи: ***ієрархічні*** та ***реляційні***. Перші подають відомості про мовні об'єкти за різними відношеннями їхнього підпорядкування (родо-видовими або "частина-ціле" тощо). Другі – представляють інформацію про той чи

інший мовний об'єкт як сукупність його атрибутів, властивостей, характеристик, ознак. Така інформація в базі даних об'єднана лише на підставі її належності певному єдиному об'єкту, за ознакою відношення, або реляції до нього. Здебільшого лінгвістичні бази даних поєднують обидва ці способи організації про мовні об'єкти, оскільки сама мовна система організована як за відношеннями підпорядкування одиниць, так і за відношеннями включення або координації (співположення).

Залежно від обраних джерел формування лінгвістичних баз даних розрізняють ***словнико-*** та ***текстозорієнтовані бази***. Тип вихідного джерела, спосіб представлення та організації в ньому мовної інформації визначають не лише структуру інфологічних моделей баз цих двох різновидів, але й спосіб роботи з ними, одержання з них потрібної інформації та вигляд машинних продуктів опрацювання таких лінгвістичних баз.

Представлення в пам'яті комп'ютера тексту традиційного (=укладеного вручну, або "паперового") словника залежить від мети його подальшого використання. Якщо користувач не планує конструювати на основі такого тексту нові лінгвістичні об'єкти, наприклад, словники нових типів, або використовувати таку базу даних для інших лінгвістичних досліджень чи практичних потреб, його цілком задовольнить так звана ***комп'ютерна копія традиційного словника***, або ***словника-оригінала***. Таке представлення тексту словника не передбачає будь-якого втручання в його структуру, спосіб виділення типів інформації або відношень між ними тощо. Такі копії можуть бути цілком придатними і для пошукових дослідницьких завдань. Однак, як доводить практика, це стосується лише словників без правої, пояснювальної частини, а отже, словників, мінімально структурованих за типами інформації, на зразок орфографічних, орфоепічних, словників-індексів морфем або морфемних структур слів, як, наприклад, 2-томний словник-довідник І.Т.Яценка "Морфемний аналіз" (К., 1980-1981) або "Морфемний словник" Л.М.Полюги (К., 1983) чи "Великий зведений орфографічний словник сучасної української лексики" (Київ-Ірпінь, 2003). Для словників з розгалуженою правою частиною, або так званих екзегетичних (пояснювальних) словників (від грецьк. ἔκθεσις "пояснення, переказ, виклад") комп'ютерні копії текстів, розрахованих на людське сприйняття мовної інформації, виявляються малопридатними для використання. Тексти таких словників для виконання різноманітних пошукових завдань вимагають втручання лінгвіста-дослідника до процесу їхнього представлення, точніше до процедури унаочнення відношень між різними типами інформації (виведення, включення, виключення, заміна, суміщення тощо) в таких словниках. Для представлення текстів таких словників у базах даних створюють на відміну від копій їхні ***комп'ютерні версії***. Комп'ютерна версія традиційного словника має спеціальну ***граматику***, або правила аналізу та синтезу його тексту за типа-

ми інформації, вираженими спеціальними компонентами словникової метамови (=мови опису інформації, вміщеної в словнику). Процедура створення такої формальної граматики полягає у встановленні параметрів лексикографічного опрацювання мовної інформації, представленої в аналізованому традиційному словнику, відношень між такими параметрами, а також виявлення їхніх метамовних еквівалентів та вироблення процедур "розкладення" або, навпаки, "збирання" тексту статті словника за певними правилами. Аналіз статті словника полягає в одержанні елементарних (мінімальних) одиниць інформації зі складених; синтез становить процес, діаметрально протилежний аналізу, але не обов'язково тотожний за складом процедур, що передусім спричинює тип оброблюваної мовної інформації. Саме наявність машинної версії зі спеціальною граматикою її опрацювання робить текст традиційного словника придатним для виконання дослідницьких завдань, зручного й ефективного одержання зі словника тієї чи іншої інформації у будь-якій комбінації її типів, а також уможливлює видобування зі словника інформації нової, прихованої, для якої в тексті традиційного словника не було вироблено спеціальних метамовних еквівалентів або вона взагалі не була структурована. Отже, комп'ютерна версія відкриває можливості для виконання різноманітних процедур лінгвістичного конструювання, передусім конструювання словників нових типів. З огляду на це комп'ютерну версію традиційного словника можна розглядати як його в базі даних, придатну для аналізу та синтезу його тексту.

Рис. 4. Схема граматики машинної версії традиційного тлумачного словника (за Л.І.Колодяжною)
(**а** – зона інформації; **х** – тип інформації)

Залежно від мети створення бази даних комп'ютерні версії традиційних словників можуть представляти їхній текст у повному/неповному та зміненому/незміненому вигляді. Наприклад, стратегія створення бази даних морфемно-словотвірного фонду Інституту мовознавства ім.О.О.Потебні НАН України спиралася на комп'ютерні версії текстів традиційних словників у повному й неповному зміненому вигляді. Інфологічна модель цієї бази даних містила слово в канонічній (=вихідній, словниковій) формі, поділене на морфеми, як об'єкт опису та сукупність інформації про нього, дібраної з традиційних словників-джерел

різного типу. Для датологічної моделі було обрано зонний принцип подання інформації, розроблений для комп'ютерних версій традиційних словників московською дослідницею Л.І.Колодяжною[26]. Він становить на відміну від табличної форми запису гнучкий спосіб розміщення інформації в пам'яті комп'ютера за зонами однорідної інформації. Зручність і економність зонного принципу датологічного проектування бази полягає в значущості наявності/відсутності самої зони, а не інформації в ній. У граматиці машинних версій традиційних словників, розробленій Л.І.Колодяжною, між складниками – реалізаторами типів інформації у статті словника – встановлено відношення слідування або залежності. Ось, приміром, як виглядає схема такої граматики для машинних версій тлумачних словників сучасної російської мови[27] (див. рис. 4)

У свою чергу, в кожній із зон інформація може в разі потреби деталізуватися, а отже, для кожної зони в граматиці Л.І.Колодяжної існують компоненти і, відповідно, їхні словникові (метамовні) еквіваленти прості і складні, однорідні і неоднорідні. Крім того, структура такої граматики відкрита: вона може поповнюватися новими зонами або змінювати структуру зон залежно від інформації того чи іншого словника. Так, приміром, академічний тлумачний словник російської мови в 4-х томах (МАС) за ред. А.П.Євгеньєвої до запозичених слів подає інформацію про їхні етимони та джерела запозичення, а отже, в статтях до таких лексем з'являється ще одна зона – зона етимології, відсутня у тлумачному словнику С.І.Ожегова. Унаслідок цього на ІІ рівні аналізу тексту словникової статті тлумачних словників (див. рис. 4) поруч із зонами $\mathbf{a_2}$ та $\mathbf{a_3}$ в граматиці машинної версії МАС з'являється зона $\mathbf{a_4}$ з її атрибутами – типами інформації: $\mathbf{x_7}$ – етимон і $\mathbf{x_8}$ – пояснення етимона, джерело запозичення слова. Наприклад, статтю до реєстрового слова ***альмави́ва*** у МАС за правилами граматики Л.І.Колодяжної у такому розширеному варіанті можна розмітити так:

-a$_1$- *альмави́ва* (реєстрове слово **-x$_1$-**)**,** -ы, *ж.* (грам.ремарки **-x$_2$-**). *Устар.* (стил.ремарка **-x$_3$-**). (зона ремарок **-a$_5$-**) (зона реєстрового слова **-a$_2$-**). Широкий мужской плащ особого покроя (дефініція **-x$_4$-**). *Альберт простился с хозяйкой и, надев истертую шапку с широкими полями и летнюю старую альмавиву---, вышел на крыльцо* (цитата **-x$_5$-**). Л.Н.Толстой. Альберт (джерело цитати **-x$_6$-**).

---

[26] Див.: **Колодяжная Л.И.** Структура словарного текста в аспекте машинной лексикографии. – Автореф. дис. ...канд.филол.наук. - М., 1986.

[27] **Колодяжная Л.И.** Зазнач. праця. – С. 7. Зведену граматику машинних версій словників тлумачного типу Л.І.Колодяжна розробила за матеріалами 4-ох найавторитетніших словників сучасної російської мови: 17-томного «Словаря современного русского литературного языка» (так званого БАС (Большого академического словаря), 4-томного «Толкового словаря русского языка» за ред. Д.М.Ушакова, 4-томного «Словаря русского языка» за ред. А.П.Євгеньєвої (так званого МАС (Малого академического словаря) та однотомного "Словаря русского языка» С.І.Ожегова.

(зона прикладу **-a_6-**) (зона дефініції **-a_3-**)

[По имени графа Альмавивы (етимон **-x_7-**) – героя комедии Бомарше "Севильский цирюльник"] (пояснення етимона **-x_8-**) (зона етимології **-a_4-**)

Кожна зона та підзона мають свій маркер – символ розмітки і розпізнавання цього складника статті під час її комп'ютерного опрацювання. Як бачимо, схема граматики Л.І.Колодяжної подає багаторівневу структуру організації інформації про слово в тлумачних словниках. Вона містить складені (об'єкти **a**) та елементарні, далі нерозкладні (об'єкти **x**) компоненти. Причому одні складені компоненти можуть, у свою чергу, входити до складу інших складених компонентів, а отже, мати різний ступінь складності, напр.: компонент $\mathbf{a_1}$ (всю статтю словника) можуть складати компоненти $\mathbf{a_2}$, $\mathbf{a_3}$ та $\mathbf{a_4}$. У свою чергу, компонент $\mathbf{a_2}$ вміщує $\mathbf{a_5}$, а $\mathbf{a_3} – \mathbf{a_6}$.

Аналіз структури статей різних словників української мови подала у своїй праці "Нариси з комп'ютерної лінгвістики" М.М.Пещак[28]. Він спрямований на виділення типів інформації, з'ясування способів їхнього оформлення у метамові того чи іншого словника і, зрештою, – на розроблення граматики для комп'ютерних версій таких традиційних словників та конструювання на їхній основі словників нових типів. В Українському мовно-інформаційному фонді НАН України на основі аналізу структури статей тлумачного "Словника української мови" в 11 тт., найбільш місткого за типами інформації про слово, створено спеціальні засоби аналізу і, відповідно, конструювання реєстрових і пояснювальних (інтерпретаційних) частин статей до слів різних частин мови. Зокрема, для дієслів виділені такі типи інформації, що становлять окремі блоки процедури аналізу або синтезу словникової статті: парадигматичні та стилістичні показники, показник перехідності/неперехідності. Кінцеву ланку аналізу словникової статті становлять конкретні значення – виразники виділених типів інформації, а синтезу – реєстрова частина в цілісному вигляді як поєднання реєстрового слова (слів) з атрибутами інформації про них[29]. Таке ж детальне структурування здійснюється й для виділення типів інформації у тлумачній, або пояснювальній, інтерпретаційній частині словникових статей. Їхні конкретні комбінації в окремих статтях зумовлені обсягом і складом семантичної структури певного слова. Ось, наприклад, які типи інформації про семантику слова ***душа́*** можна виявити в тлумачній частині статті до нього у 2-ому томі СУМу: 1) формули тлумачення окремих значень та відтінків значень з супровідними граматичними та стилістичними ремарками; 2) ілюстрації до формул тлумачення з їхніми паспортами (джерелами цитат). Окремі підстатті у складі загальної статті до цього слова станов-

[28] **Пещак М.М.** Нариси з комп'ютерної лінгвістики.– Ужгород, 1999. – С.111-166.

[29] Див. докладніше в: **Широков В.А.** Інформаційна теорія лексикографічних систем. – К., 1998.

лять тлумачення до різних типів фразеологічних одиниць – ідіом, зрощень або сталих словосполук, кожен з яких має свій спеціальний маркер-символ – ромб, трикутник або відсутність символа, а лише абзацне виділення такої одиниці і подання її напівжирним шрифтом. Кожна фразеологічна одиниця, у свою чергу, має формули тлумачення й ілюстрації до них з відповідними паспортами. Тлумачна частина статті до слова ***душа́*** має 5 формул тлумачень його основних значень та 2 формули тлумачення відтінків значень, у статті також подані тлумачення 49 фразеологічних одиниць на зразок ***Відво́дити (відвести́, одво́дити, одвести́) ду́шу, ревізька душа́, чорни́льна душа́, в одну́ ду́шу, криви́ти душе́ю, чого́ душа́ забажа́є (захо́че*** *і т.ін.*). У свою чергу, формули тлумачення також становлять об'єкт структурування для з'ясування принципів семантичної класифікації лексики в тлумачному словнику, виділення метамовних еквівалентів тих чи інших складників семантичної структури слова. Структуровані за типами семантичної інформації формули тлумачення СУМу створюють підґрунтя для конструювання нових типів пояснювальних словників, зокрема ідеографічних та семантичних. Про це докладніше йтиметься у §5 цього розділу, присвяченому комп'ютерній лексикографії.

У датологічній моделі бази даних морфемно-словотвірного фонду Інституту мовознавства ім.О.О.Потебні НАН України 7 зон інформації: 1 головна і 6 залежних, які складають так званий ***інформаційний кортеж*** до головної зони із вміщеним у ній об'єктом опису – словом, поділеним на морфеми[30]. Зони інформаційного кортежу містять інформацію про наявність слова в реєстрах словників-джерел формування фонду та про його частиномовну належність. Як джерела формування бази даних фонду відібрано 5 словників сучасної української мови з різними аспектами опису слів, що й зумовило різний спосіб представлення в цих словниках кількісного та якісного складу сучасного українського лексикону. Це: тлумачний "Словник української мови" в 11 тт. (К., 1970–1980) (далі – СУМ), 2-томний словник-довідник І.Т.Яценка "Морфемний аналіз" (К., 1980–1981), 2-томний "Частотний словник сучасної української художньої прози" (К., 1981), "Словник іншомовних слів" за ред.О.С.Мельничука (К., 1974) та орфографічна частина "Словника-довідника з правопису та словоживання" С.І.Головащука (К., 1989), яка містить багато нових слів, що не потрапили до реєстру СУМу. Отже, реєстри названих словників доповнювали один одного, а в результаті дали можливість сформувати певне об'єктивоване й різноаспектне представлення складу сучасного українського лексикону.

[30] У спеціальній літературі запис бази даних, що містить об'єкт опису та інформацію про нього, називають також ***одиницею зберігання***. Окремі складники одиниці зберігання становлять її характеристики (Див.: **Баранов А.Н.** Введение в прикладную лингвистику. – М., 2001. – С.115).

Зведений за матеріалами цих словників реєстр слів сучасної української мови – стрижень морфемно-словотвірного фонду - налічує 166385 одиниць (для порівняння – реєстр СУМу містить близько 137 тис. слів, академічного "Українського орфографічного словника" 2002-го року – понад 143 тис. слів). 2003 р. у видавництві "Перун" (Ірпінь) вийшов друком "Великий зведений орфографічний словник сучасної української лексики" (укладачі В.Т.Бусел, М.Д.Василега-Дерибас, О.В.Дмитрієв, Г.В.Латник, Г.В.Степенко), реєстр якого містить понад 253 тис. слів. Це на сьогодні вичерпний перелік українських слів, засвідчених словниками та енциклопедіями, виданими в Україні у другій половині XX і в перші роки XXI століть.

За правилами синхронного морфемного аналізу у словах виділено мінімальні значущі одиниці – морфеми – 5-ти класів: корені, префікси, суфікси (до них залучено і дієслівний постфікс **–ся**), флексії канонічних форм слів і в складних словах-композитах – з'єднувальні голосні. Для кожного класу морфем для датологічній моделі вироблено спеціальні маркери. Корінь подано у скісних дужках. Префікси в простому слові позначені знаком **&** (амперсант), якщо їх більше одного, то вони один від одного відокремлені знаком **+**. У складних словах префікси, розташовані між коренями, позначено символом **:**. Суфікси в простому слові містяться між знаками скісної дужки кореня і комою як показником межі зони слова. Якщо суфіксів більше одного, то вони також відокремлюються один від одного знаком **+**. У складних словах між коренями суфікси позначено символом **$**. Перед флексіями як у простих, так і в складних словах ставиться знак *****, цей знак без літер після нього позначає нульову флексію. У слові обов'язково відзначено наголос (наголоси – у складних словах-юкстапозитах на зразок ***дýмати-гадáти***). Для унаочнення йотації, що відбувається на морфемних швах, до запису морфемної будови слова вводиться літера **j**. До кожного слова обов'язково подано інформацію про його частиномовну належність. Крім 10 частин мови, виділених у традиційній граматиці, спеціальними символами у цій зоні позначено дієприкметники та дієприслівники як лексико-граматичні розряди слів зі своєрідними формальними, змістовими та функціональними властивостями. Символи обрано за першою літерою найменування частини мови; в разі збігу перших літер найменувань використовувалися двобуквені символи, пор.: **І** – іменник, **Д** – дієслово, **П** – прикметник, **ПС** – прислівник, **З** – займенник, **Ч** – числівник, **ПР** – прийменник, **ЧК** – частка, **С** – сполучник, **В** – вигук, **ДП** – дієприкметник та **ДС** – дієприслівник.

Кількість зон словникової інформації у конкретних записах бази даних залежить від наявності/відсутності слова в реєстрах словників-джерел її формування. Максимальна кількість зон за реєстрами словників – 5, мінімальна – 1, а отже, максимальний запис у базі містить 7 зон інформації, мінімальний – 3 обов'язкові: зони слова, інформації за

одним із словників-джерел та показника частиномовної належності. Для зони кожного словника також вироблено спеціальні маркери, зорієнтовані на назву, тип словника або прізвище його укладача: **M** – словник "Морфемний аналіз"; **T** – тлумачний словник, **F** – "Частотний словник сучасної української художньої прози" (від англ. **frequency** "частота"), **X** – "Словник іншомовних слів" (від грецьк. ξένος "чужий") та **G** – словник С.І.Головащука. Для зон словників, що мають праву, пояснювальну, частину, в датологічній моделі передбачено введення, крім позначення самого словника, змістової інформації – наповнення зони, а саме: для зон **T** та **X** – кількість значень слова, а для зони **F** – показник абсолютної частоти вживання слова у півмільйонній текстовій вибірці, на основі якої було укладено "Частотний словник сучасної української художньої прози". Ось, наприклад, як виглядає у записах бази даних морфемно-словотвірного фонду інформація про слова **комп'ютер, мова, український, лінгвістика, студент, університет, залік, іспит**:

&/комп'ю́тер/*, T1, G, X1, I
&/мо́в/*a, T6, F129, I
&/украї́н/ськ*ий, M, T2, F94, П
&/лінгв/і́стик*a, M, T1, X1, I
&/студ/е́нт*, T1, F56, I
&/університе́т/*, T2, F28, G, X3, I
&за́/лік/*, M, T1, F1, G, I
&і́с/пит/*, T2, F26, G, I

Записи дають змогу виявити ступінь вживаності того чи іншого слова, показником якого може служити наявність такого слова в одному чи більше словників. З огляду на якість реєстрів словників-джерел бази можна встановити причину того чи іншого ступеня вживаності лексеми. Наприклад, якщо слово зустрілося лише в словнику "Морфемний аналіз", словнику іншомовних слів або в словнику С.І.Головащука, реєстри яких широко подають різногалузеву термінологічну лексику, то мале поширення його в загальномовному лексиконі можна пояснити спеціалізованим характером його семантики. Якщо ж слово трапилося лише в частотному словнику сучасної української художньої прози, в реєстрі якого чимало індивідуально-авторських новотворів, розмовної, а то й жаргонної лексики, то ступінь вживаності таких лексем можна пояснити оказіональним характером таких одиниць. Наприклад, до лексем з обмеженою сферою вживання належать такі слова, відзначені лише реєстром "Частотного словника сучасної української художньої прози":

&/ра́д/іс+н+о-/бі́л/*ий, F1, П
&/фарт/о́в*ий, F1, П
&/філіжа́ноч/к*a, F1, I
&/рай//гус/а́к*, F1, I
&/ра́т/н+о, F2, ПС

&до/боро́/ти, F1, Д

або лексеми, відзначені лише реєстром "Словника іншомовних слів":

&ана́/лог/ов*ий, X1, П
&/гум/і́н+ов*ий, X1, П
&/нав//а́рх/*, X1, І
&/са́го/*, X1, І
&/сакс//го́рн/*и, X1, І
&/пір/о/фо́р/н*ий, X1, П

Максимальний ступінь вживаності мають слова, засвідчені реєстрами всіх словників-джерел, пор.:

&/абстра́кц/ij*а, М, Т3, F1, X3, G, І
&/гімна́з/ij*а, М, Т1, F12, X1, G, І
&/комфорт/а́бель+н*ий, М, Т1, F2, X1, G, П
&/математи́ч/н*ий, М, Т1, F5, X1, G, П
&/форс/ува́+ти, М, Т4, F2, X1, G, Д

Зауважимо однак, що цей критерій не завжди дозволяє віднести такі лексеми до ядра сучасного українського лексикону, оскільки принаймні в художніх прозових текстах вони мають, як бачимо, низьку частоту вживання. Отже, для з'ясування активності тієї чи іншої лексеми в сучасній українській мові показник її залучення до словників, різних за способом формування та опису реєстрів, слід узгоджувати з показниками частоти її вживання в текстах різних функціональних стилів та різної тематики.

На основі цієї основної бази даних створено кілька похідних від неї баз, баз-сателітів, які містять інформацію про морфеми окремих класів у складі її слів, а також про моделі морфемної будови таких слів. Таких баз-сателітів 6: 5 – для морфем окремих класів і 1 – для моделей морфемної будови слів, що становлять організовані сукупності символів-маркерів класів морфем. Бази даних про морфеми окремих класів побудовано у вигляді упорядкованих за алфавітом реєстрів конкретних морфем – коренів, префіксів, суфіксів, флексій та з'єднувальних голосних з показниками їхньої частоти вживання в словах основної бази. Порядкові номери морфів у реєстрах баз-сателітів важливі для подальшого опрацювання морфемної структури слів і для зберігання й перетворення основної бази даних, оскільки структури реальних слів у разі потреби можна замінити сукупностями порядкових номерів морфів, що входять до їхнього складу.

Наприклад, база префіксів, реєстр якої налічує 145 одиниць (морфів у реальних словах мови), виглядає таким чином:

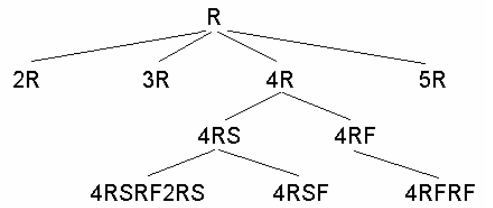
Порядковий номер	Префікс	Кількість вживань у словах реєстру[31]
1.	а	198
2.	ан	69
3.	анти	142
4.	архі	19
5.	без	1099
6.	в	3267
7.	ви	4296
8.	від	3018
9.	віді	74
10.	відо	5

Одиницею опису в базі даних про моделі морфемної будови слів є послідовність символів класів морфем у складі конкретних слів основної бази. Вони згруповані в реєстрі навколо моделей з мінімальною кількістю певних символів, а отже, сама ця база подає можливе розгортання певних ядерних моделей морфемної будови слова, подібне до того, яке ми вище представили в морфемних сітках простих слів (див. рис. 3). Наприклад, вище з морфемної сітки простих іменників ми дізналися, як в ній може ускладнюватися – "розгортатися" вліво та вправо – модель морфемної будови слова **R.** А ось як вона може розгортатися в складних словах. Коло кожної такої моделі подано кількісні показники її реалізації в словах фонду:

Порядковий номер	Модель	Кількість слів
1.	R	8823
2.	RR	1929
3.	RRR	240
4.	RRRR	14
5.	RRRRR	2
6.	RRRRS	1
7.	RRRRSRFRRS	1
8.	RRRRSF	5
9.	RRRRF	1
10.	RRRRFRF	2

Цей реєстр можна також представити у вигляді орієнтованого графа, у вузлах якого вміщено вихідні структури. Напрямок розгортання його ребер засвідчує спосіб творення похідних структур, а кількість таких ребер – ступінь їхньої складності. Зокрема, подані вище структури можна у графі розмістити в такий спосіб:

[31] Цей показник не тотожний показнику кількості окремих слів з таким префіксом, оскільки деякі префікси в одному простому чи складному слові можуть вживатися по кілька разів, напр.: ***попоходити, прапрабабуся, вибірково-вимірювальний, заготівельно-закупівельний***.



Уміщені на цьому графі символьні моделі морфемної будови слів описують такі українські слова (питомі та іншомовного походження), як ***а, бовть, авокадо, мурмурандо, фрау*** (модель **R**), ***аби-хто, біг-вість, беф-строганов, де-факто*** (модель 2R), ***а-ні-чичирк, ка-зна-де, як-не-як*** (модель **3R**), ***зоо-пале-онто-лог-ій-а, стерео-фото-грам-метр-ичн-ий*** (модель **4RSF**), ***сім-сот-п'ят-десят-и-літт-я*** (модель **4RFRF**) тощо.

Усього в цій базі 694 символьні моделі: 51 – для простих і 643 – для складних слів. Це в стисненому вигляді механізм творення морфемної будови слів сучасної української мови, а морфемні сітки, або орієнтовані графи впорядкування таких одиниць становлять моделі такого словопороджувального механізму мови.

Для роботи користувачів як з основною базою даних, так і з базами-сателітами створено спеціальний ***інтерфейс*** – засоби доступу до них і отримання з них потрібної інформації[32]. Такий засіб зв'язку між користувачем і базою даних у морфемно-словотвірному фонді становить діалогова система "МОРФОЛОГ", яка дозволяє працювати з базою в інтерактивному режимі. Її створили лінгвісти-розробники фонду (Н.Ф.Клименко та Є.А.Карпіловська) разом з математиками-програмістами (С.Г.Буригіним, М.А.Перельмутером та В.С.Карпіловським). Інтерактивний, або онлайновий режим, режим опрацювання інформації в масштабі реального часу на відміну від так званого пакетного режиму уможливлює роботу з інформацією бази з втручанням користувача в процес її опрацювання або з можливим візуальним (на дисплеї комп'ютера) контролем ходу виконання поставленого завдання. Крім того, система "МОРФОЛОГ" дає змогу користувачеві вибрати той масив інформації, який його цікавить: основну чи залежні бази, загальну інформацію про склад бази, в основній базі працювати тільки зі словом без інформаційного кортежу до нього чи із записом в базі в цілому

[32] Ми далі будемо спиратися саме на таке розуміння цього терміна, розуміння, яке відбиває специфіку спілкування з комп'ютером людини-користувача. У термінології обчислювальних систем **інтерфейс** (від англ. **interface** – буквально – між-обличчя, обличчя посередині) трактують як "елементи з'єднання та допоміжні схеми керування, використовувані для з'єднання пристроїв" або як характеристику взаємозв'язку двох програмних одиниць (див.: **Толковый** словарь по вычислительным системам. – С.243-244).

(слово+ його інформаційний кортеж), а також визначити різновид роботи з базою: виконання дослідницьких завдань, одержання з бази довідкової інформації чи редагування записів бази даних. Для роботи з окремим словом чи зі словом та інформаційним кортежем до нього створено спеціальні засоби переведення запису слова в морфемах в його орфографічний запис і навпаки. Для цих різновидів роботи з базою розроблений спеціальний інтерфейс з відповідним ***меню***. Меню називають певний виведений на екран дисплею список можливих способів (режимів) роботи з базою даних. У системі "МОРФОЛОГ" меню має вигляд дерева впорядкування типів інформації про слово, поданих в основній базі даних. У свою чергу, кожен з типів інформації, вміщених у такому дереві, становить окремий спосіб доступу й роботи з базою для одержання з неї такого роду інформації. На рис. 5 та 6 подано структуру загального інтерфейса діалогової системи "МОРФОЛОГ" та структуру спеціального інтерфейса в складі цієї системи для роботи з окремим словом та словом і його інформаційним кортежем:

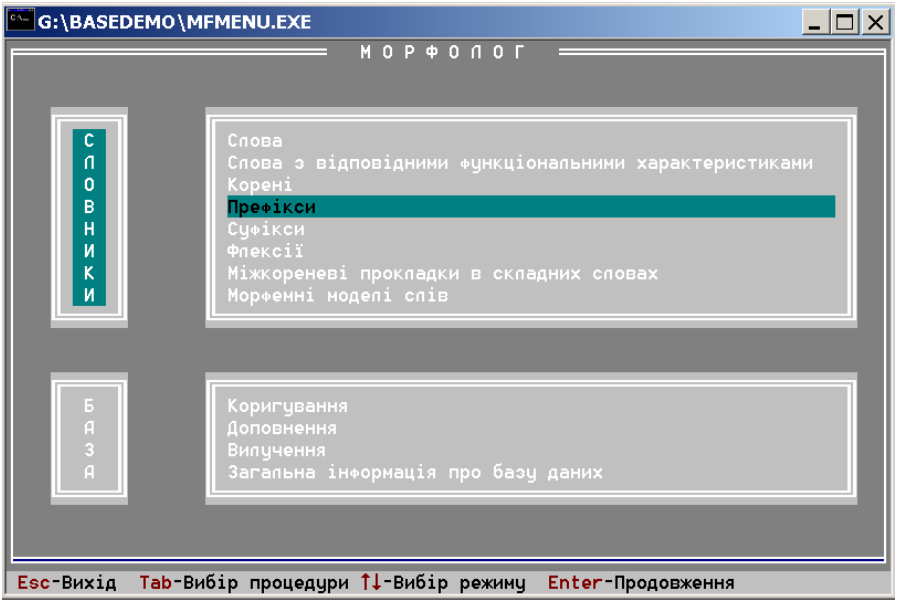


Рис. 5. Меню загального інтерфейса діалогової системи "МОРФОЛОГ"

Інтерфейс, поданий на рис. 6, дозволяє користувачеві структурувати інформацію бази даних за її типами, комбінувати типи інформації у той чи інший, потрібний для виконання певного завдання спосіб. Перші чотири різновиди роботи налаштовані на аналіз самого слова. ***Маска*** (буквена або морфемна) становить певний реальний або довіль-

ний образ буквеної або морфемної структури слова (=сукупність символів) для пошуку слів у базі даних. Маска як довільний образ пошуку нагадує відому всім зі шкільних років гру у відгадування слова за буквами. Ще не "відгадані" букви у масці пошуку замінює знак питання **?**. Якщо нас цікавить лише певний початок слова, то для позначення всіх можливих завершень слова у масці використовується символ ***** (зірочка) Так, наприклад, буквена маска пошуку **г?ро*** представляє сукупність зі 138 слів бази даних – 95 іменників, 35 прикметників, 5 дієслів, 2 прислівників та 1 дієприкметника, серед яких містяться слова ***геро́й, гірокóмпас, горо́д, горошúна, гаровúй, гіромагні́тний, гороб'я́чий, горóхуватий, героїзува́ти, городникува́ти, гороріźьблений, геро́ї́чно, горо́ю***. Додаючи до режиму пошуку інформації за буквеною маскою слова один, кілька або і всі інші режими, передбачені інтерфейсом, користувач дістає змогу структурувати одержану сукупність зі 138 слів за іншими параметрами їхніх форми, семантики або функціонування в мові. Наприклад, завдяки режиму **Маска морфемної структури** з'ясовуємо, що серед таких слів наявні прості і складні (кількакореневі) одиниці, встановлюємо моделі їхніх морфемних структур, визначаємо їхню активність для слів цієї буквеної будови.

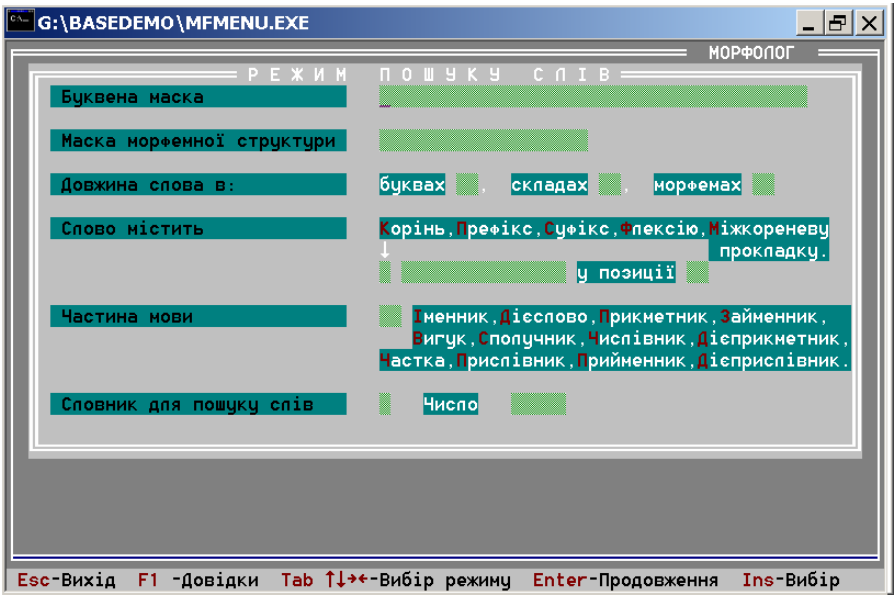


Рис. 6. Меню спеціального інтерфейса діалогової системи "МОРФОЛОГ" для роботи з основною базою даних морфемно-словотвірного фонду

Кожен з режимів роботи з базою дозволяє розкласифікувати слова за певним параметром. Наприклад, режим **Довжина слова у буквах** дозволяє одержати інформацію про інтервал можливих для слів сучасної української мови структур за кількістю букв. Встановлено, що для простих (=з одним коренем) слів такий інтервал становить 1–22 букви (***а, в, інтернаціоналізува́тися, конституціоналі́стський***); а для складних (кількакореневих) – 5–30 букв (***аби́як, якра́з, амлітýдно-і́мпульсно-модульо́ваний***).

Деталізоване структурування бази даних за типами інформації про слово, а також створені спеціальні засоби опрацювання такої інформації дають можливість описаній вище словникозорієнтованій базі даних морфемно-словотвірного фонду української мови, створеній в Інституті мовознавства ім.О.О.Потебні НАН України, виконувати цілий ряд функцій, а саме: інформаційно-довідкову, навчальну, редакційно-видавничу і власне лінгвістичну, дослідницьку.

🕮 Терміни

- **комп'ютерне середовище для лінгвіста** – бази даних про мовну систему або продукти мовленнєвої діяльності та комп'ютерні (апаратні (інструментальні (hardware)) та програмні (software) засоби доступу до них та їхнього оброблення
- **база даних (database)** – структурований і формалізований масив інформації про певну предметну галузь
  - **ієрархічна база даних –** база даних, організована на основі ієрархічних відношень між мовними об'єктами або відомостями про них, напр. родо-видових відношень, відношень "частина-ціле" тощо
  - **реляційна база даних –** база даних, відомості в якій організовані на основі їхньої належності спільному об'єкту опису або спільному атрибуту такого об'єкта
  - **словникозорієнтована (=словникова, лексикографічна) база даних** – база даних, побудована за текстами певних словників
  - **текстозорієнтована (=текстова) база даних** – база даних, побудована на основі певного корпусу текстів
  - **інфологічний етап проектування бази даних –** етап відбору інформації та її структурування
  - **датологічний етап проектування бази даних –** етап оформлення інформації відповідною мовою представлення, придатною для комп'ютерного опрацювання
    - **логічна стадія датологічного етапу –** пов'язання концептуальної інформаційної моделі з операційною системою певного типу комп'ютера, з тими СКБД (системами керу-

вання базами даних), якими він послугується, вибір форми організації даних, прийнятої і прийнятної для роботи таких СКБД
    - **фізична стадія датологічного етапу –** вибір раціональної структури організації бази даних у пам'яті комп'ютера, методів роботи з нею, виходячи з тих можливостей, які надає СКБД
- **зонний принцип запису мовної інформації –** гнучка модель запису інформації, в якій кожному типу інформації відведено окрему зону і значущою є сама наявність/відсутність таких зон
- **формалізація –** структурований опис інформації в певній формі
    - **формалізація внутрішня** – формалізація предметної галузі, у прикладній та комп'ютерній лінгвістиці – мови; створення інформаційної моделі мовної інформаціїі
    - **формалізація зовнішня** – формалізація інформаційної моделі предметної галузі; у прикладній та комп'ютерній лінгвістиці – інформаційної моделі мови, створення формальних способів її представлення для роботи з комп'ютером або іншими засобами опрацювання інформації
    - **машинний продукт (= продукт оброблення бази даних, =продукт обчислення)** – втілений у певному комп'ютерному форматі (формі представлення) результат автоматичного оброблення бази даних
    - **інтерфейс (= засоби доступу до бази даних та її ведення)** – спеціальні системи спілкування з базою даних, призначені для її поповнення або редагування, а також для одержання з неї інформації, потрібної для виконання завдання користувача
    - **меню –** виведений на екран дисплею перелік можливих способів (режимів) роботи з базою даних
- **режим оброблення бази даних або бази знань** – спосіб роботи з базою даних (базою знань), доступу до неї
    - **інтерактивний (= діалоговий, онлайновий (on-line)) режим, режим опрацювання інформації в масштабі реального часу) –** спосіб оброблення даних з можливим втручанням користувача на будь-якому етапі виконання поставленого завдання або з можливим візуальним (на дисплеї комп'ютера) контролем ходу його виконання
    - **пакетний режим –** спосіб оброблення даних у пакеті – певній послідовності завдань
- **база знань (knowledge base), інтелектуальна база даних (intelligent database) –** структурований і формалізований набір відомос-

тей про об'єкти певної предметної галузі та відношення між ними, на основі яких можна будувати судження про них, здійснювати різноманітні операції логічних умовиводів

- **комп'ютерна копія традиційного словника –** комп'ютерний аналог тексту традиційного словника, створений без додаткового втручання лінгвіста-дослідника
- **комп'ютерна версія традиційного словника –** комп'ютерний аналог тексту традиційного словника, який має спеціальну граматику його опрацювання, підготовлену лінгвістом-дослідником
  - **граматика комп'ютерної версії традиційного словника –** правила аналізу та синтезу комп'ютерної версії традиційного словника за метамовними еквівалентами типів поданої в ньому інформації

§4. Лінгвістичний алгоритм та лінгвістичний процесор

- Алгоритм, його ступені деталізації та програмна реалізація
- Вимоги до побудови алгоритму
- Способи представлення алгоритму: блок-схема (граф-схема) алгоритму: її складники і правила їх організації
- Алгоритм і процесор. Лінгвістичний процесор: лексикографічні (=словникові) та корпусні (=текстові) лінгвістичні процесори

Для виконання певного завдання комп'ютер треба забезпечити інструкцією, яка описує послідовність його дій, спрямованих на одержання бажаного результату. Таку визначену послідовність правил розв'язання завдання в математиці називають ***алгоритмом.*** Поняття це виникло ще задовго до створення комп'ютера. Саме найменування такої логічної процедури походить від латинізованого варіанта імені узбецького математика Мухаммеда бен Муси аль-Хорезмі (787 – бл. 850 рр.), автора знаменитого трактату з арифметики та алгебри "Книга про відновлення та протиставлення", який її придумав і вперше застосував у виконанні математичних обчислень. Людина може здійснювати певні операції і без алгоритму, без заздалегідь чітко визначеної послідовності своїх дій. Крім того, людина може сама виробляти певні послідовності дій, аналізувати їх і вибирати найкращу або ж таку, що дає змогу досягти бажаного результату завдяки найменшій кількості операцій. Комп'ютер без алгоритму не здатен виконувати будь-яке завдання, оскільки не має людського інтелекту з його здатністю до прийняття рішень. "Енциклопедія кібернетики" подає таку дефініцію алгоритму: **"Алгоритм** – точно визначене правило дій (програма), для якого зада-

но вказівку, як і в якій послідовності це правило слід застосовувати до вихідних даних завдання, щоб одержати його розв'язання"[33].

Алгоритм може мати різні ступені деталізації послідовності виконання правил і, відповідно, різні способи представлення: від найбільш узагальненого до максимально деталізованого, з ретельним описом всіх кроків виконання тієї чи іншої логічної операції в його складі. У кібернетиці максимально деталізований спосіб представлення алгоритму здобув назву ***машинного алгоритму***. Він становив опис послідовності виконання певних дій конкретних пристроїв комп'ютера з конкретними даними. Для представлення машинних алгоритмів послугувалися спеціальними машинними мовами – мовами машинних кодів. З розвитком програмування відпала необхідність у таких ретельних «інструкціях» комп'ютеру. Змінилося поняття програмної реалізації алгоритму, тобто його представлення комп'ютеру засобами певної мови програмування. Сучасні мови програмування високого рівня (Сі$^{++}$, ПАСКАЛЬ та ін.) зменшують кількість приписів комп'ютеру, як виконати той чи інший крок алгоритму і завдяки цьому дають можливість спростити його програмну реалізацію. Самий спосіб запису алгоритму виконання завдання, зокрема лінгвістичного, можна змінювати в напрямку його зближення з певним способом програмної реалізації, тобто з його представленням засобами певної мови програмування.

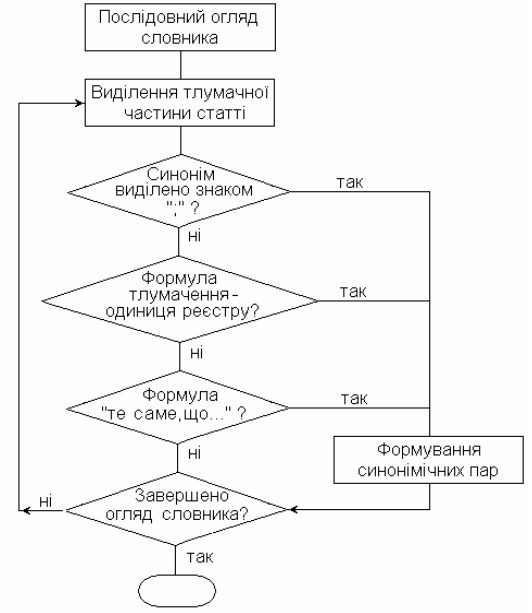


Рис. 7. Блок-схема алгоритму формування синонімічних пар дієслів за формальними показниками синонімії слів у СУМі

[33] **Энциклопедия** кибернетики. – К., 1974. – Т.1. – С.94.

Складниками алгоритму є певні дії, що виконуються за певних умов. Графічне представлення алгоритму становить його ***блок-схема***, або ***граф-схема***, яка подає хід виконання поставленого завдання. Прийняті певні правила побудови та графічного зображення таких схем алгоритмів. Умови виконання дій у блок-схемі вміщують у ромбах, а назви самих виконуваних дій – у прямокутниках, стрілками показують напрямок виконання завдання, залежність між тими чи тими умовами і діями, черговість їх виконання. За способом побудови така схема запису алгоритму становить орієнтований граф (звідки й друга поширена її назва) з чітким спрямуванням кроків виконання завдання та їхньою взаємозалежністю. На кожну умову-питання в алгоритмі можливі лише дві відповіді – **так** чи **ні** (інше їхнє зображення становлять знаки **+** та **-**). Завершення виконання завдання у блок схемі засвідчує зображення еліпса. Ось, наприклад, як можна представити алгоритм виконання певного лінгвістичного завдання, або ***лінгвістичний алгоритм*** в максимально узагальненому і відстороненому від конкретної програмної реалізації вигляді. Нижче на рис. 7 подаємо цей різновид алгоритму для виконання такого лінгвістичного завдання, як побудова синонімічних пар дієслів за формальними ознаками опису синонімічних відношень між ними в метамові СУМу (див. рис. 7):

Кожен алгоритм має певну кількість ***кроків*** свого виконання, зумовлених кількістю висунутих умов та дій, необхідних для одержання результату виконання завдання. Кожен крок описує окрему логічну операцію. Конкретне наповнення таких кроків, тобто зміст умов і підпорядкованих їм дій складають ***правила виконання алгоритму.*** Для ефективної роботи алгоритму правила його побудови повинні відповідати таким вимогам: 1) ***детермінованість (=визначеність)*** – результат, одержаний на кожному кроці роботи алгоритму повинен бути однозначним, несуперечливим і зумовленим результатами виконання попередніх дій; 2) ***дискретність*** – кроки алгоритму повинні відповідати чіткому логічному поділу процесу опрацювання вхідних даних на окремі етапи; 3) ***масовість*** – алгоритм повинен забезпечувати виконання однотипного завдання на інших даних, однотипних з тими, на основі яких його було розроблено.

Сукупність алгоритмів для опрацювання певних мовних даних з метою виконання різноманітних теоретичних і практичних лінгвістичних завдань називають ***лінгвістичним (=мовним) процесором*** (від англ. **processor** "виконавець дії"). Залежно від налаштування на спосіб представлення мовних даних – словники або тексти, а також від мети виконуваних завдань розрізняють ***лексикографічні (=словникові)*** та ***корпусні (=текстові)*** лінгвістичні процесори. У фаховій літературі можна зустріти як синоніми до цих термінів визначення ***автоматизовані лексикографічні системи*** та ***автоматизовані системи опрацювання тексту (АСОТ)***, або ***автоматизовані системи пе-***

реоблення тексту (АПТ). Про склад і можливості мовних процесорів докладніше мова піде в §§5 та 6 цього розділу та в §§3 і 4 другого розділу нашої книжки. Тут лише підкреслимо, що лінгвістичні процесори становлять сукупності алгоритмів і створених на їхній основі програм, призначених для аналізу 1) мовної інформації і 2) розв'язання теоретичних та практичних завдань завдяки саме лінгвістичному аналізу такої інформації.

Термін ***процесор*** запозичено з терміносистеми обчислювальної техніки, де він позначає обчислювальний пристрій, що керує виконанням операцій комп'ютера або, ширше, будь-якою системою приймання та передавання інформації. Частіше він виступає як синонім терміна ***центральний процесор*** (англ. **central processor**), який становить основний робочий елемент комп'ютера. Залежно від характеру виконуваних комп'ютером операцій, етапу опрацювання вхідних даних виділяють також ***препроцесор***, який займається підготовкою вхідних даних для виконання завдання, та ***постпроцесор***, що формує кінцевий продукт виконаного завдання для його роздруку або виведення на екран дисплею комп'ютера. Лінгвістичний процесор як певну систему засобів для комп'ютерного опрацювання мовної інформації також можна уявити як сукупність алгоритмів і програм для підготовки (препроцесор), опрацювання (центральний процесор) мовних даних та представлення користувачеві результату їхнього оброблення (постпроцесор). Лінгвіст відіграє провідну роль у процесі комп'ютерного опрацювання мовної інформації: він створює алгоритм виконання того чи іншого завдання, а отже визначає саму стратегію процесу оброблення мовних даних, готує вихідні дані для роботи алгоритму та аналізує одержані результати для можливого вдосконалення алгоритму.

🕮 Терміни

- **алгоритм –** чітко визначена послідовність дій, спрямованих на виконання певного завдання та одержання його результату
  - **правило виконання алгоритму –** сукупність умов та дій, підпорядкованих таким умовам, необхідних для виконання алгоритму
  - **крок алгоритму –** етап виконання завдання за правилами алгоритму, певна логічна операція
  - **мова представлення алгоритму** – спосіб запису правил алгоритму
  - **блок-схема (=граф-схема) алгоритму –** представлення правил виконання алгоритму у вигляді орієнтованого графа
  - **програмна реалізація алгоритму –** представлення алгоритму засобами певної мови програмування

- **процесор –** обчислювальний пристрій, що керує виконанням операцій комп'ютера або, ширше, будь-якою системою приймання та передавання інформації
  - **центральний процесор –** основний робочий елемент комп'ютера, який керує виконанням його операцій
  - **препроцесор –** програма, яка здійснює підготовку вхідних даних для виконання завдання
  - **постпроцесор –** програма, яка здійснює операції над вихідними даними іншої програми
  - **лінгвістичний (=мовний) процесор** – сукупність алгоритмів і створених на їхній основі програм для опрацювання певних мовних даних з метою виконання різноманітних теоретичних і практичних лінгвістичних завдань
    - **лексикографічний (=словниковий) процесор** – сукупність алгоритмів і створених на їхній основі програм для опрацювання інформації уже існуючих словників та конструювання нових комп'ютерних словників
    - **корпусний (=текстовий) процесор** – сукупність алгоритмів і створених на їхній основі програм для опрацювання інформації текстів та конструювання нових комп'ютерних текстів (первинних чи вторинних) або продуктів їхнього опрацювання: словопокажчиків, конкордансів, частотних словників, мовних картотек

§5. Комп'ютерна лексикографія: її предмет та завдання

- Комп'ютерний, або автоматичний словник і комп'ютерний варіант традиційного словника
- Типи комп'ютерних словників
- Лексикографічні (=словникові) процесори: склад та призначення

Для окреслення предмета і завдань комп'ютерної лексикографії, її зв'язків з лексикографією традиційною, яка в своїй теорії та практиці не орієнтується на комп'ютерне опрацювання мовної інформації, принципово важливо дати визначення ***комп'ютерного***, або ***автоматичного словника***. Як такий розглядаємо словник, процедури укладання якого здійснює комп'ютер. Відрізняємо комп'ютерний, або автоматичний словник від комп'ютерного варіанта чи копії традиційного, укладеного людиною словника. Останні подають лише нову, комп'ютерну форму інформації, вміщеної в традиційних словниках, а отже, це лише трансформація форми уже готового продукту лексикографічного опрацювання мовного матеріалу традиційними, некомп'ютерними методами. Комп'ютерний словник і комп'ютерний варіант традиційного слов-

ника становлять результати двох напрямків роботи у сучасній комп'ютерній лексикографії: 1) переведення у комп'ютерну форму вже існуючих словників, укладених людиною, створення на їхніх основі словникозорієнтованих баз даних різної структури та призначення і 2) розроблення понятєвого та процедурного апарату, лінгвістичних алгоритмів та процесорів для конструювання комп'ютерних словників нових типів.

В Україні вже накопичено цінний досвід роботи в обох цих напрямках. Прикладом першого може служити описана вище словникозорієнтована база даних морфемно-словотвірного фонду Інституту мовознавства ім. О.О.Потебні НАН України, бази даних 11-томного тлумачного "Словника української мови" та академічного "Орфографічного словника української мови" (перше видання – 1994 р.), сформовані в Українському мовно-інформаційному фонді НАН України[34], різноманітні частотні словники, тезауруси та конкорданси, створювані на основі корпусу текстів різних функціональних стилів у лабораторії комп'ютерної лінгвістики при кафедрі сучасної української мови Інституту філології Київського національного університету ім.Тараса Шевченка[35], термінологічні словникові бази даних, підтримувані лексикографічним процесором СЛОВО, розроблювані співробітниками університету "Львівська політехніка"[36], та ін. До цієї роботи залучено традиційні словники, різні як за складом реєстрів, так і за способом їхнього опрацювання: тлумачні, перекладні, термінологічні, морфемні, орфографічні та орфоепічні. Для забезпечення можливості виконання нових дослідницьких завдань створюються не комп'ютерні копії, а комп'ютерні версії таких традиційних словників зі своїми граматиками аналізу та синтезу текстів словникових статей. Перевага комп'ютерних версій над комп'ютерними копіями, як ми вже показували вище на прикладі використання версій традиційних словників для розв'язання дослідницьких завдань з морфеміки й словотворення, полягає у можливості їхнього багаторазового й багатоаспектного використання, автоматичної переорганізації, доповнення чи стиснення вміщеної в них інформації про ті чи інші мовні об'єкти. Продемонструємо це на прикладі використання словникозорієнтованої бази даних морфемно-словотвірного фонду Інституту мовознавства ім.О.О.Потебні НАН України для укладання комп'ютерних морфемних та словотвірних словників нових типів.

---

[34] **Ковтуненко Л.С.** Комп'ютерні аспекти лексикографічних систем // Мовознавство. – 1996. - № 4-5. – С.28-34; **Пещак М.М.** Зазнач. праця.

[35] **Дарчук Н.П., Грязнухіна Т.О.** Частотний словник сучасної української публіцистики // Мовознавство. – 1996. - № 4-5. – С.15-18.

[36] **Коссак О., Коруд О., Хвищун Л.** Система комп'ютерної підтримки словникових баз даних СЛОВО // Проблеми українізації комп'ютерів.–К., 1993.–С.73-76.

1998 р. світ побачив перший в Україні комп'ютерний "Словник афіксальних морфем української мови" (автори – Н.Ф.Клименко, Є.А.Карпіловська, В.С.Карпіловський, Т.І.Недозим), укладений за матеріалами цієї бази даних[37]. Для конструювання цього словника було розроблено спеціальні лінгвістичні алгоритми опрацювання афіксальної частини слів. Словник в цілому подає інформацію про 817 афіксальних морфів: 145 префіксальних та 672 суфіксальні. Для кожного з класів афіксальних морфем відведено окрему частину словника, причому з огляду на різний обсяг реєстрів префіксальних і суфіксальних морфів для представлення типів інформації у кожній з частин було створено свій формат словникової статті, свої способи та засоби організації інформаційних кортежів до реєстрових одиниць. ***Форматом статті комп'ютерного словника*** прийнято називати модель організації, розміщення та графічного представлення в словнику інформації про описувані в ньому мовні об'єкти. Вироблення такого формату становить разом зі створенням бази даних та лексикографічного процесора невід'ємний складник процесу укладання комп'ютерного словника.

Частини описуваного комп'ютерного словника містять спільні й відмінні типи інформації про афіксальні морфеми, що зумовлене спрямуванням алгоритмів їхнього укладання на розв'язання різних власне дослідницьких завдань. У префіксальній частині комп'ютер за встановленими формальними ознаками збирав морфи-варіанти "під дах" спільної системної одиниці – морфеми, а на основі кількісних показників вживання таких формальних варіантів у словах мови визначав морф-домінанту, що й виконував роль системної одиниці – морфеми. Наприклад, у словах бази даних виявлено три префіксальні морфи-варіанти зі спільним категоріальним значенням "спрямування дії під об'єкт": **під-, піді-, підо-**. На підставі обстеження кількісних показників їхнього вживання у словах статус домінанти такого морфного ряду надано морфу **під-** (2813 слів з 2932 обстежених), далі за спадом показників у ряду впорядковано залежні від домінанти морфи **піді-** (73 слова з 2932) та **підо-** (46 слів з 2932).

У суфіксальній частині словника на основі вивчення формальних, змістових та функціональних властивостей суфіксальних морфів (їхніх позиційних та комбінаторних (сполучувальних) характеристик) комп'ютер відокремлював омографічні морфи (такі, що могли за допомогою однієї форми (однієї буквеної структури) виражати кілька категоріальних значень) від неомографічних. Для омографічних морфів було встановлено в цілому спектри категоріальних значень, що вони їх здатні виражати. Для кожного з реалізаторів омографічного морфа з певним категоріальним значенням встановлено діагностичні формальні

[37] **Клименко Н.Ф., Карпіловська Є.А., Карпіловський В.С., Недозим Т.І.** Словник афіксальних морфем української мови. – К., 1998.

ознаки в слові. Спектри таких діагностичних ознак можуть сягати різної глибини залежно від характеру омографії певного суфіксального морфа та кількості таких морфів у слові. Наприклад, для того, щоб виявити суфіксальний морф **-ат-** у значенні "опредметнена ознака", достатньо перевірити його правого партнера у слові. Якщо це флексія **–ий**, то комп'ютер однозначно приписує реалізатору цього омографічного морфа таке значення, незалежно від конкретного кореня або префікса, що сполучається з певним коренем, пор.: ***бород-а́т*ий, по-вож-а́т*ий, не-жон-а́т*ий***. Для інших омографічних морфів перевірка на сполучуваність з безпосередніми партнерами в слові виявляється недостатньою, бо вони також мають тотожну форму, пор.: ***ро́з-ви-т-ок-0*** і ***с-пов-и́-т-ок-0*** або ***пт-а́ш-н*я*** і ***пт-аш-н*я́***. У першому випадку, щоб розрізнити реалізатори омографічного суфіксального морфа **-ок** зі значеннями "процес" і "предмет", треба до перевірки залучити сполучуваність таких морфемних структур з певними префіксами, а в другому випадку, щоб вирізнити реалізатори омографічного морфа **-н-** зі значеннями "місце" та "збірність", треба ввести до алгоритму інформацію про місце наголосу в слові, тобто вказати, наголошені чи ненаголошені праві або ліві партнери цього суфіксального морфа у слові.

Створені алгоритми аналізу морфемної структури слів і дали змогу комп'ютеру укласти морфемний словник нового типу – частотно-валентний, аналогів якому поки що немає в слов'янській не лише комп'ютерній, а й традиційній лексикографії. Реєстр префіксальної частини словника складають 74 системні одиниці-морфеми, з яких 39 неваріабельні (такі, що не мають формальних варіантів) та 35 варіабельні (з певною – від 1 до 8 – кількістю морфів-формальних варіантів). Варіабельними/неваріабельними можуть бути як питомі префікси, так і префікси іншомовного походження. Наприклад, не мають формальних варіантів такі префікси, як **ви-** *виразити*, **за-** *заспівати*, **ре-** *регенерація*, **сюр-** *сюрреалізм*. Здатність мати формальні варіанти виявили префікси **у-** *упадати* (пор. його варіанти **в-** *впадати*, **во-** *ворушити*, **вві-** *ввійти*, **уві-** *увійти*, **ві-** *вілляти*, **ув-** *увостаннє*, **ву-** *вустілка*), **син-** *синхронія* (пор. його варіанти **си-** *симетрія* та **сим-** *симфонія*) та ін.

Реєстр суфіксальної частини словника містить неомографічні (з одним категоріальним значенням) та омографічні (з кількома категоріальними значеннями) суфіксальні морфи, причому неомографічних суфіксальних одиниць майже вдвічі більше, ніж омографічних (408 проти 264). Це можна пояснити передусім флективним характером української мови, який спричинює більшу кількість формальних перетворень одиниць саме післякореневої частини слів. Не всі вони, звісно, призводять до виникнення, зокрема, суфіксальних морфів-омографів. У післякореневій частині простих українських слів спостерігаємо два типи перетворення форми залежно від омографічності/неомографічності мор-

фа-домінанти у твірному слові: 1) усунення омографічності аломорфа у похідному слові та, навпаки, 2) набуття аломорфом омографічності у похідному слові. Наприклад, суфікс **-ад-** у слові *попадя* омографічний до суфіксів з такою ж буквеною структурою, але відмінними категоріально-розрядними значеннями у словах *лимонад, рафінад, аркада, бравада, декада, лампада*. Натомість його аломорф **-адь-** у похідному слові *попадька* стає неомографічним. Діаметрально протилежний за своїм результатом процес спостерігаємо в парах слів *пупј-анок-пупј-аноч-ок*. Суфіксальний морф-домінанта **-анок-** неомографічний, проте його аломорф **-аноч-** у похідному слові стає омографічним до морфів з іншими значеннями у словах *сливј-аноч-ка, вуш-аноч-ка, корej-аноч*-ка. Для омографічних морфів укладено зведений індекс властивих їм 50 категоріально-розрядних значень різного ступеня узагальнення – від максимально граматикалізованих, властивих цілим частинам мови або – в їхніх межах – граматичним чи словотвірним категоріям, до максимально лексикалізованих, притаманних ліченим словам на зразок значення "різновид звуків" морфа **-ант-** у чотирьох словах-лінгвістичних термінах ***вібр-а́нт, сон-а́нт, спір-а́нт*** та ***кон-сон-а́нт.***

Стаття префіксальної частини словника може містити кілька підрозділів, або підстатей. Перший (**1.0**) показує сполучення префікса безпосередньо з коренем (**R**), другий (**1.1**) – його ліво-, а третій (**1.2**) – правосторонніх партнерів у слові, тобто фіксує сполуки, в яких такий префікс стоїть перед іншим префіксом або після нього. Якщо описуваний префікс має морфи-варіанти, то вони далі подані за спадом кількісних показників їхнього вживання за такою ж схемою побудови статті (див. рис. 8):

супер-, супра-			35	0.02	1	1
1.0. - [супер]R	29	0,01	1	1	I П	супер/експлі́брис/* супер/шви́д/к+іс+н*ий
1.2 - [супер]+об	3	0,00	0	0	I	супер+об/ма́з/ува+нн*я
1.2 - [супер]+ре	1	0,00	0	0	I	супер+ре/генер/а́т+ор
2.0 - [супра]R	2	0,00	0	0	I	супра/літора́ль/*

Примітка. Риска зліва від префікса вказує на відсутність лівого партнера. У квадратні дужки взято префікс – заголовкову одиницю статті, для якого встановлені ліві та праві партнери в слові. Отже, запис - **[супер]+ре** слід читати так: заголовковий префікс **супер-** не має партнерів зліва, а позиція справа від нього зайнята префіксом **ре**-.

Рис. 8. Словникова стаття до префіксальної пари **супер-, супра-** в "Словнику афіксальних морфем української мови"

Інформацію про префіксосполуки в окремих підрозділах словникової статті розподілено за 6 зонами. Перша містить показник частино-

мовної належності слова, що має в своєму складі таку префіксосполуку, друга – саме слово-приклад, третя – показник кількості слів з цією сполукою у генеральному реєстрі слів комп'ютерного морфемно-словотвірного фонду української мови, четверта – показник (у %) питомої ваги такої групи слів у генеральному реєстрі, п'ята – показник кількості слів з описуваною префіксосполукою у "Частотному словнику сучасної української художньої прози" і, нарешті, шоста зона статті подає сумарний показник абсолютної частоти вживання в текстах слів з такою сполукою (за "Частотним словником сучасної української художньої прози"). Подані в 4–6 зонах кількісні показники дають користувачеві можливість зіставляти активність та питому вагу описуваних префіксосполук у базі даних, що моделює лексикон як розділ загальної системи мови, та в текстах сучасної української художньої прози, які становлять продукт реалізації лексичної системи мови.

Формати статей до неомографічних та омографічних морфів у суфіксальній частині описуваного словника однакові. Заголовком статті є одиниця реєстру – окремий суфіксальний морф. Якщо реєстрова одиниця омографічна, то до неї додаються порядкові номери всіх категоріальних значень, властивих її реалізаторам у словах і вміщеним в згаданому вище "Індексі значень омографічних одиниць", що передує тексту цієї частини словника. Наприклад, складений суфікс **-езн-** здатен в прикметниках виступати в значенні "ознака" (номер в Індексі 24: ***помп-е́зн*ий***), а в прикметниках та займенниках – у значенні "підсилена ознака" (номер в Індексі 27: ***отак-е́зн*ий, важ-е́зн*ий, величе́зн*ий***). Реєстрова одиниця статті виглядає так: **езн (24, 27).** Відсутність цифр коло реєстрової одиниці засвідчує її неомографічність.

Пояснювальна частина статті суфіксальної частини словника поділена на 7 зон. Її стрижень (об'єкт опису та пояснення) становить друга зона. В ній подано всі ті оточення, в яких реєстровий суфіксальний морф трапився в словах бази даних. Таке оточення подано у вигляді тріади, в якій центральне місце займає аналізована одиниця реєстру (її позначено символом **#** у квадратних дужках), зліва від неї розташовується її лівий партнер, а справа – правий партнер або вказівка на відсутність його формального вираження у такій слівній структурі, наприклад, нульова флексія або абсолютний кінець слова у незмінюваних словах. Першу зону статті формують порядкові номери таких різних внутрішньослівних оточень суфіксального морфа. Морфемосполуки теж мають свій принцип упорядкування. Вихідною при впорядкуванні обрано мінімальну за номером післякореневу позицію суфіксальної одиниці в слові. Точкою відліку для впорядкування її лівих партнерів є символ **R**, що замінює конкретний корінь слова. Для впорядкування правих партнерів вихідною є позиція абсолютного кінця слова або позиція, в якій суфіксальний морф сполучається з нульовою,

формально не вираженою флексією (символ "**-**"). Далі партнери зліва та справа впорядковано за алфавітом.

Третя зона пояснювальної частини містить показники частиномовної належності слів, в яких реалізована описувана морфемосполука. Для кожного оточення суфіксальної одиниці наведено весь спектр його частиномовного функціонування, напр.: оточення **R-іш-н** властиве прикметникам (***давн-і́ш-н-ій***) та іменникам (***гор-іш-н-я́н-ин***) із суфіксом **-іш**-, а оточення **R-авл-ува** – дієсловам (***по-жв-а́вл-ува-ти***) та іменникам (***по-жв-а́вл-ува-нн-я***), що містять у своєму складі суфікс **-авл-**. Подання внутрішньослівних оточень суфіксальних морфів дає змогу користувачеві словника: 1) встановлювати позиційний розподіл в словах суфіксальних морфів з різними функціями: деривaційною для суфіксів, кваліфікативною, або класифікувальною для суфіксоїдів у непохідних словах з подільними основами на зразок **–ерo** зі значенням "особа" у словах ***ранч-е́ро*** (пор. ***ра́нчо***), ***мачет-е́ро*** (пор. ***мачете***), ***романс-е́ро*** (пор. ***рома́нс***), конструктивною, суто структурною для суфіксальних зв'язок, подібних до **-абель-** у словах ***комфорт-а́бель-н*ий, транспорт-а́бель-н*ий, чит-а́бель-н*ий***; 2) вивчати якісний склад і кількість партнерів суфіксальних морфів у слові. Найбільші комбінаторні властивості, як засвідчує матеріал суфіксальної частини Словника, виявляють суфіксальні зв'язки – формативи. Приміром, для форматива **-а-** виявлено 153 різних внутрішньослівних оточення.

У четвертій зоні розташовані символьні моделі морфемної структури тих слів, у яких засвідчено описувану морфемосполуку. Заголовкова суфіксальна одиниця статті береться в таких моделях у квадратні дужки. Це дозволяє користувачеві встановлювати інтервал її позиційного розподілу в словах мови, вивчати функціонування такої одиниці в різнотипних морфемних структурах. Наприклад, оточення **у-j-уч** трапляється в складі дієприслівників у морфемних структурах слів, описуваних моделями **PRSS[S]SS *об-мізк-о́в-у-j-уч-и*** або **PRS[S]SS *по-сві́ст-у-j-уч-и***, а оточення **R-овит-ий** – в прикметниках з морфемними структурами, описуваними моделями **R[S]F *мізк-ові́т-ий*** та **PR[S]F *не-горд-ові́т-ий***.

П'ята зона подає поділене на морфеми слово-ілюстрацію зі знаками кодування морфемних класів, прийнятими в записах бази даних морфемно-словотвірного фонду української мови.

62. АТИН

.001	/R/[#]а	I	RSF	/голуб'j/а́тин*а	8	0.0057
.002	ч[#]а	I	RS[S]F	/зай/ч+а́тин*а	3	0.0022
				Разом:	11	0.0080

Рис. 9. Формат статті до неомографічного суфікса **–атин-** зі значенням "різновид м'яса" у "Словнику афіксальних морфем української мови"

У шостій та сьомій зонах вміщено відомості про кількісні характе-

ристики морфемосполук, а саме: про кількість слів з ними в базі даних та їхню питому вагу в ній (в %). Питома вага визначається як відношення слів з описуваною морфемосполукою до загальної кількості опрацьованих слів бази даних. Показник питомої ваги одного слова, тобто мінімальний показник питомої ваги, становить 0,0007%. Його визначено з точністю до четвертої цифри після коми; при підсумовуванні цифри округляються з точністю до третього знаку після коми. У кінці пояснювальної частини подані зведені показники реалізації суфіксального морфа в словах лексикону та питомої ваги в лексиконі слів, що його містять. Ось, наприклад, як виглядає стаття до неомографічного суфікса **-атин**- (див. рис.9).

У пояснювальній частині статей до омографічних суфіксальних морфів, крім цього, подано інформацію про ті значення, які вони можуть виражати в певних внутрішньослівних оточеннях. Їх засвідчують порядкові номери таких значень в Індексі, додані до моделі відповідного внутрішньослівного оточення. У деяких випадках важко відокремити значення, властиві омографічній одиниці. Коли така сукупність нечітко розмежованих, синкретичних, значень регулярно повторювалася в серіях слів, до такого суфіксального морфа додавалися суми порядкових номерів таких значень, напр.: **-ат**- **(17+34+48**). Цей запис слід тлумачити так: суфікс **-ат**- може виражати в слові одночасно значення "сукупність, збірність", "місце" та "установа, форма управління", пор.: ***декан-а́т, каган-а́т, консул-а́т, старост-а́т*** тощо. Формат статті до омографічного суфіксального морфа **-ав-** подано на рис. 10:

3. АВ (24, 26, 46)

001	/R/ [#] - (24, 46)	І	R[S]F	/рук/а́в* (46)	1	0.0007
		П	R[S]F	/ласк/а́в* (24)	1	0.0007
003	/R/[#]-еньк (24,26)	П	R[S]SF	/ми́рш/ав+еньк*ий (24) /молож/а́в+еньк*ий (26)	8	0.0057

Рис. 10. Формат статті до омографічного суфіксального морфа **–ав-** зі значеннями "предмет", "ознака", "послаблена ознака" у "Словнику афіксальних морфем української мови"

Описаний комп'ютерний "Словник афіксальних морфем української мови" містить нові типи інформації про морфемну будову сучасного українського слова – формальні, змістові, функціональні, кількісні, частотні, а також нові способи їхньої організації та представлення. Комп'ютерна форма словника з чітким структуруванням словникових статей за допомогою застосованого зонного принципу запису мовної інформації уможливлює роботу з ним не лише як з готовим продуктом опрацювання бази даних, а й як з інформаційно-довідковою, навчальною та дослідницькою системою. Таким чином, сконструйований словник у своїй комп'ютерній формі становить базу даних-сателіт, яка хоча й пов'язана з основною базою даних, проте містить і важливу додатко-

ву інформацію про морфемний склад слова в цілому, про властивості окремих морфів у ньому та про закономірності будови й функціонування інвентаря афіксальних морфем сучасної української мови. Він може становити підґрунтя для укладання інших морфемних словників, зокрема словників морфем-синонімів або антонімів. Таких, наприклад, як суфікси на позначення особи чи місця або посилення чи послаблення вияву ознаки.

Інший напрямок досліджень з комп'ютерної лексикографії становить конструювання словників з новими лінгвістичними об'єктами як одиницями реєстру, або одиницями опису й пояснення в таких словниках. Прикладом такого типу комп'ютерних словників є створювані на основі одномовних тлумачних словників семантичні, або ідеографічні словники. Першою спробою укладення такого словника за допомогою комп'ютера став опублікований 1982 р. за редакцією Ю.М.Караулова "Русский семантический словарь: Опыт автоматического построения тезауруса: от понятия к слову"[38]. Виходу цього словника передувала багаторічна робота колективу дослідників Інституту російської мови АН СРСР, очолюваного Ю.М.Карауловим, з вивчення структури метамови тлумачних словників, вироблення принципів семантичної класифікації лексики на основі формальних ознак опису тих чи інших понять у дефініціях одномовних тлумачних словників російської мови. Теоретичне узагальнення досвіду цієї роботи викладено в низці праць Ю.М.Караулова та його колег[39]. Семантичний словник автори розглядали як різновид ***тезаурусів***, одномовних словників ідеографічного, або ідеологічного типу, в яких слова впорядковано на основі спільності виражених ними понять. Стрижень ідеографічних словників становить так звана ***синоптична***, або ***зведена схема понять*** – основа семантичної класифікації лексики. Укладати такі словники почали ще в другій половині XIX століття. Першою ластівкою серед лексикографічних праць ідеографічного типу став "Тезаурус англійських слів та висловів" П.М.Роже ("Roget's thesaurus of English words and phrases"), опублікований у 1852 р. На момент створення семантичного словника Ю.М.Караулова та його колективу, крім тезаурусу П.М.Роже, вже існували

[38] **Русский** семантический словарь: Опыт автоматического построения тезауруса: от понятия к слову. - М., 1982.

[39] Зацікавлених читачів відсилаємо до монографій: **Караулов Ю.Н.** Общая и русская идеография. – М., 1976.; **Караулов Ю.Н.** Частотный словарь семантических множителей русского языка. – М., 1980.; **Караулов Ю.Н.** Лингвистическое конструирование и тезаурус литературного языка. - М., 1981.; **Анализ** метаязыка словаря с помощью ЭВМ. – Ю.Н.Караулов, В.И.Молчанов, В.А.Афанасьев, Н.В.Михалев. – М., 1982. Див. також докладний і критичний аналіз «Русского семантического словаря» у: **Клименко Н.Ф.** Построение тезауруса с помощью ЭВМ // Украинский семантический словарь. Проспект. – К., 1990. – С.81-89.

укладені вручну ідеографічні словники для ряду європейських мов, зокрема, словники Ф.Дорнзайфа та Р.Халліга й В.фон Вартбурга – для німецької, Х.Касареса – для іспанської, П.Буасьєра – для французької мови. Після опублікування комп'ютерного "Русского семантического словаря" світ побачила низка ідеографічних словників російської мови, укладених вручну з використанням різних методик семантичного аналізу лексики та її системної організації, а саме: словник "Лексическая основа русского языка" за ред. В.В.Морковкіна (М., 1984), "Идеографический словарь глаголов русского языка" О.С.Баранова (М., 1995), "Русский семантический словарь. Толковый словарь, систематизированный по классам слов и значений. Т.1.: Слова указующие (местоимения). Слова именующие: имена существительные (Все живое. Земля. Космос)" за ред. Н.Ю.Шведової (М., 1998) та "Толковый словарь русских глаголов: Идеографическое описание. Английские эквиваленты. Синонимы. Антонимы" за ред. Л.Г.Бабенко (М., 1999).

У лексикографії, як відомо, існують два визначення терміна ***тезаурус*** (від лат. **thēsaurus** "скарб, скарбниця"). Перше, традиційне, вже подано вище. Друге сформувалося у межах комп'ютерної лінгвістики і трактує тезаурус як різновид ідеографічного словника для потреб систем автоматичного пошуку інформації. Різновидом саме такої системи стала база даних для укладання "Русского семантического словаря". Підкреслюючи відмінність свого словника від традиційних, укладених вручну ідеографічних словників, упорядники послугуються і термінологічним апаратом інформатики. Так, реєстрову одиницю "Русского семантического словаря" становить ***дескриптор*** (від лат. **dēscrīptor** "описувач" з **dē-scrībo** "описувати") – слово – виразник спільного поняття для певного об'єднання лексем, що, за словами Ю.М.Караулова, складають його лексичне оточення, своєрідний "семантичний ореол" дескриптора в словнику[40]. В системах автоматичного інформаційного пошуку дескрипторами прийнято називати стандартизовані слова або словосполуки, що виконують роль образів, ключів для індексування змісту документів й подальшого автоматичного пошуку в них потрібної інформації (про це докладніше див. далі у §5 розділу 2). Усі слова до такого об'єднання зібрані за ознакою наявності в їхніх статтях у тлумачних словниках С.І.Ожегова та Д.М.Ушакова, які стали фактичною базою для укладання «Русского семантического словаря», семантичних множників, властивих тлумаченню слова-дескриптора. Упорядковані вони за показниками кількості та питомої ваги в їхніх тлумаченнях таких множників. ***Семантичний множник***, у свою чергу, визначений як мінімальний словниковий реалізатор компонента семантичної структури слова, представлений як основа слова – складника дефініції, спільна для ряду слів з тим же коренем.

[40] **Русский** семантический словарь. – С.3.

Отже, новим лінгвістичним об'єктом виступає як саме подібне об'єднання семантично пов'язаних слів, так і інструменти конструювання цього об'єднання – дескриптори та семантичні множники. Стаття обговорюваного словника має такий формат. Заголовок статті складають дескриптор з його порядковим номером у реєстрі словника. До дескриптора подано всі семантичні множники – реалізатори компонентів його семантичної структури в словниковій дефініції з показником кількості слів, які такий семантичний множник об'єднує в статті. Саме наповнення статті "Русского семантического словаря" становлять усі слова опрацьованої бази даних, які містять у своїх словникових дефініціях хоча б один із семантичних множників дескриптора. Наприклад, дескриптору ***детство*** властиві 14 семантичних множників різної системотвірної потужності, тобто вони здатні в базі даних об'єднувати навколо дескриптора різну кількість слів. Завдяки цим множникам комп'ютер зібрав навколо вказаного дескриптора 34 слова бази даних (див. табл. 6):

Таблиця 6. Фрагмент статті з дескриптором ***детство*** в "Русском семантическом словаре" за ред. Ю.М.Караулова

275. детство
Семантические множители: 1. возраст – 1, 2. незре – 1, 3. дит – 2, 4. малоле – 3, 5. невзро – 3, 6. неразум – 3, 7. мальч – 4, 8. наив – 5, 9. реб – 5, 10. дев – 6, 11. младе – 7, 12. год – 11, 13. мал – 13, 14. лет – 16
Число слов, включенных в статью дескриптора – 34
возрастать 4*1
дети 11*3,7,9,10,13,14
женщина 11*10, 14
молодой 13*2
невеста 7*10
парнишка 2*7
подросток 7*7,10,14
прогрессивный 6*1
расти 20*12,14
роды 7*11,14
усыновить 9*7,10
ясли 13*13,14
девочка 4*9,10
детский 3*2,5,14
маленький 7*4,13

Примітка. Цифра коло слова до зірочки вказує на кількість його семантичних множників; цифра після зірочки зазначає кількість семантичних множників слова, які збігаються з семантичними множниками дескриптора

Деякі з включених до статті цього дескриптора слова нас дивують, наприклад, ***невеста*** або ***прогрессивный***. Проте після звернення до словникових дефініцій цих слів стає зрозумілою їхня поява у цій слов-

никовій статті. Зокрема, словник С.І.Ожегова подає таке визначення слова ***детство*** – "Детский возраст, детские годы", що дає можливість виділити семантичні множники **дет, возраст** та **год**. У дефініції слова ***невеста*** присутній спільний з цим словом семантичний множник **возраст**, пор. "Девушка или женщина, вступающая в брак, а также (разг.) девушка, достигшая возраста, при к-ром можно вступать в брак". Появу слова ***прогрессивный*** також пояснює наявність в його дефініції семантичного множника **возраст**, основи спільної для цілого ряду слів – компонентів словникових тлумачень: ***возраст, возраст**ать, **возраст**ающий, **возраст**ной*. Див. функціонування цього множника в дефініції слова ***прогрессивный***: "1. Постепенно усиливающийся, возрастающий". Кількість семантичних множників кожного зі слів, що потрапили в статтю дескриптора, спільних з множниками самого дескриптора – виразника описуваного поняття, дає змогу структурувати статтю за ступенем зв'язаності слів з дескриптором, а отже, за їхньою питомою вагою в мовній реалізації цього поняття. Найменш потужними виявляться ті слова, в яких лише один з множників виявився спільним, наприклад, ***прогрессивный, возрастать, парнишка.*** Найпотужнішим виявилося слово ***дети***, яке має 6 спільних з дескриптором семантичних множників: **дит, неразум, реб, дев, мал, лет** (див. дефініцію слова ***дети*** в словнику С.І.Ожегова: 1. Мальчики и (или) девочки в раннем возрасте, до отрочества (употр. в знач. мн. к "ребенок" и "дитя"). В численних рецензіях на "Русский семантический словарь"[41] серед інших критичних зауважень якраз і вказувалося на невизначеність ієрархії семантичних множників, їхньої обов'язковості/факультативності при мовній реалізації певного поняття, їхнього функціонального навантаження в словникових дефініціях, хоча для надання тому чи іншому складнику словникової дефініції статусу семантичного множника й враховувалися певні його статистичні характеристики, наприклад, таке слово в дефініціях лексем бази даних мало зустрітися не менше 7 разів. Надання семантичним множникам рівноправного статусу в процесі семантичної організації лексики, як ми пересвідчилися на прикладі складу статті з дескриптором ***детство***, призводить до включення в таку статтю поняттєво віддалених, а то й просто випадкових слів, дефініції яких мають семантичні множники-омографи на зразок **возраст** (возраста) і **возраст** (возрастающий).

Оригінальну методику семантичної класифікації лексики на основі компонентного аналізу словникових дефініцій та організації семантич-

[41] **Дерягин В.** Учить ли ЭВМ писать с ошибками? // Правда. – 1983. – 23 сент.; **Пещак М.М.** [Рецензія] // Мовознавство. – 1984. - № 2. – С.74-75. – Рец. на кн.: Русский семантический словарь; **Шайкевич А.Я.** Об автоматическом построении тезауруса на основе толковых словарей // Науч.-техн. информация. - Сер.2. – 1985. - № 4. – С.12-19.

но пов'язаних слів за ступенем узагальнення єдиного для всіх них поняття запропонувала російська дослідниця Е.В.Кузнецова. Її методика аналізу словникових дефініцій ґрунтувалася на принципово відмінній від методики Ю.М.Караулова основі. Власне компонентному аналізу дефініції передував її логічний аналіз, виділення в її складі інтегральної частини, таких ознак поняття, які пов'язують цю лексему з іншими, і частини диференційної, ознак, що відрізняють значення описуваного слова від значень інших слів, які виражають те саме поняття, конкретизують його в семантиці описуваного слова. Такий підхід дозволив уникнути при формуванні понятєвих груп випадкових лексем, пов'язаних з цими групами нерелевантними для вираження спільного поняття компонентами дефініції. Основу для формування подібних понятєвих груп закладали спільні реалізатори інтегральної частини дефініцій. Розроблена Е.В.Кузнецовою методика здобула назву ***процедури ступінчастої ідентифікації лексики***[42]. Принцип її полягає в тому, що в дефініціях слів (об'єктом аналізу Е.В.Кузнецової та очолюваного нею колективу співробітників кафедри російської мови Уральського університету (м.Свердловськ, нині – м.Катеринбург у Росії) були дієслова сучасної російської мови) виділялися реалізатори поняття, спільного для певної лексико-граматичної групи слів, та реалізатори понять, якими слова в такій групі відрізнялися одне від одного. Перші, які виражали інтегральні семантичні ознаки, було названо ***словниковими ідентифікаторами***, другі, що виражали ознаки диференційні, дистинктивні, – ***словниковими конкретизаторами***. За ступенем узагальнення спільного поняття ідентифікатори, у свою чергу, було поділено на ***родові*** та ***видові***. Результатом семантичної класифікації дієслівної лексики за методикою ступінчастої ідентифікації став «Толковый словарь русских глаголов: Идеографическое описание. Английские эквиваленты. Синонимы. Антонимы», який після передчасної смерті Е.В.Кузнецової завершили її учні[43]. На відміну від "Русского семантического словаря" цей словник укладено без використання комп'ютера, проте методика ступінчастої ідентифікації лексики завдяки її чіткій структурованості та зорієнтуванню на вивчення реалізаторів понять у словникових дефініціях, а отже, на їхню метамовну, зовнішню, формалізацію виявилася ефективною й придатною для процедури конструювання комп'ютерних словників ідеографічного типу.

На матеріалі сучасної української мови це переконливо продемонструвала Н.В.Сніжко, уклавши за допомогою цієї методики ідеографіч-

[42] **Кузнецова Э.В.** Ступенчатая идентификация как средство описания семантических связей слов // Вопросы металингвистики. - Ленинград, 1973. - С.84-94; **Кузнецова Э.В.** Лексикология русского языка. - М., 1982.

[43] **Толковый** словарь русских глаголов: Идеографическое описание. Английские эквиваленты. Синонимы. Антонимы / Под ред. Л.Г.Бабенко. – М., 1999.

ний тезаурус українських іменників[44]. Матеріалом її аналізу стали дефініції 54 тис. слів в 11-томному тлумачному "Словнику української мови". На основі ідентифікаторів (родових та видових), виявлених у їхніх словникових формулах тлумачення, сконструйовано 101 поняттєву групу[45], серед яких назви людей, опредметнених дій, предметів, місць, знарядь дії, тварин, властивостей, станів тощо. Наявність видових ідентифікаторів дає можливість структурувати групи на підгрупи. Наприклад, до складу поняттєвої підгрупи "бур'ян" у складі поняттєвої групи "рослина" ввійшли 47 слів, що в своїх дефініціях містять цей видовий ідентифікатор, пор.: ***ба́бка, вівсю́г, жовти́ло, лобода́, осо́т, реп'я́х*** та інші. Всього у складі поняттєвої групи "рослина" виділено 35 поняттєвих підгруп з видовими ідентифікаторами різного ступеня узагальнення такого спільного поняття, а саме: ***дерево, кущ, трава, чагарник, водорість, злак, яблуня, ялина, льон*** тощо.

4-ий рівень ідентифікації	***росли́на*** – 1. Організм (зона **РСІ**), який живиться неорганічними речовинами повітря й ґрунту (зона **СК**), є однією з форм існування живої матерії на Землі і разом з тваринними організмами належить до живої природи[46].
3-ій рівень ідентифікації	***де́рево*** – 1. Багаторічна (зона **СК**) рослина (зона **ВСІ**), з твердим стовбуром і гіллям, що утворює крону (зона **СК**).
2-ий рівень ідентифікації	***я́блуня*** – Садове й лісове фруктове (зона **СК**) дерево (зона **ВСІ**) родини розових з плодами перев. кулястої форми (зона **СК**).
1-ий рівень ідентифікації	***анто́нівка*** – 1. Зимостійкий сорт (зона **СК**) яблуні (зона **ВСІ**). ***апо́рт*** – 1. Один з кращих осінньо-зимових сортів (зона **СК**) яблунь (зона **ВСІ**). ***кальві́ль*** - Зимовий сорт (зона **СК**) яблуні (зона **ВСІ**). ***ране́т, рене́т*** – 1. Південний високоцінний сорт (зона **СК**) яблунь (зона **ВСІ**).

[44] **Сніжко М.Д., Сніжко Н.В.** Автоматизована система класифікації лінгвістичних об'єктів // Проблеми українізації комп'ютерів. – К., 1993. – С.69-72; **Сніжко Н.В., Сніжко М.Д.** "Ідеографічний тезаурус" як інформаційно-довідкова система при вивченні закономірностей структурно-функціональної організації лексики // Мовознавство. - 1996. - № 4-5. – С.23-28.

[45] Повний перелік сконструйованих поняттєвих груп іменників див. у: **Сніжко Н.В.** Ідеографічний тезаурус як модель лексико-семантичної системи (за наслідками автоматизованого аналізу українських іменників) // Мовознавство. – 1995. - № 6. – С.28-35.

[46] Дефініція слова ***росли́на*** містить також компоненти, що вказують на її зв'язок з родовими ідентифікаторами вищих рівнів – 5-го, 6-го та 7-го – в синоптичній схемі ідеографічного словника, а саме: ***організм*** (5-ий рівень)***, жива природа*** (6-ий рівень) та ***жива матерія*** (7-ий рівень – найвищий для слів цієї семантики в аналізованому словнику - СУМі). Зони в дефініції, що містять такі ідентифікатори вищих рівнів, можна розглядати як додаткові відсильні.

симире́нка – 1. Зимовий сорт (зона **СК**) яблуні (зона **ВСІ**).
цига́нка[2] – Зимовий сорт (зона **СК**) яблуні (зона **ВСІ**), що має соковиті плоди з темно-червоною шкіркою (зона **СК**)

Примітка. Абревіатури **РСІ, ВСІ** та **СК** позначають у дефініціях, відповідно, родові та видові словникові ідентифікатори й словникові конкретизатори. У свою чергу, словникові конкретизатори також можуть мати різний ступінь узагальнення й утворювати певну ієрархію, пор. відношення між конкретизаторами ідентифікатора ***дерево*** *садове, лісове, фруктове, родини розових, з плодами, кулястої форми.*

Принцип виділення у дефініціях словникових ідентифікаторів різного ступеня та різноманітних їхніх конкретизаторів, а отже, місця такої лексеми в ієрархії семантично пов'язаних слів продемонструємо на прикладі іменників ***росли́на, де́рево, я́блуня, анто́нівка, апо́рт, кальві́ль, ране́т, рене́т, симире́нка, цига́нка***[2].

Ретельний і всебічний аналіз статей до іменників у тлумачному словнику і підготовка їхньої відповідної комп'ютерної версії з гнучкою граматикою аналізу та синтезу уможливила конструювання семантично пов'язаних слів за всім спектром ознак їхнього опису і взагалі багатоаспектну роботу з такою базою даних як з дослідницькою та інформаційно-довідковою системою. Так, на вимогу користувача можна не лише побудувати певну поняттєву групу або підгрупу, а й сконструювати в їхніх межах синонімічні ряди з властивими їм домінантами, антонімічні пари, групи слів з певним типом розвитку вихідного значення (наприклад, "тварина→людина", "людина→предмет"), дібрати слова з певним типом морфемної або словотвірної будови, з певними стилістичними або граматичними характеристиками. Наприклад, у межах поняттєвої підгрупи "бур'ян" сконструйовані такі синонімічні об'єднання слів за тими чи іншими властивостями цього виду рослин: ***бере́зка, пові́й, пові́йка; вівсю́г, ві́всик; липу́чка, липни́к; лобода́, на́тина; миші́й, бри́ця; осо́т, осе́т, храбу́ст; різа́к, тілорі́з; тата́рник, чортополо́х***. За спільним типом семантичної структури, що об'єднує значення "місце" і "напій" згрупувалися іменники ***вишня́к, казе́нка, полуни́чник, сливни́к, сливня́к, спотика́ч, травни́к***. За стилістичною ремаркою *заст.* (застаріле) у поняттєвій групі назв людей об'єдналися найменування ***брова́р, брова́рник, бу́дник, бурла́к, валю́шник, гура́льник*** та ін.; серед назв хусток ремаркою *діал.* (діалектне) позначено іменники ***плат, плати́на***; як розмовні за допомогою ремарки *розм.* кваліфіковано назви яблунь ***квасни́ця, рене́та, цига́нка***[47]. Таким чином, укладений ідеографічний тезаурус іменників сучасної укра-

[47] Усі приклади взято з праці: **Сніжко Н.В., Сніжко М.Д.** "Ідеографічний тезаурус" як інформаційно-довідкова система при вивченні закономірностей структурно-функціональної організації лексики // Мовознавство.-1996. - № 4-5. – С.27.

їнської мови може служити базою для укладення інших типів словників: синонімічного, антонімічного, лексико-граматичних груп слів тощо.

А.Я.Середницька, творчо опрацювавши методику ступінчастої ідентифікації лексики Е.В.Кузнецової і доповнивши її власними оригінальними процедурами конструювання синоптичної (зведеної) схеми понять, уклала ідеографічний словник українських дієслів переміщення, реєстр якого становлять 54 концепти, що пов'язані із загальним поняттям "рух" і описують 5500 дієслів цієї тематичної групи, вміщені в 11-томному тлумачному "Словнику української мови"[48]. Сконструйована А.Я.Середницькою синоптична схема – інструмент розподілу дієслівної лексики в цьому ідеографічному словнику, містить концепти трьох рівнів: концептом 1-го рівня є найзагальніше поняття "рух"; на 2-ому рівні узагальнення його деталізують концепти "зміна розташування предмета у просторі", "зміна стану", "напрям переміщення", "спосіб переміщення" та "траєкторія руху". На 3-ому рівні перебувають концепти, що й становлять реєстрові одиниці словника. Так, наприклад, концепт "спосіб переміщення" конкретизують 29 концептів, серед яких "переміщати волоком", "переміщати у рідині", переміщатися стрибками", "переміщатися, переслідуючи" та інші. У свою чергу, за певними додатковими семантичними ознаками у межах статей словника можуть формуватися синонімічні ряди або антонімічні пари. Наприклад, у статті з реєстровою одиницею-концептом **Переміщатися за допомогою ніг** за ознакою "іти повільно" в синонімічний ряд вишикувалися дієслова ***брести́ 1, волокти́ся 2, ди́бати, дибуля́ти, кли́гати, лі́зти 2, пле́нтатися, плести́ся 1, повзти́ 2//, посува́ти 2*** тощо[49]. Вироблені процедури максимально формалізовані, проте самий процес укладання словника здійснено вручну. У комп'ютерному варіанті подано уже готовий продукт опрацювання мовного матеріалу. Запропонована методика аналізу дієслівних дефініцій дозволяє досить легко перетворити текст словника на базу даних, придатну для виконання різноманітних інформаційно-пошукових та дослідницьких процедур.

Підкреслимо ще раз, що перевага комп'ютерного словника перед некомп'ютерним, традиційним, полягає в тому, що 1) його укладення передбачає побудову відповідної бази даних та розроблення спеціального лексикографічного процесора; 2) сконструйований словник може, в свою чергу ставати базою для укладення нових комп'ютерних словників, а також разом з його базою даних виконувати функції авто-

[48] **Середницька А.Я.** Ідеографічний поділ дієслівної лексики в сучасній українській мові. – Автореф.дис. ... канд.філол.наук. – К., 2001.

[49] Цифра після слова позначає певне його значення, цифра з двома скісними – відтінок значення, оскільки об'єктом опису в ідеографічному словнику дієслів переміщення А.Я.Середницької є лексико-семантичний варіант слова.

матичної інформаційно-довідкової, навчальної, редакційно-видавничої та дослідницької системи.

Таку перевагу яскраво демонструє лексикографічний процесор СЛОВО, створений львівськими дослідниками О.М.Коссаком та С.Л.Маньковським для конструювання комп'ютерних термінологічних одно- і багатомовних словників. В основу стратегії створення цього процесора покладено досвід укладання "Англо-українсько-російського словника з інформатики та обчислювальної техніки". Словникова термінологічна база даних має гніздовий принцип організації. Вершиною гнізда є слово певної мови (у повідомленнях про дотеперішні версії бази – це українська, російська, англійська та німецька мови), до такого базового слова подані відповідні переклади. Якщо таке родове поняття-термін має кілька видових термінів-конкретизаторів, вони також вміщуються в гнізді з відповідними перекладами, причому спільний для всіх них родовий термін позначений символом **~**. Кожен вихідний термін має порядковий номер зліва, українські переклади також мають справа від себе порядкові номери-відсилання до додатку словника – вказівника російських термінів-еквівалентів[50]. Лексикографічний процесор становить інтерфейс, що дає можливість користувачеві працювати з базою даних в інтерактивному режимі, а отже, контролювати всі етапи побудови словникових статей. Так, наприклад, якщо в існуючій базі даних не вміщено ту чи іншу інформацію про наявні слова або інформація про якийсь термін взагалі відсутня, користувач дістає змогу ввести потрібні відомості до бази даних, а то й створити їх за розробленим форматом статті нового термінологічного словника. Лексикографічний процесор СЛОВО, крім процедур коригування бази даних та конструювання термінологічних словників, забезпечує також роботу з базою даних як з інформаційно-пошуковою, навчальною та дослідницькою системою. Наприклад, завдяки спеціальним алгоритмам пошук можна здійснювати за такими ***ключами***, або ***образами для пошуку об'єктів***: самим словом певної мови, його номером у базі даних, першою літерою або його іншомовними еквівалентами. В цілому лексикографічний процесор СЛОВО здатен виконувати такі функції: 1) запис термінів до відповідного словника (за мовою представлення); 2) коригування словників; 3) пошук слів та їхніх перекладів у базі за заданими ключами та 4) вилучення слів та/або їхніх перекладів зі словників. Формат статей конструйованих термінологічних словників подібний до формату, створеного для укладеного розробниками процесора СЛОВО "Англо-українсько-російського словника з інформатики та обчислювальної техніки". Ось, наприклад, яку інформацію містить у цьому словнику гніздо з базовим словом ***database*** "база даних":

[50] **Коссак О.М., Маньковський С.Л.** Англо-українсько-російський словник з інформатики та обчислювальної техніки. – Л., 1991.

D003.	.00.0	**database** база даних	б001.001
	.01.0	**comprehensive ~** база даних широкого користування	б001.007
	.02.0	**design ~** база даних проектування	б001.005
	.03.0	**distributed ~** розподілена база даних	б001.017
	.04.0	**enterprise ~** база даних підприємства	б001.004
	.05.0	**generalized ~** база даних загального призначення	б001.003
	.06.0	**hierarchical ~** ієрархічна база даних	б001.010
	.07.0	**integrated ~** інтегрована база даних	б001.011
	.08.0	**intelligent ~** інтелектуальна база даних	б001.012
	.09.0	**loaded ~** заповнена база даних	б001.009
	.10.0	**logical ~** логічна база даних	б001.015
	.11.0	**network ~** база даних мережі	б001.019
	.12.0	**on-line ~** інтерактивна база даних	б001.013
	.13.0	**personal ~** особиста база даних	б001.014
	.14.0	**physical ~** фізична база даних	б001.021
	.15.0	**populated ~** заповнена база даних	б001.009
	.16.0	**private ~** приватна база даних	б001.022
	.17.0	**public ~** спільна база даних	б001.016
	.18.0	**random access ~** база даних з прямим доступом	б001.006
	.19.0	**relational ~** реляційна база даних	б001.018
	.20.0	**shareable ~** база даних колективного користування	б001.002

Останнім часом у зв'язку з реалізацією загальнодержавної програми лексикографічного опрацювання української мови "Словники України" взято курс на координацію зусиль розробників словникових баз даних та лексикографічних процесорів, що працюють у різних наукових та науково-дослідних закладах України. Програма передбачає створення потужних інтегральних словникових баз з гнучкими і розгалуженими лексикографічними процесорами, що забезпечують процедури конструювання комп'ютерних словників різних типів, а також уможливлюють користувачам доступ до таких словникових систем з метою одержання довідкової інформації та виконання різноманітних теоретичних і практичних дослідницьких завдань. Цю роботу координує Науково-координаційна рада "Інформація. Мова. Інтелект" при Президії НАН України, очолювана віце-президентом НАН України І.Ф.Курасом. Разом з математиками-програмістами, фахівцями в галузі інформаційних технологій та розроблення систем штучного інтелекту у діяльності Ради найактивнішу участь беруть лінгвісти – працівники Інституту мовознавства ім.О.О.Потебні НАН України, Інституту української мови НАН України та Українського мовно-інформаційного фонду НАН України. Розроблена цією Радою науково-дослідна програма ставить на меті розгортання масштабних досліджень, спрямованих на

створення сучасних комп'ютерних інтелектуальних систем, в яких одне зі стрижневих місць посідає розроблення засобів для автоматичного опрацювання мовної інформації.

Так, до 10-річчя незалежності України колектив лінгвістів та математиків-програмістів Українського мовно-інформаційного фонду у складі В.А.Широкова, І.В.Шевченка, О.Г.Рабульця, О.М.Костишина та М.М.Пещак відповідно до Указу Президента України від 7 серпня 1999 року "Про розвиток національної словникової бази" створив комп'ютерну «Інтегровану лексикографічну систему». На компакт-диску (CD – англ. ***compact-disk***) вміщено заабеткований реєстр українських слів, який налічує понад 150 тис. різних лексем. Системні одиниці реєстру (канонічні, або вихідні форми змінюваних слів) розгортаються у 3 млн. текстових слововживань (форм слів у непрямих відмінках або інших форм змінюваних слів). Для роботи з реєстром можливі 5 режимів, у кожному з яких користувачеві надається інформація про певний аспект форми, семантики та функціонування слова. У режимі "Парадигма" вміщено відомості про словозмінні характеристики слів. Його роботу забезпечує відповідна лінгвістична база, яка охоплює 1500 парадигматичних класів змінюваних слів повнозначних частин мови (як загальних, так і власних назв – імен, прізвищ, по батькові, географічних найменувань), а з урахуванням акцентуаційних властивостей таких класів у базі близько 3000. Режим "Транскрипція" подає запис слова у фонетичній транскрипції. Основу його бази даних забезпечив академічний 2-томний "Орфоепічний словник української мови", укладений колективом співробітників Українського мовно-інформаційного фонду спільно з працівниками Інститутів мовознавства ім.О.О.Потебні та української мови НАН України. Перший том цього словника побачив світ 2001 року. Режими "Фразеологія", "Синонімія" та "Антонімія" подають відомості про різні властивості семантики слів реєстру, а саме: про їхню здатність створювати фразеологічні одиниці різного типу, входити до складу синонімічних пар або рядів, а також утворювати антонімічні пари зі словами, що мають спільну інтегральну сему. Лінгвістичні бази даних для роботи системи в цих режимах становлять комп'ютерні версії відповідних словників сучасної української мови: 2-томного академічного "Фразеологічного словника української мови" (К., 1-е вид. – 1993, 2-е вид. – 1999), 2-томного академічного "Словника синонімів української мови" (К., 1999–2000) та "Словника антонімів української мови" Л.М.Полюги (К., 2-е вид., 1999). Вони містять, відповідно, 56 тис. фразеологічних одиниць, 9200 синонімічних пар або рядів та 2200 компонентів антонімічних пар. Зручний і простий в користуванні інтерфейс робить "Інтегровану лексикографічну систему" доступною для будь-якого користувача, не вимагаючи від нього спеціальної програмістської

підготовки. Ось, наприклад, яку інформацію можна одержати з описуваної системи для загальної назви – іменника ***життя***:

Режим	**Інформація**		
Парадигма	***життя́* –** іменник середнього роду		
	Відмінок	Однина	Множина
	називний	***життя́***	***життя́***
	родовий	***життя́***	***житті́в***
	давальний	***життю́***	***життя́м***
	знахідний	***життя́***	***життя́***
	орудний	***життя́м***	***життя́ми***
	місцевий	на/у/по ***житті́, життю́***	на/у/по ***життя́х***
	кличний	***життя́****[51]	***життя́****
Транскрипція	***жи^{е}т':á***		
Фразеологія	***ва́жити життя́м, відда́ти життя́, життя́ не ми́ле, на життя́ і смерть, на кра́йнім поро́зі життя́, путі́вка в життя́*** (всього 76 фразеологізмів різної будови)		
Синонімія[52]	1. **ЖИТТЯ́** (стан живого організму – людини, тварини, рослини)**, ІСНУВА́ННЯ, БУТТЯ́** *уроч.,* **ДНІ** *чиї,* **ЖИВІ́Т** *заст.,* **ЖИВО́ТТЯ** *діал.* (людини) 2. **ЖИТТЯ́** (повсякденний спосіб існування когось), **ПО́БУТ, БУТТЯ́, ІСНУВА́ННЯ, ЖИТТЯ́-БУТТЯ́** *розм.,* **ЖИТУ́ХА** *фам.,* **ПРОЖИВА́ННЯ** *рідше,* **ЖИТІЄ́** *ірон. рідше,* **ПРОЖИ́ТОК** *діал.* **3. 3, 4. життя́** див. 1. **вік, дійсність**		
Антонімія[53]	**ЖИТТЯ́** (існування всього живого) – **СМЕРТЬ** (припинення життєвої діяльності організму, загибель) Пор. ще **ЖИВИ́Й- МЕ́РТВИЙ, ЖИ́ТИ-УМИРА́ТИ, НАРОДИ́ТИСЯ-УМЕ́РТИ**		

Структурування інформації про слово за типами й організації її в окремих базах даних реляційного типу уможливлює не лише подальше поповнення системи без зміни її загальної архітектури, а й роботу користувача з потрібною йому інформацією в зручному й гнучкому режимі доступу. Залежно від виконуваного завдання користувач може компонувати потрібну йому інформацію в будь-яких комбінаціях її типів, а також видобувати її з системи в потрібному йому обсязі.

---

[51] Позначка ***** вказує на форму слова, потенційно можливу, але рідко вживану в звичайному мовленні.

[52] Інформацію з цієї бази даних подано в скороченому вигляді (без ілюстрацій).

[53] Інформацію цієї бази даних також подаємо без пояснень та ілюстрацій. Зацікавлені читачі можуть одержати в повному вигляді відомості про синонімічні та антонімічні властивості слів, звернувшися до описуваної комп'ютерної системи або до відповідних словників, що послужили джерелами для формування її лінгвістичного забезпечення.

🕮 Терміни

- **комп'ютерний (=автоматичний) словник –** словник, процедури укладання якого здійснює комп'ютер
  - **формат статті комп'ютерного словника –** модель організації, розміщення та графічного представлення інформації про мовні об'єкти, описувані в словнику
- **комп'ютерна підтримка словникозорієнтованої бази даних –** інтерфейс, призначений для роботи зі словникозорієнтованою базою даних
  - **комп'ютерні засоби конструювання нових словників –** алгоритми та створені за ними програми й модулі лексикографічного процесора для конструювання словників нових типів

§6. Корпусна лінгвістика: предмет дослідження і завдання

- Стратегії створення текстозорієнтованих баз даних
- Повнотекстові бази (=корпуси текстів)
- Електронні картотеки (=ілюстративні бази даних, бази цитат)
- Продукти опрацювання текстозорієнтованих баз даних: комп'ютерні словопокажчики, конкорданси та частотні словники

У §2 цього розділу, характеризуючи різновиди лінгвістичних баз даних, ми поряд зі словникозорієнтованими виділили ***текстозорієнтовані бази даних***. Такими, нагадаємо, є бази, основою формування яких є різноманітні за тематикою, структурою, обсягом, жанром, мовою та часом створення тексти. З появою сканерів процес формування таких баз значно спростився і пришвидшився, проте сканери не розв'язали власне лінгвістичних проблем побудови таких баз даних. До них належать проблеми добору текстів та їхнього аналізу, а саме: виділення різних типів представленої в них інформації про мову та об'єкти позамовної дійсності, так звану внутрішньо- і позамовну інформацію. Отже, текстозорієнтовані бази даних, як і бази словникозорієнтовані, можуть залежно від способу організації в них інформації виступати як бази даних та як бази знань про мову та відбиту в ній "картину світу" (=інтелектуальні бази даних). Такі корпуси даних, добутих з текстів, та самі корпуси текстів становлять об'єкт вивчення корпусної лінгвістики – самостійної дисципліни в межах комп'ютерної лінгвістики.

На сьогодні сформувалися два напрямки створення текстозорієнтованих баз даних: формування ***корпусів текстів***, або ***повнотекстових баз даних*** і створення ***електронних картотек***, або ***ілюстративних баз даних***, або ***баз цитат***. Кожен з цих напрямків має свою специфіку у підході до організації мовної інформації в базах да-

них та в розробленні засобів їхнього опрацювання – текстових процесорів. Зауважимо принагідно, що ці два напрямки не заперечують, а доповнюють один одного, оскільки становлять наслідки опрацювання під різним кутом зору одного лінгвістичного об'єкту – тексту.

Структура і конкретне наповнення повнотекстових баз даних зумовлені характером тих теоретичних та практичних завдань, які така база покликана розв'язувати. Наприклад, за повнотою представлення функціонування мовної системи можна виділити ***фундаментальні*** та ***дослідницькі (=пошукові) корпуси текстів.*** Свого часу О.С.Герд запропонував розрізняти корпуси текстів ***реєструвального*** та ***інтерпретаційного***, або ***дослідницького типу***. Перші становлять фактичне підґрунтя для створення других. Реєструвальні корпуси подають тексти як цілісні об'єкти, як факт реалізації мовної системи. Корпуси інтерпретаційні становлять інформаційно-довідкові та дослідницькі системи, що дають користувачеві змогу одержувати з корпусу текстів потрібну йому інформацію про окремі мовні об'єкти та їхні властивості[54]. Як правило, корпуси текстів за допомогою спеціальних засобів доступу до них та їхнього опрацювання (=текстових процесорів) виконують функції як реєструвальні, або представницькі, так і інтерпретаційні, або дослідницькі чи інформаційно-довідкові.

До фундаментальних належить, приміром, один з перших корпусів текстів – Браунівський корпус американського варіанта сучасної англійської мови, створений у Браунівському університеті (США) в 1962–1963 рр. під керівництвом У.Френсіса. Цей корпус містить 500 текстів загальним обсягом 1 млн. слововживань. У кожному з текстів 2000 слововживань. Крім однорідності за довжиною, представлені в корпусі тексти однорідні також за 1) часом публікації (усі вперше опубліковані в 1961 р.) та 2) характером мовного оформлення (всі створені літературною англійською мовою). Вони охоплюють 15 жанрів американського варіанта англійської мови: від газетних статей та релігійної й фахової літератури до художньої та ділової прози. У 1980 р. з'явилася індексована (анотована) версія Браунівського корпусу, яка становила покажчик усіх представлених у його текстах слів з їхніми адресами та граматичними характеристиками.

Сучасні національні корпуси текстів мають здебільшого фундаментальний характер. Їхні розробники намагаються представити користувачам функціонування описуваної мови якомога повніше і різноманітніше. Наприклад, фундаментальний характер має корпус британського варіанта сучасної англійської мови – Британський національний корпус (British National Corpus – BNC). Над реалізацією цього проекту працював колектив видавців, лінгвістів та програмістів під проводом видав-

[54] **Герд А.С.** Типы русских текстов и организация Машинного фонда русского языка // Машинный фонд русского языка: идеи и суждения. – М., 1986. – С.74.

ництва Оксфордського університету (Oxford University Press) протягом 1991-1994 рр. Корпус постійно оновлюють і на сьогодні він налічує понад 100 млн. слововживань (100 106 008 слів) у широкому спектрі письмових та усних текстів: від газет, фахової періодики та популярної белетристики до листів, спогадів і творів учнів шкіл чи студентів коледжів та університетів, а також від записів офіційних ділових чи урядових зустрічей до записів розважальних радіопередач і телефонних розмов. У цілому в корпусі 4124 різних тексти, з яких 863 усних (близько 21%). Кожне слововживання в ньому споряджено докладною граматичною характеристикою. Укладачі BNC для порівняння спробували представити корпус у вигляді звичайної книжкової продукції і одержали вражаючі показники. Якщо видруковувати корпус на тонкому папері з розрахунку 400 слів на сторінку, то весь його обсяг у друкованому вигляді займатиме простір близько 10 м^2. Для того, щоб прочитати цю продукцію зі швидкістю 150 слів на хвилину, витрачаючи на це 8 годин щодня, знадобилося б 4 роки[55]. На сьогодні такі представницькі корпуси створено уже для цілого ряду мов: англійської, французької, німецької, китайської, датської, естонської та ін. Зі слов'янських такі корпуси вже мають російська, польська та чеська мови. З усіма цими текстами після виконання певних вимог (оплата, правові зобов'язання тощо) можна працювати в Інтернеті.

Одним з найбільш грандіозних проектів формування фундаментального національного корпусу є корпус текстів французькою мовою, створюваний в Інституті французької мови Національного центру наукових досліджень Франції (м.Нансі) для укладання словника "Скарбниця французької мови" ("Trésor de la langue française"). На сьогодні цей корпус – Frantext - містить тексти французькою мовою XVI–XX ст.ст. загальним обсягом 209 млн. слововживань. У ньому представлені 3417 текстів художньої літератури (80 % його складу) та інших функціонально-стильових різновидів (релігійна, наукова, публіцистична, епістолярна та мемуарна проза – 20 % складу корпусу) близько 1000 різних авторів. На основі цього корпусу здійснюється укладання 15-томного словника французької мови XIX–XX ст.ст., реєстр якого налічує 80 тис. лексем. У цілому для цього словника запроектовано дві частини: словник старофранцузької мови (до XVI ст.) та новофранцузької мови (від XVI ст. до нашого часу). На основі створюваного корпусу текстів паралельно формується електронна лексична картотека, загальним обсягом 250 млн. цитат. Отже, розроблені текстові процесори дають змогу не лише одержувати різноманітну інформацію з текстів корпусу, а й конструювати за заданими форматами картки лексичної картотеки для укладання словників різних типів.

[55] Докладнішу інформацію про цей корпус можна знайти на сайті Британського національного корпусу в Інтернеті: http://info.ox.ac.uk/bnc/

Для вивчення сучасного стану мовних систем, аналізу змін у їхніх лексиконі та граматичному ладі формуються так звані ***динамічні корпуси текстів***. Одним з таких корпусів є корпус текстів сучасної російської публіцистики (90-і рр. XX ст.), створюваний у відділі експериментальної лексикографії Інституту російської мови РАН[56]. Стратегія формування цього корпусу враховувала різноманітні інтереси користувачів, як лінгвістів, так і фахівців в інших галузях знань. Важливим, як вважали розробники проекту, є дотримання рівноваги між репрезентативністю, якісним характером відібраних текстів для аналізованого функціонально-стильового різновиду і часового відрізка (90-і роки XX ст.) та кількісним критерієм, обсягом представлених текстів з тими чи іншими параметрами. Як зазначає А.М.Баранов, описуючи проблеми, що постали перед розробниками цього корпусу текстів, "...текстів повинно було бути достатньо багато, щоб відобразити всі релевантні властивості проблемної галузі"[57], тобто властивості публіцистики як окремого функціонально-стильового різновиду сучасної російської літературної мови. Відмітною рисою цього динамічного корпусу, як і взагалі корпусів текстів останнього часу, є його повнотекстовий, а не вибірковий, цитатний, характер. На підготовчому етапі формування корпусу було здійснено розмітку текстів за типами вміщеної в них інформації, важливої для користувачів. Здійснене структурування текстів за типами вміщеної в них інформації і дає можливість використовувати корпус як інформаційно-довідкову та дослідницьку систему. Кожен тип інформації про вміщений у корпусі текст становив окремий параметр класифікації такого тексту та його конкретні реалізації (значення параметра), так званий ***фасет*** (від франц. **facette** "грань"). Такий спосіб організації інформації про описувані об'єкти предметної галузі (у даному випадку – про публіцистичні тексти) здобув в інформатиці назву ***фасетної класифікації***. Фасетна класифікація текстів описуваного корпусу російської публіцистики враховує такі параметри: 1) джерело (значення параметра – назви конкретних видань, звідки взято тексти); 2) автор (близько 1000 прізвищ авторів); 3) назва статті (1368 окремих назв); 4) політична орієнтація видання (наприклад, "загальнодемократична" преса, "ліва" преса); 5) жанр ("спогади", "інтерв'ю", "критика", "круглий стіл", "нарис", " проблемна стаття", "репортаж", "рецензія", "фейлетон"); 6) тема (наприклад, "внутрішня політика", "зовнішня політика", "література", "мистецтво" – загалом 39 різних назв тем); 7) час публікації (дата).

Для української мови завдання створення різноманітних корпусів текстів та текстозорієнтованих баз даних різного призначення з відпо-

[56] Докладніше про концепцію і принципи формування цього корпусу текстів див. у: **Баранов А.Н.** Введение в прикладную лингвистику. – С.131-136.
[57] Там же. – С.131-132.

відними інтерфейсами та текстовими процесорами стоїть не менш гостро, ніж формування словникових баз даних і конструювання комп'ютерних словників і в цілому комп'ютерних лексикографічних систем. 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 у таких функціонально-стильових різновидах української мови, як ділова проза, законодавчі та інші державні документи (мова дипломатії, наприклад), реклама, мова засобів масової інформації, тексти з тих фахових галузей, що до 1991 р. обслуговувалися російською мовою[58].

У відділі структурно-математичної лінгвістики Інституту мовознавства ім.О.О.Потебні НАН України корпуси текстів створюються передусім для розв'язання конкретних дослідницьких завдань, пов'язаних із статистичним обстеженням функціонування тих чи інших мовних одиниць у текстах (передусім наукових) та вивченням закономірностей структурної організації текстів. Отже, це корпуси дослідницького, пошукового характеру. Їхня організація підпорядкована передусім стратегії укладання різних типів частотних словників української мови. Такими є, наприклад, корпуси текстів сучасної української публіцистики та сучасного українського наукового стилю. Обсяг кожного з цих корпусів – 300 тис. слововживань – зумовлений потребами укладання частотних словників саме словоформ, оскільки такого обсягу текстів достатньо для того, щоб слова з вірогідними статистичними характеристиками покривали 80–90 % тексту. Зокрема, корпус текстів сучасної української публіцистики формувався на основі ряду центральних українських газет ("Урядовий кур'єр", "Голос України", "Сільські вісті", "Культура і життя", "Україна молода", "Літературна Україна", "Молодь України", "Вісті з України", "Товариш", "Республіка") та журналу "Золоті ворота" за один – 1994 – рік. Розробники корпусу зазначають, що він "є однорідним з

[58] Наприклад, до 1991 р. і навіть у перший рік після здобуття Україною незалежності мовою дипломатичного листування була переважно російська. Така ж ситуація склалася і в справочинстві багатьох сфер державного управління, бухгалтерській, банківській, рекламній справі тощо.

боку мови, але різнорідним за тематикою (тут є законодавчі акти і політичні огляди міжнародного та внутрішнього життя, новини культури, науки, спорту, новели та уривки з романів, вірші, реклама тощо) і разом з тим актуальним для нашого часу"[59]. Усі тексти введено до корпусу в цілісному вигляді, оскільки передбачена можливість його використання для розв'язання інших теоретичних та практичних завдань, а також створення на його основі різноманітних текстозорієнтованих баз даних. Корпус текстів сучасного українського наукового стилю містить тексти монографій (індивідуальних та колективних), статей у наукових збірниках та журналах з гуманітарних наук – мовознавства, літературознавства, фольклористики, етнографії, мистецтвознавства, філософії, соціології, історії, що вийшли друком протягом 90-х років.

Окремим і надзвичайно важливим завданням є створення корпусів українських текстів у звуковій формі, особливо текстів розмовного мовлення. На сьогодні цією роботою вже серйозно зайняті українські діалектологи. Наприклад, такі корпуси текстів – звукові комп'ютерні хрестоматії-фонотеки – для говірок Донецької та Полтавської областей створив колектив співробітників кафедри української мови Донецького національного університету (Л.Фроляк, З.Омельченко, В.Познанська, Н.Клименко, Н.Михайлова, В.Дроботенко). Для роботи з цими корпусами звукових текстів створено зручний і доступний для користувачів-філологів інтерфейс. Такі інтерфейси в інформатиці прийнято називати ***дружніми***. Він надає можливість працювати з матеріалами фонотеки в різних режимах, залучаючи до інформації звукової інформацію графічну. Ось, наприклад, як організовано роботу з фонотекою "Українські говірки Донеччини", яка містить понад 50 годин звучання діалектних текстів, записаних від інформантів у 65 населених пунктах Донецької області. Записи споряджено короткою передмовою про історію формування українських говірок на Донеччині та їхні структурно-граматичні особливості. Подано список населених пунктів та список інформантів, карти з позначеним ареалом побутування тієї чи іншої говірки. Користувачі дістають можливість, працюючи з конкретними записами, одночасно знайомитися з характеристикою інформантів, вивчати ареали поширення певної говірки, а також порівнювати записи між собою. Фонотека "Українські говірки Донеччини" із засобами роботи з нею становить текстозорієнтовану базу даних у звуковій та графічній формі.

Проте завдання стоїть ширше – необхідне створення фундаментальних корпусів текстів як у письмовій, так і в звуковій формі не лише для різних регіонів України, а й для різних комунікативних ситуацій, те-

[59] **Дарчук Н.П., Грязнухіна Т.О.** Частотний словник сучасної української публіцистики // Мовознавство. – 1996. - № 4-5. – С.16.

кстів, продукованих носіями з різними віковими, професійними, культурно-освітніми, статевими, політичними, релігійними характеристиками. Тільки таке фронтальне обстеження функціонування сучасної української мови дасть адекватну й вірогідну картину стану сучасного українського лексикону й граматичного ладу української мови, дозволить виявити тенденції й закономірності їхнього розвитку, а отже, закладе надійну основу для створення нової – активної, функціонально-прагматичної, комунікативної – граматики української мови, для розв'язання багатьох невідкладних завдань мовного будівництва в сучасній Україні.

Створення електронних картотек, передусім лексичних, або ілюстративних корпусів, корпусів цитат є не менш важливим завданням. Сьогодні найбільш інформативними для користувачів як лінгвістів, так і нелінгвістів виявилися картотеки, пов'язані з повнотекстовими базами даних, тобто такі картотеки, працюючи з якими користувач в разі потреби з допомогою спеціальних засобів адресації дістає змогу від окремого слова або цитати перейти до цілісного тексту, в яких вони функціонують. Саме таку стратегію побудови автоматизованої лексикографічної системи було прийнято на початку 80-х рр. минулого століття в Словниковому відділі Інституту мовознавства АН СРСР у Ленінграді (нині – Інститут лінгвістичних досліджень у м.Санкт-Петербург, Росія). Стрижень цієї системи становив корпус текстів класичної російської художньої літератури XIX–XX ст.ст., які, на переконання, розробників системи, слід було ввести й опрацювати в першу чергу, оскільки вони "можуть служити джерелом для цитування, виявлення й підтвердження значень, вживань слів і т.ін."[60], а отже, вкрай необхідні при укладанні словників різних типів як еталон, зразок літературної російської мови. Для такого корпусу текстів було створено спеціальний інтерфейс з відповідним меню роботи з текстами. Меню становить бібліографічний опис текстів за такими параметрами і їхніми конкретними значеннями: 1) ім'я, прізвище автора (авторів); 2) заголовок (назва книжки, журналу тощо), якщо джерело обробляється неповністю, то слід зазначити його частину, розділ; 4) видавництво; 5) місце видання; 6) рік видання; 7) номер випуску та 8) повна назва книжки, якщо текст вміщено в збірнику, антології; 9) кількість сторінок та 10) контрольна інформація, яка вказує на місце зберігання тексту в корпусі. Кожен параметр має свою спеціальну мітку, яка виконує роль ключа для пошуку в корпусі необхідної інформації. Перелік параметрів, а відповідно, й меню мають відкритий характер і в разі потреби можуть поповнюватися новими параметрами класифікації текстів або новими значеннями вже наявних параметрів.

[60] **Рогожникова Р.П., Чернышева Л.В.** Организация словарной картотеки на базе автоматизированной системы // Теория и практика современной лексикографии. - Ленинград, 1984. – С.23.

З допомогою спеціального текстового процесора в кілька етапів здійснюється формування електронної лексичної картотеки. Комп'ютерні продукти, одержані на кожному з етапів, крім власне дослідницької цінності для лексикографів, можуть виконувати функції самостійних супровідних баз даних з інформаційно-довідковими та дослідницькими функціями.

Першим з таких продуктів опрацювання сформованого корпусу-фактографічної основи картотеки стали словопокажчики окремих текстів. ***Словопокажчик***, або ***індекс*** тексту становить певним чином упорядкований список усіх вжитих у ньому слів з їхніми адресами та додатковою інформацією про них. В описуваній системі російських лексикографів словопокажчик містить словоформу, її адресу в тексті (сторінка, рядок), та показник абсолютної частоти вживання. Самий список можна впорядкувати за абеткою початку (прямий список) або кінця слова (зворотний, інверсійний список), а також за показниками інформації про таку словоформу – спадом частот слів, послідовністю нумерації сторінок, рядків у межах сторінки тощо. Ось, наприклад, який вигляд має словопокажчик для збірника лірики П.А.Вяземського в описуваному корпусі текстів[61] (див. табл. 7):

Таблиця 7. Словопокажчик до видання "П.А.Вяземский. Лирика". М., 1979 з електронної лексичної картотеки Інституту лінгвістичних досліджень Російської Академії наук у м.Санкт-Петербург (Росія)

Порядковый номер словоформы	Частота	Словоформа	Адрес слова (страница, строка)
310	1	вслед	43 (9)
311	2	вспомнил	29 (12), 10 (15)
312	2	вспылал	33 (6), 14 (9)
313	2	встав	23 (23), 4 (26)
314	2	встаю	36 (10), 17 (13)

На наступному етапі словопокажчики окремих текстів об'єднуються у зведений словопокажчик для всього корпусу. Завдяки йому користувач дістає можливість одержати інформацію про весь спектр вживання певної словоформи в аналізованому корпусі текстів. На етапі, який безпосередньо передує формуванню лексичної картки для слів змінюваних частин мови здійснюється лематизація словоформ, або їхнє зведення до канонічної (словникової) форми – ***леми*** (від лат. **lēmma** "заголовок; тема твору"). В алгоритмі лематизації враховано можливі форми-омографи на зразок ***бе́гу*** (форма родового відмінка однини російського іменника ***бег***)***-бегу́*** (форма 1-ої особи однини російського дієслова ***бежать***)***, их*** (форма родового відмінка особового займенни-

[61] Цей і наступні приклади взято із зазначеної вище статті Р.П.Рогожникової та Л.В.Чернишової.

ка ***они*** та присвійний займенник ***их***), суплетивні форми типу ***шел, шла, шло, шли*** к ***идти́*** або ***де́ти*** від ***ребе́нок***, ***лю́ди*** від ***челове́к***. До омографічних форм комп'ютер подає всі можливі леми (канонічні форми), наприклад, ***ели*** – ***ель, есть, брани – брань, бранить, ее – ее, она***.

Формування власне лексичної картки можна здійснювати в скороченому чи повному варіанті. Скорочений варіант картки містить слово в канонічній формі і всі виявлені в корпусі текстів його словоформи з їхніми адресами в текстах. Для канонічної форми зазначена кількість вживань, напр.:

встретить	4
встретишь	29 (3), 10 (6)
встречал	35 (26), 16 (23)

Повний варіант лексичної картки містить також контекст вживання слова певного вигляду і певної, заданої користувачем довжини. Така цитата з тексту може подавати різне лексичне оточення описуваного слова. Наприклад, для слова ***богатое*** у віршах П.А.Вяземського комп'ютер може на запит користувача видати повну строфу:

> Но, признаюсь, хотя и лестно, а робею:
> легко не согласясь с способностью моею,
> обогатить, друзья, себе и вам назло
> писателей дурных богатое число

При потребі цей контекст можна розширити за межі строфи або скоротити його до партнерів зліва та (або) справа.

Подібну стратегію формування лексичної картотеки на базі фундаментального корпусу текстів було прийнято на початку 90-х років після здобуття Україною незалежності й для дослідницької програми автоматизації й прискорення лексикографічних досліджень, розроблюваної спільно співробітниками відділу структурно-математичної лінгвістики Інституту мовознавства ім.О.О.Потебні і відділу лексикології та лексикографії новоствореного Інституту української мови НАН України[62]. На жаль, через ряд причин об'єктивного й суб'єктивного характеру створену систему формування лексичної картки слова на основі процедур автоматичного морфологічного та синтаксичного аналізу тексту так і не було запроваджено в практику словникарської роботи в цих наукових закладах. За результатами морфологічного аналізу тексту комп'ютер визначав для кожної словоформи її граматичні характеристики та здійснював лематизацію текстових форм. Ці операції давали змогу організовувати слова за частинами мови, укладати словопокажчики за ал-

[62] Див. докладніше про цю програму і створені в межах її виконання комп'ютерні продукти у праці: **Клименко Н.Ф., Русанівський В.М.** Від універсальної бази лінгвістичних знань до комп'ютерного укладання словників // Мовознавство. – 1995. - № 4-5. – С.3-10.

фавітним та алфавітно-частотним принципом, за різними параметрами опису слів у текстах, автоматично формувати картку з потрібним обсягом інформації про функціонування слова в текстах.

Створена співробітниками відділу структурно-математичної лінгвістики Інституту мовознавства ім.О.О.Потебні НАН України система автоматичного граматичного аналізу текстів АГАТ (див. про неї докладніше далі у §§3 та 4 розділу 2) містила в складі модуля морфологічного аналізу алгоритм контекстного аналізу словоформ. Його призначенням було виявлення діагностичних контекстів (оточень) для омонімічних (омографічних) словоформ, тобто таких словоформ, буквена структура яких повністю збігалася. Цей алгоритм аналізу тексту давав змогу укладати ***конкорданси*** (від лат. **concordāre** "узгоджуватися" з **cor (cordis)** "серце"), або ***словники контекстів вживання слова, його лексичних оточень***. Як зазначає О.С.Герд, "конкорданс включає ті ж дані, що й словопокажчик, але, крім того, в ньому для кожного слова подано контекст, який ілюструє це слововживання". І далі: "Словопокажчики задають синтагматику непрямо – через відсилання до тексту, а конкорданси явно – через вказівку на конкретні словоформи, що стоять поряд з даною"[63].

Вибір одиниці опису в конкордансі, як і вибір типів інформації про неї (граматичної, семантичної, статистичної), а також типу й довжини її оточення в тексті залежить від призначення такого конкордансу, його місця й функціонального навантаження в системі опрацювання текстової інформації (текстовому процесорі) і, зрештою, призначення самого такого текстового процесора. Можна виділити ***фундаментальні конкорданси***, які становлять скарбницю знань про вживання в текстах тієї чи іншої мовної одиниці, і ***дослідницькі конкорданси***, підпорядковані розв'язанню конкретного завдання. Як фундаментальні можна розглядати конкорданси, що з вичерпною повнотою описують спектр слововживань в окремому творі або у творах певних авторів. Такий фундаментальний конкорданс для поетичних творів Т.Г.Шевченка – 220 віршів українською та російською мовами, вміщених у 1–2 томах повного зібрання творів Т.Г.Шевченка у 12-ти томах (К., 1989–1990), створили за допомогою комп'ютера канадські україністи Олег Ільницький та Юрій Гавриш[64]. Одиниці опису в цьому конкордансі становить 18401 різна лексична одиниця-словоформа, в цілому корпус конкордансу містить 83731 слововживання таких словоформ у текстах загальним обсягом 22241 поетичний рядок. Слова-омографи на зразок ***господи*** (фо-

[63] **Герд А.С., Богданов В.В., Азарова И.С., Аверина С.А., Зубова Л.В.** Автоматизация в лексикографии и словари-конкордансы // Филологические науки. – 1981. - № 1. – С.73.

[64] **Конкорданція** поетичних творів Тараса Шевченка / A Concordance to the Poetic Works of Taras Shevchenko. – I-IV тт. – Edmonton-Toronto, 2001.

рма кличного відмінка від слова ***господь*** і родового відмінка від слова ***господа*** "житло") подані як одна реєстрова одиниця[65]. До кожної одиниці опису додано відомості про абсолютну частоту її вживання в текстах, а також всі її текстові оточення ("текстуальні обставини", за визначенням упорядників). Кожна така ілюстрація вживання словоформи налічує три рядки: рядок з описуваною словоформою, рядок, що йому передує, та рядок, що слідує за ним. "Цитуючи три рядки контексту (практика незвична для конкорданцій), – зазначають упорядники обговорюваної праці, – ми бажали надати текстуальним обставинам якнайбільшої ясности"[66]. Якщо рядок з описуваною словоформою стоїть на початку або в кінці вірша, а отже, не має якогось партнера, то відсутність останнього засвідчують 6 рисок (**------**). Крім того, для кожної ілюстрації зазначена її адреса в тексті: номер сторінки видання, на якій вміщено вірш, та номер рядка вірша на цій сторінці з аналізованою словоформою. Наприклад, власну назву ***Дніпро*** у віршах Т.Г.Шевченка конкорданс засвідчив у різних її формах загалом 56 разів: по 18 – у формах ***Дніпро*** та ***Дніпра***, 13 – у формі ***Дніпр***, 4 – ***Дніпрі*** та 3 – ***Дніпре***. А ось як виглядає стаття конкорданса до словоформи ***Дніпр***, де ми знайдемо як ілюстрації рядки з добре всім нам відомих поезій Т.Г.Шевченка:

010A 0001 ------ / Реве та стогне Дніпр широкий, / Сердитий вітер завива,
192B 0013 А річечка його взяла / Та в Дніпр широкий понесла, / А Дніпр у море, на край світа

У поданих прикладах перший стовпчик містить номер сторінки за першим томом (позначений символом А) повного зібрання творів Т.Г.Шевченка, на основі якого укладався конкорданс, символом В позначено другий том цього видання. Другий стовпчик вміщує номер рядка вірша, в якому міститься аналізована словоформа. Скісні риски відокремлюють один від одного рядки в ілюстрації. Для укладання цього конкордансу було спершу створено текстозорієнтовану базу даних, у якій потім комп'ютер на основі спеціальних програм укладання конкордансів, зокрема програми Micro-OCP (Oxford Concordance Program), виявив всі різні словоформи, для кожної з них дібрав всі контексти її вживання та підрахував загальну кількість таких вживань.

Фундаментальні конкорданси вже існують для творів багатьох письменників: Луция Аннея Сенеки, Данте, В.Шекспіра, П.де Ронсара, О.С.Пушкіна, М.Горького, У.Блейка, У.Уїтмена та ін., а також для ряду сакральних текстів. Наприклад, такий комп'ютерний конкорданс ство-

[65] Там же. – С.XXI.
[66] Там же.

рений для Нового Заповіту[67]. Прикладом дослідницького конкордансу може служити конкорданс в системі автоматичного морфологічного аналізу тексту, створеній співробітниками відділу структурно-математичної лінгвістики Інституту мовознавства ім.О.О.Потебні НАН України. Самі типи контекстів у такому конкордансі визначено з огляду на можливе вживання в них омонімічних словоформ. Для побудови діагностичного контексту було введено поняття ***опорних точок***, тобто таких знаків у тексті (наприклад, розділових) або словоформ (прийменників, дієслів), у сполученні з якими можна однозначно встановити граматичні характеристики омонімічних словоформ. Довжина контексту задавалася залежно від того, на якій відстані від аналізованого слова розташовувалася в тексті опорна точка. Для кожного типу омонімії словоформ у текстах (а їх було виявлено загалом 22) визначено типи діагностичних контекстів, які будує комп'ютер на етапі контекстного аналізу тексту. За бажанням користувачів такого текстового процесора побудовані конкорданси можуть зберігатися як самостійний продукт або ж будуватися кожен раз при аналізі конкретного опрацьовуваного тексту, оскільки, підкреслимо ще раз, такі конкорданси мають передусім суто практичне, дослідницьке спрямування.

Фундаментальні й дослідницькі конкорданси різняться не так за способами організації в них інформації про вживання слова в тексті, як за повнотою опису слововживань і за принципами формування самого опрацьовуваного корпусу текстів. Типовий формат статті конкордансів обох типів містить як заголовкову одиницю – текстову словоформу або сукупність словоформ під канонічною формою з певними граматичними та частотними характеристиками в корпусі текстів та контексти заданого вигляду й довжини з адресами в них описуваних словоформ. Самі такі контексти можуть бути впорядкованими за алфавітом або за зростанням номеру сторінки чи рядка сторінки в їхній адресі. Конкорданси створюють як для прозових, так і для поетичних текстів. Специфіка організації слів в кожному з цих типів текстів спричинює й специфіку організації інформації про них у конкордансах. Наприклад, виданий 1974 р. американським дослідником Й.Т.Шоу "Словник рим Пушкіна" (Shaw J.T. Pushkin's phymes. A dictionary. – The University of Wisconsin Press, 1974) у другій своїй частині – "Конкордансі рим" – подає граматичну інформацію про кожну кінцеву словоформу рядка, а також про кожну таку словоформу, яка може римуватися з описуваною, та статистичну інформацію про такі словоформи – кількість римованих і неримованих рядків зі словоформою – об'єктом опису. У записі такої інформації перша цифра позначає кількість римованих, друга неримо-

[67] **Computer-Konkordanz** zum Novum Testamentum Graece. – Berlin, New York, 1980.

ваних рядків, наприклад: ***венчали*** 2.0; ***венчался*** 1.0, ***венчанной*** 1.0[68]. Ось, наприклад, як у цьому конкордансі описано слово ***рабá***:

раб'а емв-2; **судьб'а** ежи-1; **погреб'а** ммв

де емв-2 означає: единственное число мужской род винительный падеж 2-е склонение, ежи-1 – единственное число женский род именительный падеж 1-е склонение, ммв – множественное число мужской род винительный падеж

Окремо в "Конкордансі рим" коло того сегмента слова, який римується, подано зведену інформацію про кількість римованих та неримованих рядків, наприклад, для слів ***рабá, судьбá, погребá*** такий сегментом буде кінцевий склад **бá**. Показники його активності в пушкінських віршах за даними словника Й.Т.Шоу такі: **б'а** 5 6 4, де 5 – число кінцевих різних слів з цим сегментом, 6 – кількість римованих рядків з ним, 4 – кількість неримованих рядків.

2004 р. у Львові видано «Словник рим української мови» Святослава Караванського, над яким автор працював майже півстоліття[69]. Основу реєстру цього словника становить лексика, подана в академічному орфоепічному словнику-довіднику «Українська літературна вимова і наголос» (К., 1973). Крім неї, С.Караванський включив до свого словника і поетичну, рідковживану лексику, відсутню, як правило, в нормативних словниках академічного типу. Попри всю цікавість і корисність таких слів для цілісного уявлення про можливості української римотворчості С.Караванський у передмові до свого словника закликав до обережного вживання їх «не в поезіях», до підтримування «мовної рівноваги висловлюваного»[70]. Загалом у цьому словнику вміщено близько 58 тис. текстових форм слів української мови, що виявили здатність до римування. За підрахунками С.Караванського, такі словоформи розподілені за близько 12 тис. різних слів, оскільки кожна окрема лексема - одиниця лексикону представлена в словнику в середньому 4-5 своїми словозмінними (текстовими) формами.

Вихідну ознаку впорядкування слів з певними римами у словнику С.Караванського становить якість наголошеного голосного звука у складі рими. За цією ознакою всі рими поділено на 5 родин, а саме: родини з римами, що містять наголошені голосні И, Е, А, О, У. Як додаткові ознаки розміщення слів у межах певних родин рим взято до уваги праве та ліве оточення наголошеного голосного у слові. Для

---

[68] **Герд А.С., Богданов В.В., Азарова И.С., Аверина С.А., Зубова Л.В.** Зазнач. праця. – С.75-76.

[69] **Караванський Святослав**. Словник рим української мови, укладений як лінгвографічна модель формального нагромадження звукових сигналів мовним центром людини. – Львів, 2004. – 1048 с.

[70] Там же. – С.ХХ.

абеткування слів запроваджено спеціальний звуковий алфавіт, розроблений С.Караванським на основі відомих класифікацій голосних та приголосних звуків сучасної української мови. Ось, приміром, які відомості про здатність до римування лексеми ***слово*** в різних властивих їй текстових формах, можна одержати з обговорюваного словника. В однині, оскільки наголос в усіх формах падає на основу, це слово здатне творити лише жіночі рими, тобто рими з наголосом на голосному другого складу від кінця слова, наприклад, ***сло́во, сло́ву, сло́ві*** тощо. Отже, рими до цих форм слід шукати в родині з наголошеним голосним **О.** Як виявилося, форма ***сло́во*** римується зі словами різних частин мови, пор.: ***сло́во*** і ***сло́во*** (слово в слово), ***дібро́во!,*** *це не* ***но́во,*** *от і* ***чудо́во****, раз, два і* ***гото́во****, на серці* ***барвінко́во***.

До римування, як засвідчив словник, активно втягується не лише питома українська, а й запозичена – книжна, спеціальна – лексика. В одну групу за спільною моделлю римування в межах цієї ж родини об'єдналися, наприклад, запозичені з різних європейських мов (грецької, латинської, французької) прості іменники ***апофео́з, віртуо́з, гіпно́з, курйо́з, прогно́з, туберкульо́з,*** питомі складні іменники ***бомбово́з, водово́з, паро́во́з, тепло́во́з*** на позначення, здавалось би, далеких від поезії реалій життя та питомі українські прості іменники, уподобані поетами, – ***гроз*** (род. в. множини від ***гроза***) та ***моро́з***. Врахування можливого внутрішньослівного (передусім правого в лінійному записі слова) оточення наголошених складів дало автору змогу унаявнити потенціал не тільки так званих точних (з повним збігом частини слів після наголошеного голосного), а й неточних рим, тобто рим з частинами після наголошеного голосного, що звучать подібно завдяки наявним у них приголосним або голосним з певними спільними ознаками. Наприклад, до слова ***хоро́брий*** словник С.Караванського подає не лише точну риму ***до́брий*** (тут і далі в прикладах підкреслюємо спільну частину після наголошеного голосного), але й такі неточні рими з різним ступенем близькості до цього слова, як ***ко́бра, о́браз, о́бруч, руч-о́б-руч, жало́бний, озло́бний*** та інші. Крім відомостей про моделі римування, користувачі одержать з цього словника також інформацію про сполучуваність слів з певними римами, про їхні можливі фонетичні, граматичні та акцентуаційні варіанти. Цей надзвичайно цінний матеріал узагальнено в **Індексі граматичних категорій**, доданому до словника (с.984-1027).

Зважаючи на такий всебічний характер моделювання українських рим, можна цілком погодитися з С.Караванським, що його словник, крім свого прямого призначення, «може правити за орфографічний словник, за словник вимови та наголосів, за фразеологічний і за гра-

матичний словник»[71] української мови. Його також можна вважати різновидом словника сполучуваності слів – конкордансу – української мови. Зацікавлених читачів відсилаємо до вміщеної в словнику докладної передмови укладача, в якій ретельно розглянуто весь спектр структурних ознак, використаних для аналізу та організації слів у корпусі словника, а також накреслено перспективи його використання для вивчення властивих українській мові способів асоціювання слів за певними ознаками їхньої звукової будови, встановлення закономірностей українського римування та ресурсів римотворення.

Для з'ясування функціонального навантаження того чи іншого слова в тексті, його важливості для текстів певного функціонального різновиду мови, стилю, певної тематики та будови вкрай необхідні частотні характеристики. Їх містять так звані ***частотні словники***. Частотним називають словник, в якому для кожного слова подано кількість його вживань у тексті, або частоту його появи в такому тексті[72]. Укладанням частотних словників займається окрема лінгвістична дисципліна – ***статистична лексикографія.*** Обсяг текстів, на основі яких встановлюють частотні характеристики певних одиниць, називають ***вибіркою***. Масив мовних фактів (текстів, інвентарів мовних одиниць певних типів), з якого роблять таку вибірку, має назву ***генеральна сукупність***. В.С.Перебийніс запропонувала класифікувати частотні словники за 1) одиницями опису; 2) обсягом аналізованих текстів; 3) типом текстів; 4) повнотою представлення лексики певного корпусу опрацьованих текстів; 5) способом упорядкування одиниць опису; 6) типами статистичних (частотних) характеристик і повнотою їхнього представлення у словнику. Наприклад, за одиницями опису розрізняють частотні словники слів, словоформ, словосполучень, морфем, буквосполук; за обсягом вибірки такі словники поділяють на великі (з вибіркою в 1 млн. і більше слововживань), середні (обсяг вибірки від 400 тис. до 999 тис.), невеликі (з вибіркою від 100 тис. до 399 тис.) та мікрословники з вибіркою, меншою за 100 тис. слововживань. Самі ж вибірки можуть різнитися за функціональним стилем, жанром, тематикою, часом створення чи видання обстежуваних текстів, їхньою належністю певному автору. Для одержання вірогідних статистичних характеристик обстежених одиниць генеральна сукупність текстів або будь-яких масивів мовних одиниць (реєстрів словників, наприклад) повинна бути однорідною. У лінгвостатистиці – галузі лінгвістики, яка для вивчення мовних явищ застосовує статистичні методи, – прийнято розрізняти однорідність лінгвістичну та статистичну. В.С.Перебийніс визначила для текстів, які складають лінгвістично однорідну генераль-

[71] Там же.

[72] Див.: **Перебийніс В.С.** Частотний словник // Енциклопедія "Українська мова". – К., 2000. – С.724-725.

ну сукупність, такі необхідні ознаки: спільні 1) час написання; 2) жанр; 3) тематика; 4) належність одному автору, тобто спільні авторський стиль, манера письма[73]. Статистичну однорідність генеральній сукупності забезпечує її обстеження за єдиними статистичними методиками і подальший аналіз на підставі єдиних статистичних (частотних) характеристик. У свою чергу, вибірка з такої генеральної сукупності повинна бути репрезентативною, тобто такою, яка б давала можливість одержати максимально наближені до реальних показники розподілу в аналізованих текстах тих чи інших одиниць та особливостей їхнього функціонування в них. "Статистика, – зауважує В.І.Перебийніс, – дає в руки дослідника правила побудови вибірки, яка хоч і не буде охоплювати всі тексти, але буде відбивати їхні якості"[74]. Залежно від мети статистичного обстеження мовного матеріалу дослідник може будувати різні типи виборок: ***механічні, випадкові*** та ***типові***, або ***зональні***. Вони різняться між собою як за способом організації, так і за тими ознаками текстів, що дослідник обирає за визначальні при структуруванні генеральної сукупності. Вибірки різних типів не заперечують, а доповнюють одна одну. Механічна вибірка прагне до рівномірного розподілу аналізованих одиниць по всій генеральній сукупності. Її доцільно застосовувати при обстеженні відносно невеликих за обсягом текстів. Випадкова вибірка формується з допомогою таблиці випадкових чисел. Типова ж, або зональна вибірка становить обстеження з допомогою таблиці випадкових чисел текстів, однорідних за якимись з перелічених вище ознак, а саме: за часом створення чи публікації, жанром, стилем, будовою, автором, мовою оформлення. Випадкові та типові (зональні) вибірки формують для аналізу великих за обсягом текстів або для обстеження сукупностей текстів, різних за обсягом. Докладніше з методиками побудови виборок різних типів і статистичним апаратом їхнього аналізу зацікавлені читачі зможуть познайомитися в уже згаданій книжці В.С.Перебийніс, спеціально написаній для лінгвістів, що не мають спеціальної математичної підготовки. У нашому викладі обмежимося визначенням основних параметрів укладання частотних словників та статистичних показників мовних одиниць, а також обговоренням корисності статистичних характеристик для лінгвістичних досліджень.

За способом подання мовних одиниць, для яких встановлено частотні характеристики, виділяють ***повні*** словники, що містять всі одиниці, вжиті в аналізованих текстах, та ***неповні***, що містять лише одиниці з частотою, яка дорівнює або перевищує певний заданий ***поріг*** – граничний показник частоти. Частотні словники можуть різнитися і за типами представлених у них статистичних характеристик. ***Абсолютна***

[73] **Перебийніс В.І.** Статистичні методи для лінгвістів: Навч. посібник. – Вінниця, 2002. – С.15-17.
[74] Там же.- С. 17.

частота слова – це кількість всіх його вживань в окремій підвибірці вибірки або у вибірці в цілому. ***Відносну частоту*** слова засвідчує відношення абсолютної частоти його вживання у вибірці до загальної кількості слів у ній. ***Середню частоту*** вживання слова обраховують як відношення суми абсолютних частот його вживання в окремих підвибірках вибірки до суми таких підвиборок. ***Показник міри коливання середньої частоти в тексті*** дає змогу встановлювати розподіл обстежуваної одиниці в тексті, значущість, стабільність її появи в ньому та в текстах подібної до нього будови, тематики, стилю, жанру та інші. Різні види частотних характеристик доповнюють та уточнюють один одного. Наприклад, середня частота дозволяє уточнити показники абсолютної частоти вживання, якщо вони істотно різняться у підвибірках, а відносна частота, у свою чергу, уточнює обидві такі частотні характеристики одиниць, якщо неоднаковими за довжиною виявляються самі підвибірки у межах вибірки.

У 1981 р. було опубліковано 2-томний "Частотний словник сучасної української художньої прози" за ред. В.С.Перебийніс. Цей словник укладав вручну колектив співробітників відділу структурно-математичної лінгвістики Інституту мовознавства ім.О.О.Потебні НАН України. Об'єктом опису в ньому були окремі слова, причому як повноправні реєстрові одиниці до словника були введені дієприслівники. Загальна вибірка текстів для укладання цього словника становила 500 тис. слововживань у текстах творів 25 українських письменників, які були опубліковані в період між 1945–1970 рр. Було відібрано твори майстрів сучасної української прози, різноманітні за будовою та тематикою, письменників, мовну палітру котрих дослідники вважали показовою і зразковою для стану української літературної мови того періоду. Отже, цей частотний словник є повним, середнім словником слів, для яких встановлені такі статистичні характеристики: абсолютна частота вживання в загальній вибірці, абсолютна й відносна частота вживання окремо в авторській та прямій мові (мові персонажів творів), середня частота вживання, показник міри коливання середньої частоти, показник кількості джерел, у яких зустрілася така одиниця, та кількості мінімальних виборок, в яких встановлювалася середня частота. Для кожної одиниці вказано її частиномовну належність, для змінюваних слів подані показники граматичних характеристик, незмінювані слова позначені символом *. Текст словника складають 5 списків одиниць: 1) слів з частотою, яка дорівнює 2 і більше, 2) слів з частотою 1, 3) слів, що зустрілися більше 20 разів хоча б в одному джерелі (тексті) словника, 4) слів та словоформ, упорядкованих за спадом частоти вживання в прямій та авторській мовах. Перші три списки становлять так звані ***алфавітно-частотні списки*** слів, провідним принципом побудови яких є розміщення одиниць в алфавітному порядку. Два інших списки називають у статистичній лексикографії ***ранговими***: зі спадом частоти

слова у таких списках збільшується його ранг, або порядковий номер у такому списку. У першому списку для змінюваних слів встановлено канонічну форму, під якою об'єднано всі зафіксовані в текстах словоформи. Загалом у цьому частотному словнику подано статистичні характеристики близько 33 тис. різних слів сучасної української мови. Серед них, як ми вже згадували у §3 цього розділу в описі джерел формування бази даних морфемно-словотвірного фонду української мови, чимало рідкісних слів, оказіоналізмів, неологізмів, індивідуально-авторських слововживань на зразок ***аеродромчик, арифметик, безсердеччя, будбанк, батько-герой, батько-поштар, вертиголова, кулацюра, безвечірній, бульдожий, білоштанний, біло-спокійний, виразно-розмірений.*** Трапляються серед них і діалектизми (***банякуватий***, ***варгатий***) і історизми (***амант, адамасковий***), і навіть слова-покручі – кальки з російських лексем (***авоська*** (пор. із ***сітка, сіточка***)***, балбес*** (пор. з ***бовдур, телепень***)***, кукушка*** (пор. із ***зозуля***)***, нерозбериха*** (пор. з ***безлад, плутанина***)***, бандпосібник*** пор. ***посібник*** з ***помічник, помагач***). Поява кальок та покручів у реєстрі словника зумовлена суцільною вибіркою лексем для опису. Тексти подано без жодних лакун, а подібні слова в творах відіграють роль стилістичних маркерів мови тих чи інших персонажів. За показником абсолютної частоти вживання кожному слову приписаний певний ранг, тобто його місце в ієрархії одиниць мови за ступенем активності їхнього вживання в текстах сучасної української художньої прози. Очолили такий ранговий список слова службових частин мови, а також особові займенники. Ось які лексеми опинилися на перших його 6-ти позиціях (рангах): 1 – ***не*** – 4473 (цифра зліва від слова – показник його рангу в частотному списку, а цифра справа – показник абсолютної частоти його вживання в опрацьованому текстовому масиві); 2 – ***я*** – 3748; 3 – ***а$_1$*** *сп. (сполучник)* – 3231; 4 – ***в*** – 3180; 5 – ***і$_1$*** *сп. (сполучник)* – 2506; 6 – ***ти*** – 2476. Найвищі ранги серед дієслів мають лексеми ***бути*** (16-ий ранг, абсолютна частота вживання 1073) та ***знати*** (22-ий ранг, абсолютна частота вживання 728). Серед іменників найактивнішою в досліджених текстах виявилася лексема ***люди*** (55-ий ранг, абсолютна частота вживання 357). Ось, наприклад, яку інформацію можна почерпнути зі статті до слова ***думка*** в цьому словнику. У цілому форми словозмінної парадигми цього іменника жіночого роду зустрілися в опрацьованих текстах 445 разів, відповідно 82 рази в прямій та 363 рази в авторській мові. Окремі форми цього слова, за свідченням словника, побутують в аналізованих текстах з такими показниками абсолютної частоти вживання, тобто вживання і в прямій, і в авторській мовах: ***ду́мка*** – 63; ***ду́мки*** – 42; ***ду́мку*** – 71; ***ду́мкою*** – 19; ***ду́мці*** – 26; ***думкі́*** *наз. мн.* – 61; ***думо́к*** – 44; ***думка́м*** – 3; ***думкі́*** *зн. мн.* – 47; ***думка́ми*** – 37; ***думка́х*** – 32. Відносна частота найактивнішої в цій парадигмі форми називного відмінка однини ***ду́мка*** в опрацьованому масиві художніх прозових текстів ста-

новить 0,000126, або 63/500000, де 500000 слововживань – довжина текстової вибірки, в якій встановлено частотні характеристики слів, поданих у реєстрі "Частотного словника сучасної української художньої прози".

Встановлені частотні характеристики слів створюють надійну основу для визначення ядра, найбільш активних і значущих слів для текстів української художньої прози 40–70-х років минулого століття, дають можливість визначити обсяг, склад і закономірності реалізації в цих текстах українського лексикону. "Частотний словник сучасної української художньої прози" становить "портрет" текстів цього функціонально-стильового різновиду української літературної мови з погляду кількісних та якісних характеристик їхнього лексичного наповнення. 1998 р. на базі 2-томного "Частотного словника сучасної української художньої прози" з прямим абеткуванням одиниць, або абеткуванням за початком слів було укладено з допомогою комп'ютера й видано у паперовому вигляді однотомний "Обернений частотний словник сучасної української художньої прози"[75], який становить один з можливих граматичних словників сучасної української мови з розгорнутою інформацією про особливості вживання тих чи інших словоформ у прямій та авторській мові художніх прозових творів.

Методика статистичного опрацювання текстів та вироблений формат статті "Частотного словника сучасної української художньої прози" були використані в двох комп'ютерних частотних словниках, укладених співробітниками відділу структурно-математичної лінгвістики Інституту мовознавства ім.О.О.Потебні НАН України разом з працівниками лабораторії комп'ютерної лінгвістики при кафедрі сучасної української мови Київського національного університету ім. Тараса Шевченка, – частотних словниках сучасної української публіцистики та сучасного українського наукового стилю. За обсягом опрацьованих текстових виборок (300 тис. слововживань кожна) це невеликі словники. Їхньою відмітною особливістю є застосування автоматичних процедур лематизації, або виведення канонічних форм змінюваних слів на основі результатів автоматичного морфологічного аналізу текстів. Автоматично здійснювалося й саме укладання корпусу словників: формування реєстру одиниць та статистичні підрахунки. Словники мають ту ж будову, що й їхній еталон – "Частотний словник сучасної української художньої прози". Їхній текст також містить алфавітно-частотні та рангові списки слів. Новим є визначений в цих словниках статистичний показник – ***коефіцієнт стабільності слова*** – показник стабільності вживання слова в корпусі текстів, або показник регулярності появи слова в тексті. Цей статистичний показник враховує як частоту слова в тексті, так і

[75] **Обернений** частотний словник сучасної української художньої прози. – К., 1998.

характер його розподілу в ньому, а отже, виводиться з трьох релевантних частотних характеристик: абсолютної та середньої частоти вживання слова й коефіцієнту коливання середньої частоти. Коефіцієнт стабільності становить надійний показник значущості певного слова в аналізованому корпусі текстів, його функціонального навантаження в ньому. Слова з високим коефіцієнтом стабільності можна вважати ключовими для досліджуваного тексту. Останні роки співробітники лабораторії комп'ютерної лінгвістики при кафедрі сучасної української мови Київського національного університету ім. Тараса Шевченка працюють над створенням інтегрованого Частотного словника сучасного українського поетичного мовлення. Фактичну базу для його укладання становлять твори провідних українських поетів другої половини XX – початку XXI століття: М.Вінграновського, Ліни Костенко, І.Драча, І.Калинця, В.Коротича, В.Стуса та ін. Це словник комплексного типу, який подає інформацію про частотні характеристики окремих словоформ, їхні контекстні оточення, ступінь новизни в українському лексиконі, морфемну та словотвірну будову, здатність брати участь у творенні тих чи інших стилістичних фігур[76].

Статистична лексикографія, як і сучасна лексикографія в цілому, все більше тяжіє до створення словників комплексних типів, таких, які б в описі тих чи інших мовних одиниць унаявнювали різні ознаки їхніх форми, змісту або використання в тексті. Прикладом частотного словника комплексного типу може служити Словник мови Достоєвського, створюваний співробітниками Інституту російської мови ім.В.В.Виноградова Російської Академії наук під керівництвом А.Я.Шайкевича. Комплексний підхід до опису слів відбивають як реєстр цього словника, так і сукупність статистичних характеристик опису реєстру. Словник поєднує рангові списки слів та словосполук зі списками алфавітно-частотними. У реєстрі поруч з окремими словами подано (вперше в статистичній лексикографії) бінарні сполуки, а саме: певне слово в таких списках представлене з його лівими або правими сусідами в текстах творів Ф.М.Достоєвського: художніх, публіцистичних та епістолярних. Для таких випадків спільного вживання слів у текстах обчислено спеціальний статистичний ***показник текстового зв'язку слів***. Він виявився цікавим і значущим для прогнозування спектрів можливої сполучуваності того чи іншого слова, його текстових оточень. Наприклад, для слова ***глаза*** статистично значущим у текстах Ф.М.Достоєвського виявилося спільне вживання зі словами ***смотреть, обводить, впиться, зоркий, скрыться*** та ін.[77]. Показники абсолютної,

[76] Докладнішу інформацію про цей словник зацікавлені читачі можуть знайти на сайті в Інтернеті за адресою: www\proling.com

[77] **Шайкевич А.Я.** О статистическом словаре языка Достоевского // Русский язык в научном освещении. – М., 2001. - № 2. – С. 122-149.

середньої та відносної частоти вживання в текстах різних жанрів, періодів творчості письменника обчислено не лише для окремих слів, а й для їхніх сполук з іншими словами до цілих фраз включно, напр.: ***Бог, Бог в помочь, отдать Богу душу, вот Бог, а вот порог*** або ***выходить, выходить за, выходить замуж, выходить из себя, выходить сухим из воды***. Одержані частотні характеристики засвідчують динаміку функціонування окремих шарів російського лексикону в творах Ф.М.Достоєвського різного часу, тематики, жанрів, дають змогу створити цілісну статистичну модель ідіолекту цього письменника.

🕮 Терміни

- **повнотекстова база даних (=корпус текстів) –** упорядкована сукупність текстів у цілісному вигляді
  - **фундаментальна повнотекстова база даних (=корпус текстів) –** максимально повна й представницька сукупність текстів певною мовою
  - **дослідницька (пошукова) повнотекстова база даних (=корпус текстів) –** сукупність текстів, за обсягом і якістю джерел достатня для виконаннях певних класів дослідницьких завдань
- **електронна картотека (=ілюстративна база даних, база цитат) –** організована сукупність мовних одиниць (слів, словосполук тощо) з інформацією про їхнє вживання в тексті
- **словопокажчик (=індекс) –** упорядкований список усіх вжитих у тексті або корпусі текстів слів з їхніми адресами та додатковою інформацією про них
- **конкорданс (=словник контекстів вживання слова, його текстових оточень) –** словник, який подає контексти вживання певної мовної одиниці (слова, частини слова, словосполуки), її лексичні оточення в тексті
- **лінгвостатистика** – галузь лінгвістики, яка для дослідження мовних явищ застосовує статистичні методи
  - **статистична лексикографія** – розділ лінгвостатистики, який займається укладанням частотних словників
  - **частотний словник –** словник, в якому для кожного слова подано кількість його вживань у тексті, або частоту його появи в такому тексті
    - **генеральна сукупність** – сукупність текстів або інших джерел мовної інформації, на основі яких роблять вибірку і обчислюють частотні характеристики мовних одиниць

- **вибірка** – обсяг текстів генеральної сукупності, на основі яких встановлюють частотні характеристики певних мовних одиниць
  - **підвибірка** – мінімальна вибірка у складі вибірки, в межах якої обчислюються частотні характеристики мовної одиниці
  - **механічна вибірка –** вибірка, укладена шляхом рівномірного розподілу досліджуваних мовних одиниць у межах генеральної сукупності
  - **випадкова вибірка –** вибірка, укладена з допомогою таблиці випадкових чисел
  - **типова (зональна) вибірка –** вибірка з текстів, що мають спільні риси (час створення, будова, тематика, автор, жанр, мова написання), створена з допомогою таблиці випадкових чисел
- **частота** – показник кількості вживань мовної одиниці у підвибірці або вибірці (якщо текст короткий)
  - **абсолютна частота –** загальний показник кількості вживань одиниці у підвибірці вибірки або у вибірці
  - **середня частота –** показник середньоарифметичної кількості вживання одиниці у вибірці, обчислюваний як відношення суми абсолютних частот вживання у підвибірках до кількості таких підвиборок у вибірці
    - **показник міри коливання середньої частоти –** показник зміни значення середньої частоти у підвибірках вибірки
  - **відносна частота** – показник відношення суми абсолютних частот вживання певної одиниці у підвибірках до довжини вибірки (суми довжин підвиборок)
  - **поріг частоти –** граничний показник частоти (верхній чи нижній)
- **алфавітно-частотний список –** список слів, розміщених в алфавітному порядку, з певними частотами
- **ранговий список –** список слів, упорядкованих за спадом частот, з приписаним кожному слову **рангом** – його порядковим номером у списку

- **корпусна лінгвістика –** дисципліна в межах комп'ютерної лінгвістики, предметом досліджень якої є розроблення понятєвого та процедурного апарату для формування корпусів текстів (=текстозорієнтованих баз даних) та їх аналізу

§ 7. Комп'ютерний фонд української мови в Інституті мовознавства ім.О.О.Потебні НАН України

- Історія становлення
- Джерельна база
- Принципи формування
- Здобутки
- Перспективи розбудови

Поява комп'ютера і можливість моделювання з його допомогою різноманітних мовних об'єктів покликала до життя ідею створення ***комп'ютерних,*** або ***машинних фондів національних мов***. Вони мислилися як, з одного боку, скарбниці відомостей про будову та функціонування мови, а з другого, – як якісно нове фактичне підґрунтя для вивчення мови та опрацювання мовної інформації. Перші такі фонди з'явилися за кордоном (у США, Великобританії, Італії, Франції, Німеччині, Швеції) ще в середині 50-х років минулого століття. У §6 ми познайомилися з одним з таких фондів – корпусом текстів французької мови Національного центру наукових досліджень Франції у м.Нансі, створеним для укладання словника "Скарбниця французької мови" ("Trésor de la langue française")[78]. Залежно від конкретних завдань, які ставили перед собою розробники таких комп'ютерних систем опрацювання мовної інформації, вони будувалися як словнико- або текстозорієнтовані бази даних про мову.

У колишньому Радянському Союзі над проблемою формування машинних фондів національних мов серйозно й планомірно почали працювати з кінця 70-х років минулого століття. Одним з перших ідею комп'ютерного фонду російської мови як фундаментальної автоматизованої системи опрацювання інформації в мовній формі, що містила б "повний словник та генератор словоформ, а також формалізований тлумачний словник (тезаурус) російської мови", висловив академік А.П.Єршов, фахівець у галузі створення діалогових систем спілкування з ЕОМ[79]. Обґрунтовуючи ідею машинного фонду російської мови, А.П.Єршов виділив у формуванні її моделі як зовнішні, позалінгвістич-

[78] Зацікавлених читачів відсилаємо до праць, в яких подана ширша інформація про зарубіжні машинні фонди мов: **Розенцвейг В.Ю.** Опыт создания национальных лексикографических служб за рубежом // Машинный фонд русского языка: идеи и суждения. – М.: Наука, 1986. – С.75-84; **Тулдава Ю.А.** О машинных фондах русского языка за рубежом // Там же. – С.141-142; **Казакевич О.А.** Автоматизация лексикографических работ. Автоматические словари (обзор зарубежных публикаций) // Научн.-техн.информация. - Сер.2. – 1985. - № 9. – С.25-29.

[79] **Ершов А.П.** Методологические предпосылки продуктивного диалога с ЭВМ на естественном языке // Вопросы философии. – 1981. - № 8. – С.115.

ні, так і внутрішні, власне лінгвістичні чинники. До перших він відніс проблеми навчання мови, видавничу справу, розроблення систем взаємодії людини з комп'ютером. Надзвичайно важливе значення створення машинних фондів мов для розвитку самої лінгвістики А.П.Єршов вбачав у нових потужних можливостях глибше пізнати природу мову, ставити і успішно розв'язувати нові, фундаментальні, раніше недосяжні для лінгвістів теоретичні й практичні дослідницькі завдання, які надавали такі комп'ютерні системи представлення й опрацювання мовної інформації. Якісно новий виток у розвитку лінгвістичних досліджень уможливлює детальне структурування у комп'ютерних фондах мовних даних, цілісне представлення мови як взаємоналаштованих підсистем мовних одиниць різних рівнів. Така комп'ютерна модель системи російської мови, на думку вченого, становитиме формальну систему, "що повинна бути адекватною та рівнооб'ємною живому організму мови, але водночас вона повинна бути анатомічно відпрепарованою, розійнятою, доступною для спостереження, вивчення та зміни"[80].

У лютому 1983 р. у м.Вороново під Москвою зібралася перша Всесоюзна конференція з проблем створення машинного фонду для автоматизованої системи лексикографічних досліджень. Ухвала конференції, в якій взяли участь і українські вчені – працівники Інститутів мовознавства ім.О.О.Потебні та кібернетики ім.В.М.Глушкова АН УРСР (В.С.Перебийніс, М.М.Пещак, І.П.Білецька(Севбо)), підтримала пропозицію Наукової ради з лексикології та лексикографії та Секції лінгвістичних проблем опрацювання інформації Наукової ради з комплексної проблеми «Кібернетика» АН СРСР створити проект Машинного фонду російської мови як системи комплексної автоматизації лінгвістичних досліджень та розробок. Цей проект мав стати типовим для розроблення машинних фондів мов інших народів СРСР[81]. Виконання проекту передбачало розв'язання чотирьох фундаментальних завдань:

1. Створення академічних словниково-граматичних баз даних.
2. Формування автоматично поповнюваних словопокажчиків та словників на базі текстів ділових та розмовних стилів, а також текстів науково-технічної літератури й документації.
3. Об'єднання в єдиний мовний фонд даних про загальновживану російську мову та даних термінологічних фондів.
4. Створення фонду лінгвістичних алгоритмів та програм, включаючи процесори російської мови.

[80] **Ершов А.П.** Машинный фонд русского языка: Внешняя постановка // Машинный фонд русского языка: идеи и суждения. – М., 1986. – С.12.

[81] Зі станом реалізації цього проекту в різних фондах мов народів колишнього Радянського Союзу на момент його розпаду можна познайомитися в праці: **Казакевич О.А.** Машинные фонды языков народов СССР // Научн.-техн. информация. - Сер.2. – 1989. - № 10. – С.25-32.

В Україні на час вироблення проекту створення машинного фонду національної мови вже понад 15 років здійснювалися дослідницькі роботи, пов'язані з пошуками можливостей моделювання мовних явищ за допомогою комп'ютера. Такі дослідження провадили лінгвісти й математики-програмісти в науково-дослідних установах та вузах Києва, Харкова і Львова. В Інституті мовознавства ім. О.О.Потебні АН УРСР наприкінці 60-х років був створений відділ структурно-математичної лінгвістики на чолі з проф. В.С.Перебийніс, який активно працював над виробленням формалізованих моделей опису та інтерпретації мовних явищ. Саме цей колектив виявився на момент прийняття проекту створення машинного фонду мов найбільш ідейно й організаційно підготовленим для його здійснення на матеріалі російської та української мов. Група цього відділу, очолювана В.С.Перебийніс, основне ядро якої впродовж кількох десятиліть складали М.П.Муравицька, Т.О.Грязнухіна, Н.П.Дарчук, Л.В.Орлова, В.І.Критська, Л.І.Комарова, Т.К.Пуздирєва, Т.І.Недозим, продовжуючи роботу над створенням систем автоматичного морфологічного та синтаксичного аналізу російських наукових текстів, почала активне формування повнотекстових баз даних та укладання на їхній основі різноманітних частотних словників та конкордансів. Їхнє створення уможливило розроблення комплексної системи опрацювання текстової інформації на основі різноаспектного морфологічного та синтаксичного аналізу текстів і формування словникових баз даних (електронних лексичних картотек з автоматичними процедурами лематизації та генерування текстових форм слів) для автоматичного укладання словопокажчиків, різнотипних частотних словників та конкордансів. Після набуття Україною незалежності в 1991 р. група продовжила цю роботу на матеріалі україномовних текстів різної тематики – наукових, художніх, публіцистичних. На сьогодні повнотекстова база даних фонду містить близько 700 тис. слововживань, оснащена процедурами орфографічного контролю текстів, аналізу їхньої морфологічної, синтаксичної та семантичної структури. До друку підготовлені частотні словники двох функціональних стилів сучасної української мови – публіцистичного та наукового. Разом з виданим 1981 р. 2-томним "Частотним словником сучасної української художньої прози" вони становлять зведення відомостей про особливості функціонування сучасного українського лексикону в текстах різних жанрів, будови та тематики. Для опрацювання створених баз даних розроблені спеціальні текстові процесори, які включають системи комп'ютерного аналізу тексту на різних рівнях його будови: морфологічному, синтаксичному, логіко-семантичному (система АГАТ). До складу таких процесорів ввійшли також системи автоматичного орфографічного контролю й редагування текстів та система машинного російсько-українського та українсько-російського перекладу (системи РУТА і ПЛАЙ). Про них мова піде докладніше далі у відповідних параграфах розд.2 книги.

Інша група працівників цього відділу, очолювана д.ф.н. М.М.Пещак (до складу її входили також Н.Ф.Клименко, І.Ф.Савченко, Г.М.Ярун, Є.А.Карпіловська, Н.В.Сніжко, В.А.Шкуров, І.В.Цимбалюк), від кінця 70-х років активно розробляла проблему формалізації аналізу семантики слів. Їхні дослідження, виконані із застосуванням структурних та логіко-математичних методів аналізу змісту слів, взаємодії їхньої формальної й семантичної структури, спиралися на ретельне вивчення метамов словників різного типу, передусім тлумачних, що мають найбільш показові і розгалужені метамови, або мови опису реєстрових одиниць. Результати здійснених досліджень були узагальнені в колективних монографіях "Формалізовані основи семантичної класифікації лексики"[82] та "Лексична семантика в системі "людина-машина"[83]. Накопичений досвід опрацювання цієї проблеми уможливив розроблення проекту комп'ютерного семантичного словника української мови, який було оприлюднено наприкінці 80-х років у підсумковій роботі цього колективу "Украинский семантический словарь: Проспект"[84]. На жаль, цей масштабний задум залишився нереалізованим передусім через недостатньо чітку постановку самого завдання та відсутність стратегії формування комп'ютерної версії 11-томного тлумачного "Словника української мови", придатної для конструювання словника нового типу – семантичного як різновиду ідеографічних словників. Далася взнаки і нестача ресурсів для здійснення такого задуму, як матеріальних (техніка, витратні матеріали), так і людських (кадри лінгвістів, підготовлених для розв'язання завдань комп'ютерного моделювання мовної інформації, та математиків-програмістів з досвідом алгоритмічного та програмного опрацювання завдань підвищеної логіко-класифікаційної складності, до яких, безперечно, належать завдання моделювання будови та функціонування мовних об'єктів). Як показав досвід наступних років розбудови комп'ютерного фонду української мови, доцільно було починати з часткових лінгвістичних завдань, налаштованих на однорідні об'єкти певної підсистеми чи рівня мовної системи. Такий підхід уможливлював на кожному кроці розбудови фонду одержання замкнених циклів опрацьованої мовної інформації, а відтак – і створення ефективних комп'ютерних моделей для роботи з нею, успішне виконання певних класів дослідницьких завдань.

Саме таким шляхом з 1988 р. пішов новий колектив дослідників відділу структурно-математичної лінгвістики, очолений д.ф.н. Н.Ф.Клименко. До його складу ввійшли Є.А.Карпіловська, Л.І.Комарова, Н.В.Сніжко. Пізніше до них приєдналася Л.П.Кислюк. У різний час з цим колективом працювали математики-програмісти В.С.Карпіловсь-

[82] **Формалізовані** основи семантичної класифікації лексики. – К., 1982.
[83] **Лексична** семантика в системі "людина-машина". – К., 1986.
[84] **Украинский** семантический словарь: Проспект. – К., 1990.

кий, Г.В.Колєнов, С.Г.Буригін, М.А.Перельмутер, Т.І.Недозим. Дослідники поставили перед собою основне завдання: створити базу даних про морфемну будову сучасного українського слова на матеріалі найбільш показових і різнотипних за способом опису лексикону словників сучасної української мови і розробити засоби автоматизованого укладання морфемних та словотвірних словників, комп'ютерні моделі аналізу та синтезу слів. За характером поставленого завдання створюваний комп'ютерний фонд і було названо ***морфемно-словотвірним фондом української мови.*** Наприкінці 1991 р. було завершено формування бази даних у форматі та обсягах, передбачених першою чергою реалізації проекту побудови цього фонду. Сам проект і теоретичне осмислення перших результатів його опрацювання було висвітлено у низці публікацій його розробників, як колективних, так і індивідуальних[85]. Від початку реалізації цього проекту паралельно здійснювалися формування словникозорієнтованої бази даних морфемно-словотвірного фонду української мови та виконання власне дослідницьких лінгвістичних завдань. База даних формувалася як генеральний реєстр слів сучасної української мови, зведений за матеріалами 5 авторитетних та різних за способом опису лексикону словників: 11-томного тлумачного "Словника української мови" (К., 1970–1980), 2-томного словника-довідника І.Т.Яценка "Морфемний аналіз" (К., 1980–1981), 2-томного "Частотного словника сучасної української художньої прози" (К., 1981), "Словника іншомовних слів" за редакцією О.С.Мельничука (К., 1974) та орфографічної частини "Словника-довідника з правопису та слововживання" С.І.Головащука (К., 1989). На сьогодні цей зведений реєстр налічує 166385 слів з відомостями про їхню морфемну будову, кількість властивих їм значень, абсолютну частоту вживання у півмільйонній текстовій вибірці сучасної української художньої прози та частиномовну належність. Докладно про способи опису та представлення у

[85] Зацікавлених читачів відсилаємо до праць: **Н.Ф.Клименко, Є.А.Карпіловська, Л.І.Комарова та ін.** Морфемно-словотвірний фонд української мови як дослідницька та інформаційно-довідкова система // Мовознавство. – 1990. - № 6. – С.41-50; **Н.Ф.Клименко, Е.А.Карпиловская, Л.И.Комарова.** Машинные словари морфемно-словообразовательного фонда украинского языка // Актуальные проблемы компьютерной лингвистики. – Тарту,1990. – С.54-57; **Н.Ф.Клименко, Є.А.Карпіловська.** Морфемні структури слів у сучасній українській літературній мові // Мовознавство. – 1991. - № 4. – С.10-21; **Карпіловська Є.А.** Морфемна сітка як інструмент дослідження будови слова // Українське мовознавство. – 1992. – вип.19. – С.100-110. Підсумковий огляд стану розбудови комп'ютерного морфемно-словотвірного фонду української мови за 6 років (1988-1994) опублікований англійською мовою у "Журналі Міжнародної асоціації з квантитативної лінгвістики": **Klymenko N.F., Karpilovs'ka E.A.** Computer Morpheme-Word-Formative Database of the Ukrainian Language and Its Applications // Journal of Quantitative Linguistics. – 1994. - Vol.1. - No.2. - P.113-131.

морфемно-словотвірному фонді мовної інформації йшлося у §3 цього розділу. За матеріалами фонду було укладено комп'ютерні "Словник символьних моделей морфемної будови слова", "Словник афіксальних морфем української мови" (виданий у паперовому вигляді 1998 р.), "Кореневий гніздовий словник української мови" Є.А.Карпіловської (К., 2002). Теоретичне осмислення результатів здійснених досліджень викладено в низці публікацій цього колективу, оприлюднених впродовж 90-х років. Серед них: колективна монографія Н.Ф.Клименко та Є.А.Карпіловської "Словотвірна морфеміка сучасної української літературної мови" (К., 1998), монографія Н.Ф.Клименко "Основи морфеміки сучасної української мови" (К., 1998, 1-е вид., 2000, 2-е вид), монографія Є.А.Карпіловської "Суфіксальна підсистема сучасної української літературної мови: будова та реалізація" (К., 1999), спільна доповідь Н.Ф.Клименко та Є.А.Карпіловської на XII Міжнародному з'їзді славістів у Кракові "Морфеміка слов'янських мов як об'єкт типологічного вивчення"[86] та інші праці. Останні роки члени колективу працювали над реалізацією спільного проекту комп'ютерного "Шкільного словотвірного словника української мови" з найпродуктивнішими словотвірними гніздами слів. У версії словника, підготовленій для друку, 127 найчисельніших словотвірних гнізд українських іменників, дієслів, прикметників, числівників, у складі яких описано понад 16 тис. слів активного складу сучасного українського лексикону. Словник незабаром має побачити світ у серії "Словники України".

Комп'ютерний "Кореневий гніздовий словник української мови: Гнізда слів з вершинами – омографічними корен" Є.А.Карпіловської – перший не лише в українській, а й у слов'янській лексикографії морфемно-словотвірний словник гніздового типу. Він подає чотири типи гнізд слів: 1) ***словотвірні***, в яких слова пов'язані відношеннями мотивації, або формального й семантичного виведення з певного базового слова гнізда; 2) ***кореневі***, слова в яких пов'язані відношенням кореляції, або співвіднесення за спільним коренем і всі становлять рівноправні базові одиниці в такому гнізді; 3) мішані, ***кореневo-словотвірні***, кореневі гнізда, деякі з базових одиниць яких виявили словотвірну активність та 4) ***потенційні***, або ***нульові***, в яких реалізовані лише базові слова. До реєстру словника як вершини гнізд були відібрані лише кореніомографи, або корені з різним узагальненим лексичним значенням, але тотожною буквеною структурою, та їхні формальні варіанти. Коренів-омографів виявилося 1820, вони об'єднуються в 653 групи, чисельність яких коливається в інтервалі 2–14 різних коренів зі спільною буквеною структурою. Найпотужнішим щодо омографічності виявився корінь з буквеною структурою **пол-**. Його реалізовано як у питомих, так

[86] **Клименко Н.Ф., Карпіловська Є.А.** Морфеміка слов'янських мов як об'єкт типологічного вивчення // Мовознавство. – 1998. - № 2-3. – С.117-135.

і в запозичених словах сучасної української мови, а в гніздах він може виступати і як морф-домінанта, і як аломорф-залежний варіант домінанти, пор.: ***пол(а́), по́л(е), пол-ови́н(а), пол-ова́ти, пол-и́ц(я), пол-і́й, моно-по́л-ij(а)*** (від грецьк. πωλέω "продаю") та ***екстра-пол-я́ціj(а)*** (від лат. **polio** "поліpую, опоряджую"), ***по́л-ик←***_піл_, ***метр-о-по́л-ij(а)←***_по́ліс_ (від грецьк. **πόλις (πόλιδος)** "місто"), ***ватерпол-і́ст←***_ватер-по́ло←по́ло_ (від англ. **polo** з тібет. **pulu** "гра у м'яч"), ***випо́л-ювати←***_ви́-полоти←поло́-ти_, ***пол-я́к←***_По́ль-щ(а)_ та ***біл-о-по́л(я)←***_топо́л(я)_ (від **біл**[а]+**о**+[то]**поля**). Корені-омографи формують омопарадигми. У деяких з таких коренів омографічними можуть бути і залежні формальні варіанти – аломорфи. А отже, крім омопарадигм, у системі мови можуть вишиковуватися і омопари або оморяди коренів. Сукупності коренів-омографів з їхніми можливими формальними варіантами формують ділянки сучасного українського лексикону, на яких всі слова виявляються пов'язаними за коренем спільної форми або спільних форми та значення. Ось, наприклад, яку впорядковану ділянку подають 19 гнізд українських слів, вершини яких містять корені-омографи з формою **пол-** або пов'язані з такими вершинами через формальні варіанти цих коренів (див. рис. 11):

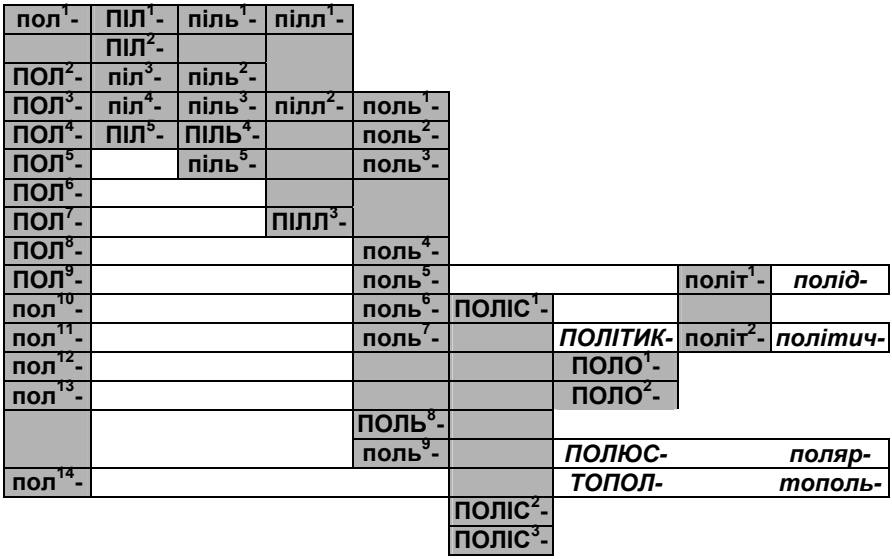


Рис. 11. Модель організації в сучасному українському лексиконі гнізд слів з коренями-омографами **пол-** та їхніми формальними варіантами

Разом з формальними варіантами коренів-омографів (їхніми домінантами, аломорфами чи дублетами) реєстр словника налічує 2594 корені, які реалізовані в корпусі з понад 71 тис.слів сучасної українсь-

кої мови. Всі слова об'єднані в 1330 гнізд, серед яких 888 словотвірних, 25 кореневих, 306 мішаних, коренево-словотвірних та 111 потенційних (нульових). Кореневий гніздовий словник Є.А.Карпіловської подає одну з можливих моделей будови сучасного українського лексикону: його двоярусну організацію за 1) коренями спільної форми та значення у гніздах різних типів та 2) за коренем спільної форми – коренем-омографом – у надгніздових єдностях різної чисельності. Багатошаровість організації лексики за омографічними коренями виявляється не лише на рівні вершин гнізд, а й на рівні окремих слів у складі самих гнізд – композитів та юкстапозитів, тобто результатів осново- та словоскладання, які містять кілька коренів-омографів. Так, наприклад, з двома коренями-омографами **ВОД-**, реалізованими у базових словах словотвірних гнізд ***вода*** та ***водити***, безпосередньо або опосередковано пов'язані ще 106 гнізд різного типу, які об'єднали 1415 слів. Скажімо, гнізда з базовими словами ***зерно*** і ***білий*** не пов'язані між собою на рівні своїх вершин-коренів, оскільки корені **зерн-** і **біл-** не перебувають у відношеннях омографії. Натомість такі гнізда опосередковано пов'язані через слово-композит ***білозірка*** "різновид солі з крупними білими зернами", яке входить до складу кожного з них. Таким чином, корені-омографи здатні відігравати системотвірну роль на різних рівнях організації слів: на рівні внутрішньо- або міжгніздовому, а також на рівні вершин гнізд або слів, що вони об'єднують. Укладений словник може становити основу для створення інших словників інтегрального типу, або словників, об'єктом опису в яких служать певні угруповання слів: тлумачних, тематичних, ідеографічних, перекладних. Основою для "збирання" лексики в таких словниках можуть бути різноманітні спільні ознаки форми та/або змісту слів.

До складу продуктів опрацювання бази даних морфемно-словотвірного фонду ввійшли і інші типи комп'ютерних словників української мови. Вище вже йшлося про укладений Н.В.Сніжко ідеографічний словник іменників української мови та про ідеографічний словник дієслів переміщення української мови, створений А.Я.Середницькою. Здійснено також спроби створити комп'ютерні версії "Словника староукраїнської мови XIV–XV ст."[87] та словників лінгвістичних термінів Є.В.Кротевича, Н.С.Родзевич та Д.І.Ганича й І.С.Олійника[88]. Останню роботу Є.А.Карпіловська виконувала разом зі студентом-дипломником

[87] **Білевич Т.Л.** Принципи ствоорення машинної версії "Словника староукраїнської мови XIV-XV ст." // Проблеми українізації комп'ютерів. – Львів, 1993. – С.14-15.

[88] **Карпіловська Є.А.** Термінологічний підфонд у складі морфемно-словотвірного фонду української мови /принципи формування та можливості використання/ // Україномовне програмне забезпечення. Матеріали 4-ої та 5-ої Міжнарод. науково-практ. конф. "УкрСофт". - Львів, 1995. - С.161-162.

українського відділення філологічного факультету Київського державного університету ім.Т.Г.Шевченка В.Кушніренком, який на основі створеної комп'ютерної версії цих термінологічних словників розробив модель тезаурусу лінгвістичних термінів, алгоритмічні процедури його побудови та здійснив програмну реалізацію процесу укладання такого комп'ютерного словника нового типу.

На сьогодні комп'ютерний фонд української мови, створений співробітниками відділу структурно-математичної лінгвістики Інституту мовознавства ім.О.О.Потебні НАН України, має розгалужену архітектуру. Він складається з трьох основних модулів-підфондів: текстової бази, генерального реєстру українських слів, зорієнтованого на 5 словників-джерел, та підфонду лінгвістичних процесорів (текстових і словникових), що виконують роль трансляторів-єднальних ланок між фактографічними базами даних. Окремий модуль фонду становлять базисателіти основних модулів, що містять продукти виконання різноманітних дослідницьких завдань опрацювання основних баз даних: словники, словопокажчики, конкорданси, таблиці сполучуваності одиниць різних мовних рівнів, комп'ютерні граматики текстів тощо. Схематично архітектуру цього фонду у сучасному вигляді можна зобразити так:

Рис.12. Архітектура комп'ютерного фонду української мови Інституту мовознавства ім.О.О.Потебні НАН України

Накопичений в україністиці досвід створення лінгвістичних баз даних, формування машинних копій та версій різнотипних традиційних ("паперових") словників, розроблення лінгвістичних словникових та текстових процесорів ставить на часі завдання об'єднання існуючої інформації на єдиній концептуальній та методико-процедурній основі в загальнодержавний комп'ютерний фонд української мови, який виконував би всі властиві такій інституції функції: інформаційно-довідкову, дослідницьку, навчальну та редакційно-видавничу. Запорука успішного розв'язання цього невідкладного завдання – співробітництво усіх фахівців – лінгвістів та математиків-програмістів, які нині працюють в цій галузі, вироблення правової основи для такої співпраці.

🕮 Терміни

- **комп'ютерний (=машинний) фонд мови –** сукупність баз даних та знань про будову та функціонування певної національної мови

- **комп'ютерного (=машинного) фонду мови підфонди (=модулі) –** складники комп'ютерного (=машинного) фонду мови: окремі лінгвістичні бази даних та знань (словнико- та текстозорієнтовані), продукти їхнього опрацювання (словники, комп'ютерні граматики, словопокажчики, конкорданси тощо), електронні лексичні картотеки, корпуси текстових ілюстрацій та лінгвістичні процесори (словникові й текстові), діалогові системи для роботи з ними в індивідуальному режимі або в режимі потрібної користувачеві комбінації модулів фонду
- **комп'ютерного (=машинного) фонду мови архітектура –** спосіб організації в пам'яті комп'ютера модулів фонду та конфігурація режимів роботи з ними
- **комп'ютерного (=машинного) фонду мови функції –** різновид виконуваних за інформацією фонду завдань
  - **інформаційно-довідкова –** одержання довідкової інформації як про систему й функціонування мови, так і про способи їхнього моделювання у фонді
  - **навчальна –** навчання мови та методам її комп'ютерного аналізу за допомогою інформації фонду
  - **дослідницька –** виконання власне лінгвістичних, пошукових завдань, спрямованих на одержання якісно нової інформації про мову або якісно нових продуктів представлення мовної інформації
  - **редакційно-видавнича –** автоматизоване редагування й укладання словників або інших продуктів опрацювання інформації, вміщеної у фонді, а також автоматизоване перевидання комп'ютерних копій і версій словників або текстів – джерел формування фонду

ІІ. ЛІНГВІСТИЧНІ ІНТЕЛЕКТУАЛЬНІ КОМП'ЮТЕРНІ СИСТЕМИ

§1. Природний інтелект (=інтелект людини) і штучний інтелект (=інтелект комп'ютера) як його модель

- Інтелект (= розум, мислення) як інструмент пізнання дійсності
- Складники інтелектуальної (=розумової, мисленнєвої) діяльності людини: пізнання, розуміння, зберігання (=пам'ять), генерування знань та вміння їх застосовувати (=обмін знаннями)
- Підходи до створення систем штучного інтелекту
- Машина та тест Тьюринга

У поданому на початку нашої книжки визначенні комп'ютерної лінгвістики як самостійної мовознавчої дисципліни було підкреслено її міждисциплінарний, комплексний характер. Така двоїстість статусу КЛ спричинена належністю її за об'єктом і предметом дослідження – мовою – до лінгвістики, а за інструментом її дослідження – комп'ютером – і зорієнтованими на нього процедурами опису й аналізу мовного матеріалу – до інформатики та комплексу дисциплін, спрямованих на створення комп'ютерних систем з так званим штучним інтелектом, або штучним розумом. При цьому ***штучний інтелект (artificial intelligence)***, або ***штучний розум (artificial mind)*** визначають як певну комп'ютерну модель природного інтелекту, або інтелекту людини. Результати мисленнєвої діяльності людини, щоб стати приступними для передачі іншим особам та для їхнього сприйняття ними, повинні бути унаочненими. Це і відбувається завдяки мові, яка фіксує в своїх одиницях наслідки розумових процесів людини, або їх вербалізує (буквально – "ослівнює", пор. з лат. **verbum** "слово"). Усі інші знакові системи фіксації мисленнєвої діяльності людини виступають стосовно мови як вторинні, а отже, місце мовної компоненти в комп'ютерних інтелектуальних системах і значення комп'ютерної лінгвістики як галузі, яка займається створенням такої компоненти, важко переоцінити і хибно недооцінювати. Як ми вже з'ясували раніше, обговорюючи проблеми моделювання, для створення будь-якої моделі треба спершу зрозуміти будову натурного об'єкта – її аналога, прообразу моделі, виділити його визначальні (релевантні) і менш значущі (нерелевантні з певного погляду або для певного типу моделі) ознаки. Отже, для того, щоб зрозуміти підходи і стратегію створення систем штучного інтелекту, проблематику розроблення таких систем, треба дослідити феномен природного інтелекту, проаналізувати те його розуміння, яке склалося на сьогодні в науках гуманітарного циклу і визначає рівень розвитку ці-

лого комплексу наук про людину як саме про **homo sapiens**, або ж **людину, яка мислить, або розуміє**.

Слово ***інтелект*** латинського походження (від лат. **intellēctus** "поняття, розум, пізнання, спостереження" з **intellego** "розуміти, пізнавати, спостерігати"). Двотомний академічний "Словник синонімів української мови" включив лексему ***інтелект*** до ряду еквівалентів слова ***розум*** у першому з властивих йому значень: "Здатність людей логічно мислити, пізнавати об'єктивну дійсність, засвоювати, запам'ятовувати і т.ін.". Крім того, цей словник виділяє ще таке значення слова ***розум***, як "здатність окремої людини запам'ятовувати, оцінювати обстановку, події тощо та відповідно діяти і поводитися; звичайний, природний стан людської свідомості". У тлумачному "Словнику української мови" слову ***інтелект*** дано таку дефініцію: "Розум, здатність людини думати, мислити; // Рівень розумового розвитку (IV, 35–36). Отже, ***інтелект, розум*** лінгвістичні словники тлумачать як субстанцію, або певну здатність мозку людини керувати її діями, рішеннями, поведінкою в певних ситуаціях. Таку здатність характеризують певні визначальні ознаки, що дозволяють судити про ступінь розвитку цієї здатності в певної людини, або, інакше кажучи, засвідчують ступінь розвитку її інтелекту. При цьому визначення ***розумна людина*** вказує на певну норму розвитку інтелекту, на встановлення якої в тій чи іншій людській спільноті великою мірою впливають соціальні, етнокультурні, психоментальні чинники, у цілому рівень духовного розвитку певного суспільства, специфіка взаємодії в його духовному житті економічних, правових, політичних, моральних настанов та приписів. Для ступеня розвитку інтелекту людини, як засвідчують дослідження нейропсихологів, філософів, когнітологів, вирішальними виявилися такі ознаки її мисленнєвої діяльності:

1. Повнота та адекватність сприйняття дійсності.
2. Уміння правильно ставити завдання розумової діяльності, добирати потрібну й всебічну інформацію про досліджувані об'єкти дійсності, структурувати певну предметну галузь – об'єкт розумової діяльності, а саме: віднаходити її релевантні ознаки і встановлювати зв'язки між ними.
3. Відповідно до норм та уявлень певного суспільства оцінювати, категоризувати й класифікувати одержану інформацію про певну предметну галузь; робити несуперечливі й адекватні висновки та узагальнення.
4. Віднаходити рішення, які дозволяли б якомога повніше й успішніше розв'язувати поставлені завдання й досягати в найпростіший спосіб мети розумової діяльності.

Якість процесу реалізації інтелекту як здатності, власне розумової діяльності характеризують такі критерії, які також визначають рівень інтелекту людини:

1. Швидкість здійснення необхідних процедур.
2. Вибір оптимального способу виконання завдання, найраціональнішого шляху до досягнення поставленої мети.
3. Уміння дібрати й організувати необхідні знання про досліджувані об'єкти з належною для виконання певного завдання повнотою та всебічністю.
4. Здатність будувати адекватну та несуперечливу модель досліджуваного об'єкта або процесу з високим ступенем пояснювальної та передбачувальної сили, тобто модель, придатну для ефективного опису, аналізу та пояснення одержаної інформації про модельовані ділянки дійсності.

Уважно проаналізувавши подані переліки ознак, ми знайдемо відповідне місце на окремих етапах розумової діяльності поняттям, що в свідомості носія української мови пов'язані з поняттями ***інтелект*** та ***розум.*** Це поняття, позначені лексемами ***свідомість, мислення, пізнання, пам'ять, інтуїція, кмітливість, обдарування, сенс, рація, глузд (здоровий глузд), розсуд, розмірковування, умовивід*** і навіть **– *геній***. Усі вони так чи інакше пов'язані з процесом аналізу, оцінки й використання в розумовій діяльності людини інформації, яку вона сприймає з навколишньої дійсності, з процесом створення в мозку людини знань про себе і довкілля. Утім, не слід забувати про те, що поняття ***інтелект***, як і саме поняття ***людина***, міждисциплінарне. Залежно від розуміння людини як біологічної, фізичної, хімічної, психічної, соціальної і т.ін. сутності кожна з відповідних галузей знання пропонує своє власне визначення інтелекту, виходячи з власного аспекту вивчення цієї сутності, призначеного для цього поняттєвого й процедурного апарату. Наприклад, психологи за інтелект людини вважають те, що вимірюють і визначають інтелектуальні тести. Вони запропонували для визначення інтелекту людини спеціальний кількісний показник її розумового розвитку, рівня знань та обізнаності, так званий ***коефіцієнт інтелектуальності*** (англ. **Intelligence quotient**, скороч. **IQ**), одержуваний на основі спеціально розроблених тестів[89].

Термін ***штучний інтелект*** (***artificial intelligence***) з'явився в науковому обігу 1956 р. під час роботи двомісячної літньої школи з проблем комп'ютерного оброблення інформації, організованої в Дартмутському коледжі (м.Хановер, штат Нью-Гемпшир у США). Деякі фахівці зрідка використовують і менш поширені терміни, як наприклад, ***когнологія*** (=наука про мислення, від лат. **cōgnitio** "пізнання, вивчення,

[89] Ідея і методика кількісного виміру розумового розвитку дитини на основі спеціальних тестів належать швейцарському психологу Альфреду Біне (1857-1911). Саме поняття ***коефіцієнту інтелектуальності*** для кількісного виміру інтелектуальних здібностей людини до наукового обігу запровадив німецький психолог Вільям Штерн (1871-1939).

уява, поняття" з **cōgnōsco** "пізнавати")[90]. Словники з інформатики та обчислювальної техніки подають такі дефініції терміна ***штучний інтелект***: "**Artificial intelligence** (штучний інтелект) – здатність автомата виконувати функції мозку"[91]. Чи не вперше сполуку ***штучний інтелект*** як самостійне гасло неспеціального словника подає "Толковый словарь русского языка конца XX века: Языковые изменения" за ред. Г.М.Скляревської в такому визначенні: "**Штучний інтелект** – В інформатиці – наукова проблема створення методів, які дозволили б використовувати людські уявлення для оперування структурами символів"[92]. У науково-популярній літературі знаходимо більш приступне для користувачів-нефахівців у галузі інформатики тлумачення цього терміна, яке ми й приймемо за основу наших подальших роздумів: «**Штучний інтелект –** розділ інформатики, пов'язаний з розробленням програм, що імітують здатність людини до розмірковування та навчання"[93]. Отже, сьогодні дослідження, спрямовані на створення штучного інтелекту, або штучного розуму, становлять предмет дисципліни в межах інформатики, яка займається створенням програм і цілісних комп'ютерних систем, спрямованих на моделювання різних проявів здатності людини мислити і розв'язувати певні логічні завдання. Фахівці чітко відокремлюють їх від тих завдань, розв'язати які можна за допомогою формалізованого алгоритму, математичного числення у чистому вигляді.

Тут ми стикаємося з парадоксом, а можливо, із закономірністю розвитку суспільства й людини. На нього вже неодноразово звертали увагу фахівці. З накопиченням знань про навколишню дійсність, удосконаленням технічних засобів їхнього опрацювання й використання, все більше завдань з розряду "інтелектуальних" переходить до розряду "формалізованих", або "механічних", "рутинних", тих, які піддаються численню і не вимагають спеціальних "інтелектуальних" комп'ютерних засобів. Історія людства при цьому кожен раз в тій чи іншій формі повторюється в життєвому циклі окремої людини. Досить порівняти коло

---

[90] Цей термін використовував, наприклад, у своїх працях американський дослідник Дж. Мак-Карті (J. McCarthy). Властивий йому комплексний підхід до розв'язання проблем штучного інтелекту і саме розуміння цього об'єкта дослідження як певної людинозорієнтованої сутності, на нашу думку, яскраво демонструє назва однієї з робіт Дж. Мак-Карті, виконаної спільно з Л.Ернестом та Д.Редді й опублікованої 1968 року – **"Комп'ютер з руками, очима та вухами"** (A computer with hands, eyes and ears – З російським перекладом цієї праці читачі можуть познайомитися в зб. "**Интегральные** роботы". – М., 1973. – Вип.1).

[91] **Англо-**русский словарь по вычислительной технике**.** - М., 1974. - С.201. Тлумачення подаємо у власному перекладі.

[92] **Толковый** словарь русского языка конца XX века: Языковые изменения. – Санкт-Петербург, 2000. – С.271. Тлумачення подаємо у власному перекладі.

[93] **Язык** компьютера. – М., 1989. – С.232. Тлумачення подаємо у власному перекладі.

завдань, які вміє розв'язувати 3-річна дитина й випускник загальноосвітньої школи, вже не кажучи про фахівця у певній галузі. Саме в цій динаміці й різноплановості інтелектуальної діяльності людини полягає основна складність побудови ефективних комп'ютерних моделей природного інтелекту, систем штучного інтелекту. В.Л.Стефанюк, редактор російського перекладу монографії "Штучний інтелект" відомого американського дослідника Ерла Ханта, так оцінив цю ситуацію: "Несталість словосполуки "штучний інтелект" почасти пояснюється тим, що в ній використано поняття інтелекта (іноді його перекладають словом "розум"), хоча створюється враження, що все менше і менше надії залишається на те, що поняттю "інтелект людини" в близькому майбутньому буде дано точне визначення, придатне водночас і для філософів, і для математиків, і для психологів, і для звичайних людей. На заваді виробленню такого визначення, прийнятного для вивчення будь-яких проявів розумової діяльності істоти або автомата, стоїть і відсутність на сьогодні загальної теорії мислення, нерозгадана таємниця життя й дії людського мозку. Ми так мало ще знаємо про мислення, свідомість, мозок людини, а без цього неможлива успішна побудова їхніх моделей для автоматів.

Одна з очевидних складностей полягає в тому, що досі на практиці в будь-якому конкретному прикладі інтелектуальної діяльності завжди трапляється так, що, щойно приходить повне розуміння процесу вибору рішення в межах цієї діяльності, прикметник "інтелектуальна" припиняють застосовувати"[94]. Показово між тим, що термін ***штучний інтелект*** народився лише в еру комп'ютерів, в еру персоніфікованих автоматів – роботів. Адже ідеї про "мислячі машини" вирують у людському суспільстві з давніх-давен. Згадаймо хоча б статуї єгипетських богів, які промовляли свої пророцтва вірним, або дельфійського оракула давніх греків чи корабель, яким Одіссей керував за допомогою голосу. До речі, саме в "Одіссеї" Гомера чи не вперше знаходимо дієслово κυβερνω "правити, керувати, вести корабель", яке збереглося і в сучасній грецькій мові й втілилося в інтернаціональний термін ***кібернетика*** – наука про керування, зв'язок і перероблення інформації. Причину досить пізньої появи терміна ***штучний інтелект*** можна вбачати в тому, що створення комп'ютерів як програмно керованих автоматів, по-перше, дало можливість зімітувати послідовно етапи мисленнєвої діяльності людини, представивши їх як окремі формалізовані процедури, а по-друге, уможливило формалізацію різних типів інформації про навколишню дійсність, приступних людині, інформацію, сприйману не лише мозком, а й різними органами чуття – зором, слухом, дотиком, і формувати на основі її опрацювання різні моделі знання про саму людину і її оточення.

[94] **Хант Э.** Искусственный интеллект. – М., 1978. – С.5.

Недарма для узагальненого представлення процесу комп'ютерного моделювання розумової діяльності людини англійський математик Алан Матисон Тьюринг запропонував у 1936 р. у своїй статті "Про обчислювані числа" гіпотетичний пристрій – прообраз комп'ютера, який міг читати, тобто сприймати, розуміти, або розпізнавати мовну інформацію у графічній формі, а також писати й стирати символи, тобто приймати рішення, представляти їх у певній графічній формі і замінювати одне прийняте рішення на інше. Роботу такого уявного пристрою, який згодом дістав назву "машина Тьюринга", можна було описати найпростішим алгоритмом. На кожному кроці роботи такого алгоритму дія «машини Тьюринга» визначена її поточним станом, або тим, що вона уміє робити, "знає" на цей момент, фігурально висловлюючись, "рівнем її розумового розвитку, інтелекту", та символом, який вона зчитує на цьому кроці роботи, тобто надходженням нового завдання для її "розумових здібностей, інтелекту". Так Алан Тьюринг змоделював у найпростішому вигляді, крок за кроком, процес розумової діяльності людини. Алану Тьюрингу, до речі, належить і саме знамените формулювання проблеми моделювання людського інтелекту – "Чи може машина мислити?"[95], відповідь на яке й шукають розробники систем штучного інтелекту. Очевидно, що для одержання на це питання чи то позитивної, чи то негативної відповіді, треба спершу окреслити, що ми будемо розуміти під процесом "машинного мислення":

- успішне й передбачуване виконання створеної людиною програми розв'язання певного завдання,
- неочікувані результати виконання, нову несподівану, не передбачену програмою інформацію, надану комп'ютером;
- відмову працювати за програмою, вказівку на ті чи інші не враховані людиною причини неможливості виконання завдання за таким алгоритмом, унаочнення комп'ютером прорахунків у нашій формалізованій моделі виконання поставленого інтелектуального завдання, створеній людиною?

Отже, відповідь на поставлене вище питання залежить від спрямування тих завдань, які вимагають застосування інтелекту і які людина передає комп'ютеру для виконання. А.М.Тьюринг запропонував також один з можливих критеріїв оцінки рівня "машинного мислення". Розроблена ним процедура відома як "тест Тьюринга". Ідея на позір досить проста: якщо людина, спілкуючись з машиною, не зможе помітити, що її співрозмовник – автомат, то можна визнати, що така машина мислить, як людина, тобто має інтелект, подібний до інтелекту людини. При цьому як умова "чистоти критерію" висувається ряд обме-

[95] Таку назву мала його книжка, що вийшла друком 1950 року, її переклад російською мовою побачив світ 1960 р.

жень, вимог до ретельності виконання тесту: передусім – контакт між людиною і машиною мусить бути не прямим, а опосередкованим, тобто їхнє спілкування здійснює певний пристрій. Крім того, це основне обмеження може мати ряд додаткових, підсилювальних умов проведення експерименту, наприклад, окреслення для спілкування певної предметної галузі, мови спілкування, її складу тощо. Як відзначають фахівці-практики, "незважаючи на свою суб'єктивність, тест Тьюринга як критерій інтелектуальності має великі переваги. Він спирається на порівняння з людиною. Доки в нас немає загальної теорії мислення, визнати дещо мислячим ми можемо, тільки порівнявши це "дещо" з людиною – єдиною істотою, що має цю характеристику апріорі"[96]. Саме так потрактували цю проблему і письменники-фантасти брати Стругацькі, вивівши у своєму відомому романі "Понеділок починається в суботу" мислячу машину "Алдан", яка вміла не лише рахувати в будь-якій метричній системі і розв'язувати логічні завдання, заперечуючи принцип вилученого третього, а й грати в японські шахи й "чет-нечет", а після того, як в неї вселили чиюсь безсмертну душу, іноді друкувати на виході: "Думаю. Прошу не заважати".

На сьогодні в галузі штучного інтелекту накреслилися такі напрями, в розбудові яких є певні здобутки й перспективи:

- доведення теорем;
- розпізнавання образів (слухових і зорових);
- теорія ігор;
- адаптивне, динамічне й евристичне програмування;
- прийняття рішень;
- природна мова та її машинне розуміння, або спілкування з ЕОМ природною мовою;
- системи з самоорганізацією та саморегулюванням, синергетика;
- роботика;
- створення комп'ютером музики;
- самонавчальні мережі;
- оброблення даних, представлених природною мовою;
- вербальне й концептуальне навчання.

Саме такий перелік основних курсів входить, наприклад, до навчальних програм зі спеціальності "Штучний інтелект" університетів США. Взагалі, за свідченням Е. Ханта, таких програм існує вже понад 100. Безперечно, нас цікавлять напрямки штучного інтелекту як наукової дисципліни, пов'язані з опрацюванням інформації природною мовою. Він неминуче для нас, фахівців-лінгвістів, постане в двох іпостасях – комп'ютерна модель інтелектуальної діяльності звичайного мовця і

[96] **Панков И.П.** Искусственный интеллект // Прикладное языкознание. Отв.ред. А.С.Герд. – Санкт-Петербург:, 1996. – С.93.

комп'ютерна модель інтелектуальної діяльності мовця-лінгвіста, фахівця, що має спеціальні знання про мову. Інакше ці моделі можна визначити як моделі самої мови – об'єкта вивчення і моделі діяльності дослідника, який вивчає мову. Ставлячи перед собою завдання створення таких комп'ютерних моделей, ми впритул підходимо до обговорення таких важливих понять, пов'язаних з нашим предметом, як ***знання*** та ***інформаційна модель об'єкта пізнання,*** або ***розумової діяльності людини.*** Невипадково, підсумовуючи шлях, який пройшла наука у створенні штучного інтелекту, дослідники окреслюють його як шлях від систем з так званою "чорною скринькою" через системи "машин знань" до інтерпретаційної ідеології моделювання природного інтелекту[97]. Цей шлях засвідчує глибше проникнення розробників систем штучного інтелекту в сутність мисленнєвої діяльності людини як джерела й приймача інформації про довкілля. Якщо в системах "чорної скриньки" опрацьовувалися дані, а самий хід їхнього опрацювання залишався за сімома замками – його замінювали уявлення про це того чи іншого дослідника, то в системах з "машинами знання" вже опрацьовувалися судження про дані, знання про дійсність, в системах з інтерпретаційною ідеологією до уваги береться весь спектр знань про довкілля, про ту чи іншу його модельовану ділянку, які можуть впливати на сприйняття та розуміння інформації і спричинювати різні способи її опрацювання, різну її інтерпретацію дослідником.

🕮 Терміни

- **інтелект –** здатність до логічного мислення та прийняття рішень
  - **інтелект природний (=інтелект людини) –** здатність людини до розумової діяльності, або здатність до логічного мислення та прийняття певних рішень
  - **інтелект штучний (artificial intelligence) (=штучний розум) (artificial mind) –** певна комп'ютерна модель природного інтелекту, або інтелекту людини
  - **коефіцієнт інтелектуальності (Intelligence quotient**, **IQ**) – показник рівня інтелекту, розумової здатності людини, обчислюваний на основі спеціальних інтелектуальних тестів
    - **"машина Тьюринга" –** програмований гіпотетичний пристрій, запропонований англійським математиком А.М.Тьюрингом, для моделювання послідовності найпростіших розумових операцій людського мозку
    - **"тест Тьюринга"** – алгоритм діалогу з автоматом, запропонований А.М.Тьюрингом для моделювання мисленнєвих процесів, реалізованих у діалозі між людьми

[97] **Баранов А.Н.** Введение в прикладную лингвистику. – С.312-314.

§2. Лінгвістичні проблеми створення баз знань

- Знання – основа і результат розумової діяльності
- Знання декларативні (=знання **що**) і процедурні (=знання **як**)
- Логічні моделі організації та способи представлення знань у системах штучного інтелекту
- Статичні моделі організації знань: фрейм, семантична сітка
- Динамічні моделі організації знань: сценарій, скрипт, план
- Організація знань в експертних системах

Інформатику визначають як науку "про технологію побудови, аналізу та використання людино-комп'ютерного (програмного) знання"[98]. На підставі саме такого розуміння предмета досліджень з інформатики, Л.С.Козачков основним завданням цієї наукової дисципліни вважає розроблення засобів та методів побудови, аналізу й узагальнення інформаційних моделей певної предметної галузі, або певного об'єкта керування[99], а відмітну рису знання в інформатиці вбачає в його особливому інтегруванні, узагальненні, формалізації та програмній реалізації[100]. Отже, ***знання*** – це вміння використати інформацію про певну предметну галузь для її аналізу, розпізнавання, це вміння оперувати такою інформацією, даними. Одержати знання про ту чи іншу предметну галузь – це вміти не лише виявити її об'єкти та їхні класифікаційні ознаки, але й пов'язати об'єкти між собою, побудувати їхню логічну (поняттєву) модель, придатну для комп'ютера. Таким чином, ***знання*** – це судження людини про саму себе або про об'єкти довкілля, твердження про наявність/відсутність у досліджуваних об'єктів чи явищ певної ознаки (ознак), зв'язку між об'єктами довкілля або між людиною і довкіллям. Учені-логіки виділяють три типи побудови суджень і, відповідно, три можливі типи побудови знання:

1. Судження про належність, існування певних ознак у певного об'єкта. Такі судження реалізують так звані ***табличні*** або ***списокові моделі знання***, представлені у вигляді двомірних матриць, одна з колонок яких містить об'єкти, а друга – приписувані їм ознаки.
2. Судження про підпорядкування одних об'єктів іншим, родо-видові відношення між ними. Це так звані ***тезаурусні***, або ***ієрархічні моделі знання***.
3. Судження про відношення між об'єктами, їхній зв'язок у дійсності за певними ознаками. Їх реалізують ***реляційні моделі знання***.

---

[98] **Козачков Л.С.** Прикладная логика информатики. – К., 1990. – С.6.
[99] Там же.
[100] Там же. – С.8.

Інтелектуальні бази даних, або бази даних з правилами виведення з них певних суджень знайшли своє застосування у комп'ютерних системах розв'язання проблемних завдань, які здобули назву ***експертних систем.*** Це системи, призначені опрацьовувати відомості про ту чи іншу проблемну галузь або ситуацію за певними встановленими правилами виведення логічних суджень. Свою назву вони одержали з огляду, по-перше, на те, що експерти – фахівці в певній галузі знання – беруть участь у створенні бази знань-основи роботи таких систем, а по-друге, сама така комп'ютерна система виконує надалі функції людини-експерта. Комп'ютер у таких системах "вміє" оцінювати інформацію, робити внаслідок її опрацювання висновки з того чи іншого питання. Отже, експертні системи подають якісно інший рівень "комп'ютерного мислення": комп'ютер здатний, як людина, робити різного роду припущення, умовиводи, спираючись на неповні або й навіть неточні відомості[101]. Робота над створенням таких інтелектуальних комп'ютерних програм опрацювання інформації в США розпочалася ще наприкінці 50-х років минулого століття. На сьогодні експертні системи довели свою ефективність у багатьох галузях фундаментальних та прикладних наук. Одними з найбільш розвинених є експертні системи з медичної діагностики, наприклад, система "Кадуцей" (кадуцей (від лат. **cādūceus**) – назва символу лікування – жезла бога Меркурія, покровителя гімнастики, торгівлі й красномовства, обвитого двома зміями). В діалоговому режимі людина-лікар відповідає комп'ютеру на все більш детальні питання про симптоми хвороби пацієнта, результати його обстеження та зроблені аналізи. Це триває доти, доки комп'ютер на основі оцінки одержаних даних не зробить певні висновки, умовиводи і не запропонує лікарю один чи кілька можливих діагнозів. Причому, такі діагнози видаються на екран за зростанням їхньої ймовірності. Інтерактивний режим роботи з експертною системою дає лікарю можливість коригувати вже введені дані, доповнювати їх новими відомостями, а отже, ставити комп'ютеру нові питання і змушувати його до пошуку нових умовиводів і рішень. Такі ж експертні системи існують для геологорозвідувальних, хімічних досліджень, досліджень з молекулярної біології (системи "Проспектор", "Дендрал", "Молген", "Генезис" та ін.). Фахівці вважають, що завдяки успішному впровадженню методів як індуктивного, так і дедуктивного опрацювання інформації експертні системи становлять сьогодні "приклад найбільш успішного практичного застосування того напрямку інформатики, яке відоме під назвою "штучний інтелект"[102]. Різновидом експертних систем можна по праву вважати і ті системи, які створено для розв'язання дослідницьких лінгвістичних завдань, наприклад, системи опрацювання текстової інформа-

[101] **Язык** компьютера. – М., 1989. – С.108-109.
[102] Там же. – С.110.

ції, аналізу та синтезу мовних одиниць, де присутні елементи як індуктивного, так і дедуктивного підходів до опрацювання даних і одержання на цій основі певних умовиводів.

У теорії і практиці комп'ютерної лінгвістики створено логічні моделі і конкретні форми представлення для всіх типів знання. Кардинальна різниця між ними полягає, на думку фахівців у галузі штучного інтелекту, в тому, що знання першого типу становлять так зване "знання що", або декларативне знання, а знання другого типу – "знання як", або процедурне знання. Для кожного з цих типів знання розроблені свої ***логічні моделі організації***, або ***структури представлення знань***. Логічною моделлю декларативних знань є, зокрема, так званий ***фрейм*** (від англ. **frame** "рамка"). Фрейм здебільшого визначають як структуру представлення знань про ознаки та будову об'єктів – ***термів*** певної предметної галузі або певної типової ситуації, а отже, як логічну модель організації декларативних знань, знань про те, "що" становить модельований об'єкт. Саме так розумів фрейм американський дослідник М.Мінський, який одним із перших почав використовувати таку логічну структуру представлення знань для моделювання дій автомата, що вмів пересувати предмети у просторі: "Фрейм, – твердив М.Мінський, – це структура даних, призначена для представлення стереотипної ситуації"[103]. Фрейм можна будувати для окремих термів або для їхньої сукупності, пов'язаної певними відношеннями. Фрейм для кожного терма у такій сукупності має власну будову. Такі фрейми для термів у складі більших сукупностей називають ***підфреймами***, або ***вкладеними фреймами***. Вони можуть пов'язуватися ієрархічними відношеннями ("рід-вид", "частина-ціле") і тоді загальний фрейм, що об'єднує такі підфрейми матиме ієрархічну будову. Проте терми можуть і не пов'язуватися безпосередніми відношеннями, а виступати як складники однієї предметної галузі або як атрибути (ознаки) однієї стереотипної ситуації. У таких випадках всі підфрейми у складі загального фрейма перебувають у відношеннях кореляції, а самий такий фрейм має реляційну будову. Таким чином, фрейм – це сукупність певних термів, важливих або визначальних для тієї чи іншої ділянки дійсності: предметної галузі або стереотипної, типової ситуації, логічна "рамка" для організації їхнього змісту або наших знань про них.

Терми у складі фрейма в цілому та окремих його підфреймів становлять певні вузли, так звані ***термінальні вузли.*** Ознаки термів, значення параметрів їхнього опису дістали назву ***слоти*** (від англ. **slot** "щілина, люк"). Слоти, власне кажучи, і є тими нішами, в які "пакують", вміщують знання про певний терм. Конкретне значення слота – це його ***зміст***. Декларативні знання про той чи інший терм залежно від за-

[103] **Минский М.** Структура для представления знаний // Психология машинного зрения. – М., 1978. – С.254.

вдання, яке ставить перед собою дослідник, вивчаючи ту ділянку дійсності, яку моделює фрейм, можуть бути суто мовними, або знаннями про назву того поняття, яке виражає терм, або ж позамовними, енциклопедичними, або знаннями про характер вживання, особливості побутування реалії або явища з такою назвою в певному соціальному середовищі. Енциклопедичні знання можуть включати різні етнолінгвістичні, лінгвокультурні, соціополітичні, психоментальні компоненти, так звані "фонові" знання, знання про тло, на якому висвічується зміст певного терма – моделі реалії чи явища.

Наприклад, для такої ділянки дійсності, як "житло" носії української мови, очевидно, передусім виділять терм ***будинок*** з рядом термів, пов'язаних з ним відношенням "частина-ціле" або відношенням належності до спільної предметної галузі "житло, місце проживання людини". Встановлюючи ознаки, за якими можна описувати зміст цього терма, ми неминуче пересвідчимося у важливості позамовної інформації для побудови адекватного й ефективного фрейма. Наприклад, істотною для змістової будови терма ***будинок*** є ознака "місцевість, де знаходиться модельований об'єкт", причому тут важлива не лише опозиція "місто-село чи взагалі позаміська зона", а й опозиція за ознаками рельєфу місцевості "рівнинна-гірська". Наявність окремих найменувань будинків за тією чи іншою ознакою конкретизації загального поняття, вираженого словом-назвою терма, засвідчує важливість такої значеннєвої класифікації для носіїв української мови, вказує на особливості концептуалізації світу в українській мові. Пригадаємо, що тільки будинок у сільській місцевості в прямому значенні можна по-українському назвати ***хата***, хоча в побутовому мовленні ця лексема вживається у розширеному, позбавленому вихідних змістових компонентів значенні "взагалі житло" (пор. *У мене в хаті тепло* (про міську квартиру). Так, наприклад, фіксує семантичну структуру слова ***хата*** СУМ: **хата –** 1. Сільський одноповерховий житловий будинок і 4. *розм.* Квартира (у 1 знач.) (СУМ XI 29–30) (пор. також з **квартира –** 1. Частина житлового будинку, що складається звичайно з однієї або кількох кімнат, кухні, передпокою тощо, з окремим ходом) (СУМ IV 130), а отже, слово ***хата*** у такому значенні позначає житло, але не будівельну споруду, як аналізоване нами поняття, позначене словом ***будинок***, і має фрейм іншої будови. Діалектна лексема ***колиба*** – це житло гірських жителів – чабанів та лісорубів – у Карпатах. Вживаючи ж назву ***курінь***, українці мають на увазі тимчасовий будинок для сторожів на баштанах, городах, який звичайно будували з недовговічного матеріалу. Б.Д. Грінченко у своєму словнику вказує на те, що слово це прийшло в український лексикон з побуту запорізьких козаків, які так називали місце, де разом мешкали члени одного військового загону. Згодом цю назву спільного житла певної кількості козаків стали вживати для позначення військової одиниці – складника Запорізького війська. Разом з тим міський будинок істотно

відрізняється від сільського і за зовнішнім виглядом, і за внутрішнім плануванням, особливо якщо цей об'єкт розглядати в історичній перспективі. Пригадаємо хоча б такі неодмінні складові традиційного українського сільського будинку – ***хати***, як ***прізьба, ґанок, сіни, світлиця, піч, мисник.*** Самі назви будинків, як ми вже бачили на прикладі колиби, також несуть виразний відбиток певного етнокультурного середовища, місцевості, історичного періоду розвитку спільноти, що їх будувала. В українській мові поруч з питомими словами, що позначали житлові будівлі українців, є ціла низка запозичених з різних мов слів, які позначають будинки, житло, властиві іншим народам. Наприклад, ***вігва́м*** (житло індіанців Північної Америки)***, коте́дж*** (від англ. **cottage,** спершу – селянський будинок) – невеликий заміський житловий будинок з ділянкою землі; будинок, розрахований на одну сім'ю)***, шале́*** (невеликий селянський будинок у горах Швейцарії)***, ю́рта*** (переносне житло кочових і напівосілих народів Центральної і Середньої Азії)***, яра́нга*** (переносне житло кочових чукчів і коряків, а також деяких груп евенків та юкагирів)[104] та ін***.*** Проте, як бачимо, суто лінгвістичної інформації про ці різновиди будинків нам явно не вистачає для побудови такого фрейму, який би унаочнив всі різновиди знань про цей терм, які нам можуть знадобитися при аналізі певної ситуації з тим чи іншим словом на позначення будинку.

Терм	Слот	Зміст слота
БУДИНОК	форма	восьмикутник, круглий
	матеріал	цегляний, дерев'яний, зі шкур
	місце збудування (характер рельєфу місцевості)	міський, сільський, гірський
	належність певному народу (етнічній групі)	гуцули (етнічна група українців), народи північно-східного Сибіру, швейцарці, індіанці Північної Америки, англійці, американці
	належність певній професійній групі	чабани та дроворуби (гуцули)
	характер покрівлі	конічна, пірамідальна
	кількість поверхів	одно-, дво-, багатоповерховий
	розміри	малий, середніх розмірів, великий
	якість матеріалів та оздоблення	багатий-убогий, комфортабельний-некомфортабельний, затишний-незатишний
	час, в який придатний для проживання	зимовий, літній, сезонний, цілорічний

Рис. 13. Структура фрейма для терма ***будинок***

[104] Визначення подаємо за тлумачними та енциклопедичними словниками, словниками іншомовних слів.

Без відомостей енциклопедичного характеру, без "фонових" знань ніяк не обійтися. Їх подекуди, крім фахових джерел (наприклад, літератури про архітектуру та будівництво) можна почерпнути і в тлумачних словниках. Наприклад, словник Б.Д.Грінченка подає таку дефініцію слова ***колиба***: 2) Зимнее жилище древорубов-гуцулов; строится в форме восьмиугольника, без окон и потолка, с отверстием в верхушке пирамидальной крыши для прохода дыму (II, 268). А ось, приміром, які істотні доповнення до поданої вище дефініції слова ***яра́нга*** в "Словнику іншомовних слів" за редакцією О.С.Мельничука можна почерпнути з "Советского энциклопедического словаря" (М., 1989, вид. 4-е випр. та доповн.): "Переносное жилище у народов сев.-вост.Сибири – круглое в плане, с конич. кровлей из шестов, покрытых оленьими шкурами". На підставі вивчення ознак будинку, описаних у дефініціях словників різних типів, можна сформувати, принаймні, такий фрейм цього поняття (див. рис. 13).

Сукупність слотів, а також перелік їхніх змістів окреслюють властиву тій чи іншій мові картину реалізації загального поняття (концепта) "будинок" у певному наборі термів-його реалізаторів, або картину концептуалізації цього поняття, скажімо, в українській мові. Причому, для кожного терма властивий зміст не окремого слота, а певна сукупність таких змістів. Наприклад, для терма ***колиба*** властивий такий набір слотів: "форма", "характер покрівлі", "належність певному народу (етнічній групі)", "належність певній професійній групі", "час, в який придатний для проживання" та змістів таких слотів, відповідно: восьмикутник без вікон та стелі, пірамідальний дах з отвором для виходу диму, гуцули, чабани та дроворуби, зимовий. Отже, для кожного з термів-назв конкретних житлових будівель можна за певним набором слотів сформувати власний підфрейм, а всі вони об'єднуються у загальний фрейм реляційної будови, оскільки всі такі терми між собою перебувають у відношеннях рівноправної кореляції, належності спільному поняттю. Проте будинок може служити не для житла, а для виконання певних виробничих або духовних функцій, пор. ***склад, пакгауз, вокзал, аеропорт, церква, собор, дзвіниця, каплиця, трапезна, усипальниця***. Таким чином, з введенням до складу загального фрейма цих термів поповниться набір властивих йому слотів та їхніх конкретних змістів. У свою чергу, поняття "будинок" містить цілу низку понять, які з ним перебувають у зв'язках включення, "частина-ціле". Це поняття, пов'язані із певними частинами будинку: ***поверх, підмурівок, стіна (мур), дах, горище, вікно, ґанок, під'їзд, вестибюль*** тощо. Кожному з таких складників також можна приписати певні ознаки, а отже, й визначити певні набори їхніх слотів. Внаслідок такої диференціації термів загальний фрейм набуває багатошарової ієрархічної будови. На кожному з рівнів такої ієрархії, в межах окремих підфреймів, терми можуть мати реляційну будову.

Фрейм становить статичну модель знання, яка унаочнює будову, внутрішню організацію об'єктів. Відомості про спосіб зв'язку між об'єктами дає інша логічна модель організації знань, інша статична структура їхнього представлення – ***семантична сітка***. Вона належить до так званих реляційних логічних моделей, оскільки унаочнює саме відношення між об'єктами. Семантична сітка є результатом сіткового моделювання предметної галузі або стереотипної ситуації, оскільки подає об'єкти у властивих їм відношеннях на відміну від фрейма як наслідку моделювання рамкового, яке обмежене унаочненням будови самого об'єкта. За принципом побудови і способом застосування вона близька до описаної нами вище і зображеної на рис.3 морфемної сітки. Як і морфемна, ***семантична сітка*** – це орієнтований граф, у вершинах, або вузлах якого розташовані об'єкти – терми, а ребра графа, або зв'язки між вузлами сітки вказують на характер відношень між термами. Ребра у семантичній сітці, як і в сітці морфемній, мають властивість так званої ***рекурсивності*** (від лат. **recursio** "круговерть"), тобто здатність до зворотного зв'язку, а отже, сітку залежно від мети дослідження можна "розгортати" від основного терма (чи основних термів) до залежного (залежних) або "згортати", просуваючись у зворотному напрямку. Семантична сітка виявилася ефективним засобом моделювання об'єднань лексем, що виражають спільне поняття, а отже, належать до спільних концептуальних чи лексико-семантичних полів, понятєвих чи тематичних груп лексики. З'ясування особливостей будови і реалізації таких об'єднань дає важливу інформацію для розроблення ефективних моделей автоматичного опрацювання текстової інформації. Саме такій меті служать семантичні сітки в дослідженнях лексики, здійснених відомим українським дослідником Е.Ф.Скороходьком[105].

Побудувати семантичну сітку можна лише, виконавши певні процедури аналізу й опису окремої сукупності однорідних об'єктів, а саме: 1) виявивши всі складники такого лексико-семантичного об'єднання, 2) визначивши їхню семантичну будову та 3) семантичні відношення між ними. При цьому можна використати кілька способів унаочнення семантичної сітки, представлення такої логічної моделі описуваної предметної галузі: 1) слівний; 2) слівно-символьний; 3) символьно-числовий; 4) аналітичний; 5) списковий; 6) матричний; 7) графічний, або геометричний. Принципи побудови семантичної сітки продемонструємо на прикладі змодельованої Е.Ф.Скороходьком структури лексико-семантичного об'єднання слів та словосполук сучасної англійської мови на позначення об'єктів підвищеного рельєфу[106]. Вихідний матеріал для

[105] Див. докладніше про ці дослідження в монографії: **Скороходько Э.Ф.** Семантические сети и автоматическая обработка текста. – К., 1983.

[106] Зацікавлених читачів відсилаємо до праці Е.Ф.Скороходька, в якій подано детальний опис цього дослідження: **Сетевое** моделирование лексики // Испо-

аналізу було дібрано з тлумачних словників англійської мови. Вирішальною для включення слова до складу досліджуваного об'єднання була наявність в його дефініції реалізаторів інтегральної семи "природний підвищений об'єкт". Напрямок зв'язків між такими словами визначала наявність якогось з них у дефініціях іншого слова. Слово, присутнє в дефініції іншого слова, називалося його ***семантичним компонентом***, а саме описуване слово – ***семантичним дериватом*** слова, присутнього в його дефініції. Щодо відношень між значеннями слів, то Е.Ф.Скороходько запропонував значення (семему, або сукупність окремих сем) описуваного слова вважати ***семантичною похідною***, а семему слова, яке його описує, – ***семантичним складником***. Як бачимо, застосована методика аналізу словникових дефініцій лексем близька до методики ступінчастої ідентифікації лексики Е.В.Кузнецової, представленої читачам у §5 розд. 1. До складу аналізованого лексико-семантичного об'єднання ввійшли такі англійські слова, як ***bank*** "насип"***, bluff*** "урвище, обрив"***, cliff*** "стрімчак, бескид"***, climb*** "підйом"***, crag*** "скеля"***, downs*** "пагорбкувата місцевість"***, eminence*** "узвишшя"***, escarpment*** "ескарп"***, height*** "високе місце"***, hill*** "пагорб"***, mountain*** "гора"***, peak*** "пік"***, plateau*** "плато"***, top*** "вершина" та ін. Зауважимо, що разом з підвищеними об'єктами природного походження до цього об'єднання включено і штучні, створені людиною підвищені об'єкти на зразок ***escarpment*** "ескарп – побудована людиною протитанкова загорода". На основі аналізу дефініцій слів ***height, hill, mountain, hilltop, top eminence*** встановлено, що слово ***hilltop*** "вершина пагорбу" становить семантичний дериват слова ***top***, а слова ***height, mountain*** є семантичними компонентами слова ***hill***. У свою чергу, ***height*** є семантичним компонентом слова ***mountain***. Спосіб графічного представлення цих змістових зв'язків між словами на позначення підвищеного рельєфу зображено на рис. 14:

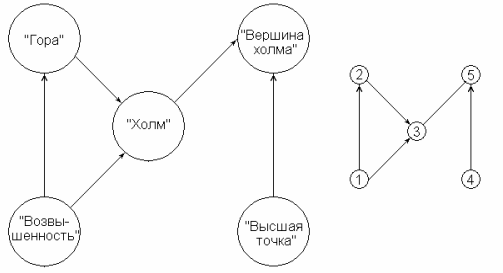


Рис. 14. Фрагмент семантичної сітки зі словами ***height*** "високе місце"***, mountain*** "гора"***, hill*** "пагорб"***, hilltop*** "вершина пагорбу" та ***top*** "вершина, вища точка" (за моделлю Е.Ф.Скороходька)

У цілому сітка цього лексико-семантичного об'єднання зображена на рис. 15:

льзование ЭВМ в лингвистических исследованиях. – К., 1990. – С.129-161. З цієї ж праці Е.Ф.Скороходька взято рис. 14 та 15.

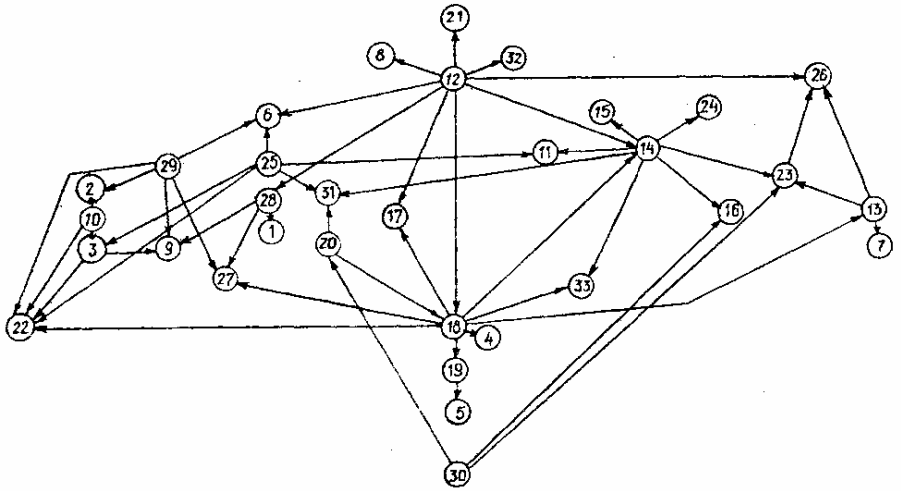


Рис. 15. Семантична сітка англійських слів на позначення об'єктів підвищеного рельєфу (за моделлю Е.Ф.Скороходька)

Примітка. Номери в сітці позначають слова в списку складників аналізованого лексико-семантичного об'єднання:

1. ***bank*** "насип"
2. ***bluff*** "урвище, обрив"
3. ***cliff*** "стрімчак, бескид"
4. ***climb*** "підйом"
5. ***col*** "депресія, гірський прохід"
6. ***crag*** "скеля"
7. ***downs*** "пагорбкувата місцевість"
8. ***eminence*** "узвишшя"
9. ***escarpment*** "ескарп"
10. ***face*** "поверхня (схилу)"
11. ***fell*** "пустинна пагорбкувата місцевість"
12. ***height*** "високе місце"
13. ***high land*** "гориста місцевість"
14. ***hill*** "пагорб"
15. ***hillside*** "схил пагорбу"
16. ***hilltop*** "вершина пагорбу"
17. ***massif*** "гірський масив"
18. ***mountain*** "гора"
19. ***mountain range*** "гірська гряда"
20. ***peak*** "пік"
21. ***plateau*** "плато"
22. ***precipice*** "урвище"
23. ***ridge*** "хребет"
24. ***rise*** "невеликий пагорб"
25. ***rock*** "скеля"
26. ***saddle*** "сідловина"
27. ***sierra*** "гірський ланцюг"
28. ***slope*** "схил, скат"
29. ***steep*** "крутий"
30. ***top*** "вершина"
31. ***tor*** "скеляста вершина пагорбу"
32. ***uplands*** "узгір'я"
33. ***valley*** "долина"

Крім графічного, як ми зазначили вище, можливі і інші засоби унаочнення зв'язків між членами цього об'єднання. Е.Ф.Скороходько подав їх так (пор. з рис. 15):

- Слівний запис: «Семема "високе місце" є семантичним складником семеми "гора";
- Слівно-символьний запис: "високе місце"→"гора";
- Символьно-числовий запис: 1→2;
- Аналітичний запис: X_1RX_2, де X_1 та X_2 – об'єкти, R – відношення між ними;
- Списковий запис: 1→2,3
  2→3
  3→5
- Матричний запис:

Семантичний складник	**Семантична похідна**				
	"високе місце"	**"гора"**	**"пагорб"**	**"вища точка"**	**"вершина пагорбу"**
"високе місце"	0	1	1	0	0
"гора"	0	0	1	0	0

Фрейм і семантична сітка моделюють статичні, нерухомі об'єкти довкілля. "Оживляють" статичні моделі інші логічні моделі знання – ***динамічні***, або такі моделі, які представляють об'єкти довкілля не лише у певних відношеннях, а у тих процесах, які з ними або між ними відбуваються, моделюють рух об'єктів. Для найменування таких моделей КЛ навіть використала термін кінематографії – ***сценарій*** (від лат. **scaena** "сцена, поле діяльності", **scaenārius** "який відбувається на сцені" з давньогрецьк. **skēnē** "шатро, намет"). Отже, ***сценарій*** – це динамічна модель представлення знань про об'єкти, яка побудована на певній послідовності логічно пов'язаних ***сцен*** – ситуацій, в яких між об'єктами або з об'єктами відбуваються певні дії. Іншим різновидом динамічних моделей знання є ***скрипти*** (від лат. **scrīptum** "писане, описане" з **scrībo** "писати, описувати") – послідовність дій з об'єктами, пов'язаних причиновими зв'язками, що мають загальноприйнятий, загальнозрозумілий, стереотипний характер. Скрипти для опрацювання концептуальної інформації запропонував американський учений Роджер Шенк. Свою теорію моделювання інформації про певні ситуації дійсності за допомогою скриптів він так і назвав ***"теорія концептуальних залежностей"***[107]. Приміром, зв'язок дій, представлений у реченнях: *Світає. Час збиратися в дорогу,* для носіїв української мови

[107] Див. докладніше в його праці: **Шенк Р.** Обработка концептуальной информации. – М., 1980.

прозорий і загальнозрозумілий: піднялося сонце, почався робочий день, засвітла зручніше вирушати в дорогу, аніж у темряві. Проте, зв'язок може бути і несподіваним, а сама така ситуація відрізнятися від типової, звичної. Такою, очевидно, є послідовність дій, представлена в реченнях: *Сумна звістка вразила її. Вона засміялася*, адже звичніша інша поведінка людини в цій ситуації: мовчання, розгубленість, плач, стогін, зойк, знепритомнення, взагалі будь-які прояви горя, страждання внаслідок одержаного сумного повідомлення. Однак саме сценарії та скрипти – найбільш відкриті для інформації лінгвокультурного, етнолінгвістичного характеру, оскільки поведінку людини в певних ситуаціях, її ставлення до інших людей та до дійсності в цілому регулюють приписи моралі, етики, культури, звичаї та традиції, усталені в певному суспільному середовищі в певний період його існування. Скажімо, відчуття болю змушує людину стогнати, кричати, це природньо і загальнозрозуміло. Проте ми знаємо, як по-різному ставиться оточення до такої поведінки в цих ситуаціях чоловіків і жінок. Якщо така поведінка в жінок зустрічає розуміння і співчуття, то у чоловіків, навпаки, всіляко вітають прояви мужності, вміння стерпіти біль без якихось зовнішніх проявів страждання. Тут для дослідника вкрай важливими виявляються різноманітні "фонові" знання, що стосуються модельованої ситуації, врахування особливостей сприйняття такої моделі тим, кому вона адресована. Наприклад, у західному культурному середовищі білий колір – є символом чистоти, цнотливості, саме в біле вбираються дівчатка для першого причастя та наречені для взяття шлюбу. Натомість на Сході, наприклад, у Китаї, білий – колір жалоби[108]. Отже, моделюючи певні ситуації, доводиться враховувати і певні усталені уявлення про символіку, особливості сприйняття тих чи інших реалій дійсності, поведінку людей в певних спільнотах. Врахування цього чинника дедалі більше вимагає такий рух у лінгвістиці, як мовна коректність. С.Г.Тер-Мінасова подає цікаві й показові приклади поширення цих тенденцій у сучасній англійській мові, що призвело до істотних змін у сучасному англійському лексиконі. Найбільш показовий у цьому плані прикметник ***black*** "чорний"[109]. Ми також стикаємося з намаганнями дотримати мовної коректності у нашому повсякденному спілкуванні. Це виявляється, зокрема, у функціонуванні таких кліше, як "особи літнього віку" замість "старі люди", "піти на заслужений відпочинок" замість "піти на пенсію за віком" або "робота не відповідає вимогам" замість "робота виконана по-

[108] Цікаву й важливу інформацію про особливості реалізації концептів таких понятєвих груп, як назви кольорів, тварин, птахів, рослин в українській, російській, англійській та китайській мовах зацікавлені читачі знайдуть у монографії української дослідниці І.О.Голубовської. Див.: **Голубовская И.А.** Этнические особенности языковых картин мира. – К., 2002.

[109] **Тер-Минасова С.Г.** Язык и межкультурная коммуникация. – М., 2000.

гано, непрофесійно, недбало" чи "неспортивна поведінка", тобто "поведінка, що не відповідає етичним нормам, прийнятим у спортивному середовищі, порушує їх".

🕮 Терміни

- **знання –** судження людини про саму себе або об'єкти довкілля, твердження про наявність/відсутність у них певної ознаки (ознак), зв'язку між об'єктами довкілля або між людиною і довкіллям
  - **знання декларативне (= знання "що") –** знання про будову об'єктів або явищ, про те, що вони становлять
  - **знання процедурне (= знання "як") –** знання про застосування, функціонування об'єктів або явищ, про те, як вони діють
- **логічні моделі знання (= структури представлення знання)**
  - **статичні логічні моделі знання** – моделі, що організують знання про будову об'єктів в стані спокою
    - **табличні (спискові) логічні моделі знання –** структура представлення знання, в якій об'єкту приписані певні ознаки (атрибути)
    - **фрейм –** рамкова структура представлення знань про певну предметну галузь або певну типову ситуацію, в якій окремому об'єкту – терму приписані набори параметрів його опису з їх можливими значеннями
      - **терм (= об'єкт, термінальний вузол)** – об'єкт, залучений до представлення знань про певну предметну галузь або певну типову ситуацію
      - **підфрейм (=вкладений фрейм) –** фрейм залежного терма, підпорядкований фрейму основного терма
      - **слот –** параметр опису терма, його ознака, атрибут
        - **зміст слота –** конкретне значення слота
    - **реляційні логічні моделі знання –** моделі організації знань, побудовані на основі різних відношень між термами: виведення, взаємовиведення, паралельного виведення з третього терма тощо
      - **тезаурусні (= ієрархічні)** – моделі організації знань, побудовані за родо-видовими відношеннями термів або відношеннями "частина-ціле"
      - **семантична сітка –** модель організації знань, що визначає терми (об'єкти) та напрямок залежності між ними у вигляді орієнтованого графа
        - **семантичний дериват –** найменування (слово або словосполука) в семантичній сітці, описуване дефініцією з іншим найменуванням

        - **семантична похідна –** значення семантичного деривата
    - **семантичний компонент –** найменування в семантичній сітці, яке описує інше найменування, входить до складу його дефініції
        - **семантичний складник –** значення семантичного компонента
        - **рекурсивність –** здатність об'єкта до зворотного зв'язку з іншими об'єктами
- **динамічні логічні моделі знання** – моделі, що організують знання про дії з об'єктами, або про об'єкти в стані руху, дії, функціонування
    - **сценарій –** модель представлення знань про об'єкти, яка побудована на певній послідовності логічно пов'язаних сцен (типових ситуацій) з ними
        - **сцена –** ситуація, в якій між об'єктами або з об'єктами відбуваються певні дії
    - **скрипт** – послідовність дій з об'єктами, пов'язаних причиновими відношеннями

§3. Автоматичний морфологічний аналіз тексту (АМА)

- Системи автоматичного перероблення тексту (АПТ), або автоматизовані системи опрацювання тексту (АСОТ)
- Модулі систем АПТ, або АСОТ – аналоги рівнів будови та розуміння тексту
- Підходи та стратегія створення систем автоматизованого морфологічного аналізу (АМА) тексту
- Модулі системи АМА: доморфологічний, флективний та контекстний аналіз тексту

Опрацювання текстової інформації незмінно залишається провідним завданням комп'ютерної лінгвістики. Наші знання про дійсність втілюються у певній вербалізованій 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. Принципову відмінність стратегій граматичного аналізу тексту в традиційній та комп'ютерній лінгвістиці Ю.М.Марчук визначив так: "...У комп'ютерній лінгвістиці поняття морфологічного аналізу є поняттям операційним. Якщо у традиційній

лінгвістиці до морфологічного аналізу належить те, що характеризує форму і відповідає на питання "що" класифікують, то в обчислювальній (прикладній) лінгвістиці важливо не "що", а "як" одержують ту чи іншу інформацію, тобто з форми слова у тексті"[110]

Системи автоматичного перероблення тексту (АПТ), або автоматизовані системи опрацювання тексту (АСОТ) становлять на сьогодні один з основних різновидів ***лінгвістичних інтелектуальних комп'ютерних систем***, систем які моделюють розумову діяльність людини при розв'язанні певних теоретичних або практичних завдань. Отже, системи АПТ, або АСОТ – лінгвістичні інтелектуальні системи, призначені для аналізу будови тексту на морфологічному, синтаксичному й логіко-семантичному рівнях та ідентифікації складників тексту в термінах відповідних модулів комп'ютерної граматики. У стратегії створення систем комп'ютерного аналізу текстової інформації можна виділити два основні підходи, дві різні технології опрацювання тексту. Перший підхід – ***словниковий*** – передбачає створення допоміжних лінгвістичних баз даних: словників, зведень правил перетворення форми тих чи інших одиниць, визначення їхньої ідентичності для виконання розроблених алгоритмів. Другий підхід одержав назву ***безсловникового***, або ***"незалежного"***, він передбачає представлення всіх потрібних відомостей про мовні одиниці у вигляді алгоритмічних правил. Згадані підходи аж ніяк не заперечують один одного. У кожного з них є свої недоліки й переваги. Обрання при створенні певної системи АПТ, або АСОТ одного з них як провідного зумовлює тип мови тексту, характер поставленого перед розробниками такої системи завдання, можливості самої комп'ютерної техніки (передусім обсяг пам'яті комп'ютера та його швидкодія), а також особливості використовуваного програмного забезпечення. Ефективними виявилися системи, що вдало поєднують переваги обох підходів, оскільки відмова від допоміжних баз даних спричинює ускладнення структури лінгвістичних алгоритмів, водночас спрощення алгоритмів обертається незручностями при роботі з занадто громіздкими, розгалуженими допоміжними базами даних.

Вихідним модулем систем АПТ (АСОТ) є модуль ***автоматичного морфологічного аналізу тексту (АМА)***. Внаслідок його здійснення 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. Здебільшого це слова, належні до різних граматичних категорій у межах окремих частин мови. Ю.М.Марчук, один з провідних у колишньому Радянському Союзі фахівців у галузі комп'ютерної лінгвіс-

[110] **Марчук Ю.Н.** Основы компьютерной лингвистики. – М., 2000. – С.43.

тики, зокрема, машинного перекладу, запропонував виділяти такі типи АМА залежно від характеру його лінгвістичного забезпечення та способу "розпізнавання" морфологічної структури тексту[111]:

1) АМА зі словником основ слів;
2) АМА зі словником словоформ;
3) АМА методом логічного множення;
4) АМА за допомогою таблиць

Найбільш поширений завдяки своїй інформативності на сьогодні перший тип АМА, який ґрунтується на достатньо показових для лексикону певної мови словнику основ (для флективних слов'янських мов він налічує по кілька сотень тисяч основ), а також на допоміжних таблицях з правилами формальних перетворень основ та їхньої сполучуваності з окремими флексіями. АМА другого типу – зі словником словоформ – виявився придатним для опрацювання текстів мовами з бідною морфологією, або, інакше кажучи, з обмеженою чи стандартизованою варіативністю форми слів у процесах словозміни, слово- або формотворення. Третій тип АМА орієнтований на використання словника основ з автоматичним зняттям за допомогою процедури логічного множення омографії флексії. Цей тип АМА розробив на початку 60-х років минулого століття ленінградський математик С.Я.Фітіалов[112]. Процедура логічного множення, запропонована С.Я.Фітіаловим, полягає у визначенні функцій, реалізованих у слові окремими графемами флексій, та встановленні таких граматичних характеристик, що є спільними для всіх складників такої флексії. Наприклад, Ю.М.Марчук демонструє хід виконання процедури логічного множення в системі С.Я.Фітіалова на прикладі встановлення граматичних характеристик російського слова ***столом***, виражених двографемною флексією **–ом**[113]**.** Після пошуку у словнику основ встановлено, що слово ***стол-ом*** містить основу **стол-** та флексію **-ом** і належить до граматичного класу іменників. Аналіз функціонування графеми **м** в іменниках російської мови дозволив встановити такий спектр її реалізації: орудний відмінок однини іменників чоловічого та середнього родів (***род-ом, сел-ом***), давальний та орудний відмінки множини іменників всіх трьох родів (відповідно, ***род-ам, сел-ам, сестр-ам*** та ***род-ами, сел-ами, сестр-ами***). З'ясування спектру реалізації в іменниках другої графеми аналізованої флексії **-о-** та зіставлення (логічне множення) спільних та від-

[111] **Марчук Ю.Н.** Зазнач. праця. – С.45.

[112] Зацікавлених проблемами формальної морфології читачів відсилаємо до праці **С.Я.Фітіалова**: О построении формальной морфологии в связи с машинным переводом // Тезисы конф. по обработке информации, машинному переводу и автоматическому чтению текста. – М., 1961.

[113] **Марчук Ю.Н.** Зазнач. праця. – С.47-48.

мінних функцій дозволяє встановити, що флексія **–ом** властива граматичним підкласам орудного відмінка однини іменників чоловічого та середнього родів. Цей приклад яскраво демонструє й недоліки такого типу АМА. Як бачимо, завдяки йому не вдається усунути омографію значень чоловічого та середнього родів.

Хід реалізації модуля АМА в системах АПТ (АСОТ) розглянемо докладніше на прикладі системи **А**втоматичного **Г**раматичного **А**налізу **Т**ексту (**АГАТ**), створеної українськими лінгвістами – співробітниками відділу структурно-математичної лінгвістики Інституту мовознавства ім.О.О.Потебні НАН України для опрацювання наукових текстів російською та українською мовами[114]. Як матеріал для пробної версії цієї системи було обрано російськомовні тексти науково-технічних рефератів з кібернетики та обчислювальної математики, оскільки їм була властива достатньо стандартизована морфологія, що могла правити за основу системи АМА. Крім того, як пояснили свій вибір самі автори системи АГАТ, у промислових системах застосовують компромісний варіант модуля АМА, що поєднує тип АМА зі словником основ з типом безсловникового АМА. При цьому "словники використовують при аналізі слів, що зустрілися раніше (у попередній вибірці при створенні словника), а нові слова аналізують без словника слів за аналогією. Враховуючи специфіку реферативного тексту – інформувати про нові дослідження, можна було припустити, що з надходженням нових текстів у рефератах виявиться багато незафіксованих у словнику слів. Крім того, оскільки систему опрацювання тексту замислено як дослідницьку, її цілі включали перевірку гіпотези про ефективність АМА без словника слів, тому що програмна реалізація такого шляху простіша"[115] Розробники системи АГАТ як стратегію її створення обрали безсловниковий, або "незалежний" тип АМА, запропонований свого часу московським дослідником Г.Г.Бєлоноговим[116]. У мовах флективного типу з розгалу-

[114] Ґрунтовний виклад стратегії створення цієї системи з докладним обговоренням окремих розв'язуваних дослідницьких завдань можна знайти у таких працях цього колективу науковців: **Морфологический** анализ научного текста на ЭВМ. – К., 1990; **Грязнухіна Т.О.** Лінгвістичне забезпечення автоматизованих систем управління // Мовознавство. – 1983. - № 5. – С.39-43; **Грязнухіна Т.О., Нікула М.В.** Система автоматичного морфологічного аналізу українського наукового тексту // Проблеми українізації комп'ютерів. – К., 1993. – С.42-46; **Олексієнко Л., Дарчук Н.** Лематизація парадигм іменників української мови // Там же. – С.62-65.

[115] **Автоматический** морфологический анализ текста // Использование ЭВМ в лингвистических исследованиях. – К., 1990. – С.74.

[116] Див. докладніше: **Белоногов Г.Г., Новоселов А.П.** Автоматизация процессов накопления, поиска и обобщения информации. – М., 1979; **Белоногов Г.Г., Калинин Ю.П., Поздняк М.Ф., Яфаева Г.М.** Алгоритм многоступенчатого мор-

женою системою словозміни, до яких належать російська та українська, інформація про граматичні значення зосереджена в кінці слова і формально виражена флексією або формотворчим суфіксом на зразок суфіксів дієприслівників **-учи/-ючи** або **-ачи/-ячи** чи суфіксом інфінітива дієслів **-ти**.

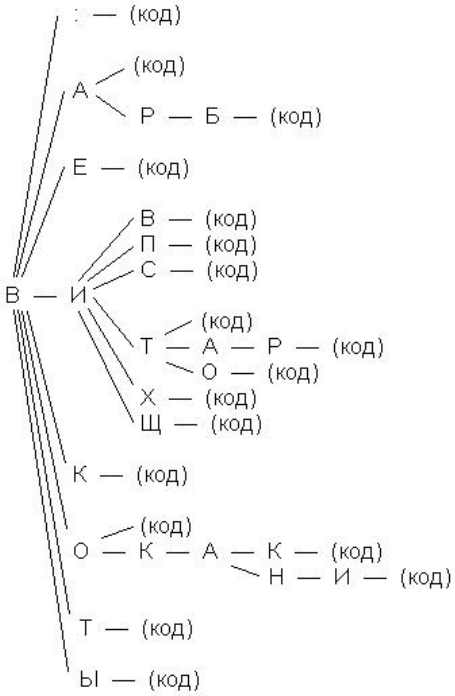


Рис. 17. Фрагмент алгоритму визначення граматичних класів слів у вигляді дерева квазіфлексій з кінцевою літерою **В**

На першому етапі дії системи АГАТ і виконувалася процедура аналізу слів за їхніми кінцівками. Цей етап здобув назву етапу ***флективного аналізу (ФА)*** тексту. Для його реалізації було створено допоміжну таблицю кінцівок слів, які дозволяли однозначно встановлювати граматичний клас слів тексту, або їхню частиномовну належність. Такі кінцівки одержали назву ***квазіфлексій***, оскільки власне флексія не завжди давала змогу однозначно встановити частиномовну належність слова і доводилося "рухатися вглиб" слова вліво (від його кінця до початку) доти, доки кінцівка дозволяла розв'язати таке завдання в межах окремого слова. Таким чином до власне флексій додавалися суфікси або їхні частини, корені та їхні частини, а то й слово в цілому виступа-

фологического анализа русских слов // Научн.-техн. информация. – Сер 2. - 1983. - № 1. – С.6-10.

ло в ролі квазіфлексії. За допомогою таблиці квазіфлексій, яка вміщує близько 20 тис. одиниць, після роботи модуля флективного аналізу АМА в складі системи АГАТ правильно визначено частиномовну належність 98% слів тексту (пробний масив для перевірки роботи цього модуля вміщував 140 тис. слововживань). Принцип використання списку квазіфлексій в межах процедури флективного аналізу тексту продемонструємо на прикладі сукупності квазіфлексій з кінцевою літерою **В** (див. рис. 17 і 18).

Як засвідчив аналіз пробного масиву тексту, найдовша квазіфлексія з цією літерою, що дає змогу однозначно встановити частиномовну належність слова, налічує 6 літер (***инаков*** – квазіфлексія короткого російського прикметника ***одинаков***). Слово в системі АМА трактується як послідовність літер від пробілу до пробілу. Отже, якщо після літери **В** іде пробіл, то такій текстовій одиниці присвоюється код (буквений символ) граматичного класу прийменників. Якщо зліва замість пробілу розташовано літеру, то залежно від її графічного втілення квазіфлексія з кінцевою **В** може бути 2–6-буквеною.

Квазі-флексія	ЛГК	Приклад	Кількість різних словоформ	Частота в текстах
:В	прийменник	в	1	3383
АВ	іменник	состав, глав, прав	3	37
БРАВ	дієприслівник	выбрав	1	1
ЕВ	іменник	деревьев, случаев	12	74
ВИВ	прислівник	добавив	1	1
ПИВ	дієприслівник	накопив	1	1
СИВ	іменник	курсив	2	9
РАТИВ	дієприслівник	сократив	1	1
ТИВ	іменник	коллектив, альтернатив	4	13
ОТИВ	прийменник прислівник	против, напротив	2	4
ХИВ	іменник	архив	1	1
ЩИВ	дієприслівник	обобщив	1	1
КВ	іменник	букв	1	6
ОВ	іменник	шагов, основ, останов, знаков	359	3148
КАКОВ	займенник-прикметник	каков	1	1
ИНАКОВ	короткий прикметник	одинаков	1	1
ТВ	іменник	средств	18	257
ЫВ	іменник	взрыв	1	1

Рис. 18. Фрагмент словника квазіфлексій лексико-граматичних класів слів (ЛГК) з кінцевою літерою **В**.

Наприклад, літера **А** в поєднанні з кінцевою літерою **В** дозволяє однозначно розпізнати іменники (***состАВ, глАВ, прАВ***), натомість літера **О** зліва від кінцевої **В** розпізнати іменники (***шагОВ, оснОВ, останОВ, знакОВ***), проте для визначення належності слова до граматичних класів займенників-прикметників та коротких прикметників необхідні, відповідно, 5-ти- (***КАКОВ***) та 6-буквені квазіфлексії (***одИНАКОВ***). Вибір тієї чи іншої квазіфлексії для визначення певного граматичного класу слів зумовлений так званим принципом переваги: квазіфлексії надають код того класу слів, переважну більшість яких така квазіфлексія описує. Так, квазіфлексії **АВ** та **ОВ** в сукупностях слів з такими кінцевими буквосполуками властиві переважно саме іменникам. У цьому можна переконатися, звернувшись до кількісних показників вживання слів з такими кінцівками, поданими у словнику квазіфлексій лексико-граматичних класів слів системи АГАТ (див. рис. 18).

Як бачимо, в апараті системи АГАТ поняття граматичного класу значно ширше за поняття частини мови. Оскільки комп'ютерна граматика на відміну від граматики традиційної має виразне операційне спрямування, то деякі лексико-граматичні розряди слів у межах традиційних 10 частин мови, а також деякі типи одиниць з огляду на їхні особливі функції в структурі тексту виділено в окремі граматичні класи. У системі АГАТ для російськомовних науково-технічних реферативних текстів таких класів 22. Наприклад, як окремі граматичні класи виділені займенники, що відмінюються за зразком іменників та прикметників, короткі та повні дієприкметники або прикметники, а також ініціальні або звукові абревіатури, формули, символи, цифри, власні імена, іншомовні невідмінювані слова. Текстові одиниці, що не потребують визначення граматичних характеристик, на зразок формул або цифр, вилучаються з опрацьовуваного системою АГАТ текстового масиву на етапі так званого ***доморфологічного аналізу***. Після виконання процедури ФА комп'ютер кожному слову в тексті присвоїв коди граматичних класів та граматичних підкласів. Код для змінюваних слів становить двобуквений символ: перша буква позначає граматичний клас, а друга – граматичний підклас слова. Незмінюваним словам, а також формулам, абревіатурам, цифрам присвоєно однобуквені символи, що вказують лише на їхній граматичний клас. Проте, як довели вищеподані приклади, флективний аналіз не в усіх випадках дає можливість однозначно встановити граматичні підкласи в межах граматичних класів слів. Так, наприклад, омографічними залишилися іменники ***состАВ, глАВ*** та ***прАВ***, що мають різні значення категорій відмінка, числа та роду, а також іменники ***деревьЕВ, случаЕВ, коллекТИВ, альтернаТИВ*** та ін.

Усунути омографію таких квазіфлексій дозволяє наступний етап АМА – етап ***контекстного аналізу (КА)*** словоформ тексту, тобто аналіз таких слів у певному текстовому оточенні за опорними точками. Вище у §6 розділу 1, присвяченому текстозорієнтованим базам даних,

вже йшлося про принцип роботи модуля АМА системи АГАТ на цьому етапі. Тут розглянемо його докладніше. Нагадаємо основні складники процедури контекстного аналізу. За опорні точки обираються такі слова або пунктуаційні знаки, які дають для певної словоформи так званий діагностичний контекст, тобто контекст, в якому вдається однозначно встановити граматичний підклас слова з омографічною квазіфлексією. Отже, на етап КА потрапляють ті слова, яким після виконання процедури ФА комп'ютер присвоїв коди 13 граматичних класів з неоднозначними показниками граматичних підкласів. Наприклад, для словоформ російських іменників, що мають у наукових текстах омонімію значень називного та знахідного відмінків, діагностичними виявилися контексти, в яких перед омонімічною словоформою (зліва від неї) в тексті розташовані такі опорні точки: дієслова з часткою **-ся**, короткі дієприкметники, кома з наступним підрядним сполучником, крапка, знак питання, крапка з комою. Якщо комп'ютер знаходить одну з перелічених опорних точок, то словоформі приписується значення називного відмінка, напр.: *предлагается* ***модель*** – им. (вин.)[117], ***Алгоритм предназначен*** – им. (вин.) ед.ч., *отсюда следует, что* ***описание*** *не является полным* – им. (вин.) ед.ч.[118].

Крім процедури аналізу контекстів за опорними точками, для усунення омонімії, наприклад, сполук прикметник+іменник, у цьому модулі АМА застосовано й процедуру логічного множення граматичної інформації аналізованого слова та слова з його оточення в тексті. Так, для встановлення граматичних підкласів слів в атрибутивній словосполуці ***цветовой гаммы*** спершу визначено граматичні класи кожного з цих слів. Зокрема, прикметник ***цветовой*** може мати граматичні значення називного й знахідного відмінків однини в чоловічому роді та родового, давального, орудного й місцевого відмінків однини в жіночому роді, а іменник ***гаммы*** – значення жіночого роду, родового відмінка однини та називного й знахідного відмінків множини. Зіставивши спектри функціонування цих двох слів, виявимо, що спільними для них є значення жіночого роду та родового відмінка однини. Розроблення модуля КА в системі АГАТ, крім суто практичного, має й важливий теоретичний вихід, оскільки в процесі його створення з'ясовано обсяг і типи омонімії, властивої словам різних частин мови, виявлено формальні засоби усунення омонімії, а також з'ясовано функціональне навантаження омографічних квазіфлексій в тексті. Так, найбільш відкритим для омонімії в сучасній російській мові виявився клас іменників. У досліджених науково-технічних текстах омонімічні іменники становлять 45–50% тек-

[117] У дужках підкреслено не реалізоване в цьому контексті граматичне значення словоформи.

[118] Приклади взято з: **Автоматический** морфологический анализ текста. – С.81.

стових слововживань. Іменникам властиві 22 типи омонімії. Найбільш різноманітною (9 різних типів) виявилася омонімія двох граматичних підкласів слів, пор.: 1) називний-знахідний відмінки однини (***узел, модель, слово***); 2) родовий-знахідний відмінки однини (***автора***); 3) називний-знахідний відмінки множини (***вопросы***); 4) родовий-знахідний відмінки множини (***читателей***); 5) давальний-місцевий відмінки однини (***виду, форме***); 6) орудний відмінок однини-давальний відмінок множини (***слепым***); 7) родовий-місцевий відмінки множини (***данных***); 8) називний відмінок однини жіночого роду-родовий відмінок однини чоловічого роду (***графа***); 9) давальний відмінок однини чоловічого роду-знахідний відмінок однини жіночого роду (***графу***)[119]. Результати застосування модуля КА засвідчили його високу ефективність – для майже 91% слів, що залишилися омонімічними після роботи модуля ФА, вдалося усунути омонімію граматичних підкласів. Ті 9% словоформ, для яких версія модуля КА, перевірена на пробному текстовому масиві, не дала можливості зняти омонімію граматичних підкласів, становлять цінний матеріал для вдосконалення алгоритмів КА, передусім для розширення інвентаря опорних точок.

Лінгвістичне забезпечення модуля АМА в системі АГАТ – алгоритмізовані правила визначення граматичних класів та граматичних підкласів слів – мають як суто практичне, прикладне, так і теоретичне значення для вивчення граматичного ладу мови. Вони складають один з основних розділів комп'ютерної граматики мови – ***комп'ютерну морфологію***. Завдяки своєму операційному характеру комп'ютерна морфологія унаявнює механізми побудови й функціонування тих чи інших одиниць мовної системи. Специфіку опису морфологічних процесів у комп'ютерній граматиці продемонструємо на прикладі процедури лематизації текстових одиниць, тобто процедури зведення формальних варіантів слова у тексті до його певного усталеного інваріанта – ***леми***, або канонічної (вихідної, словникової) форми слова в системі АГАТ (від лат. ***lētta*** "заголовок; тема твору")[120]. Підставою для створення такої автоматичної процедури стали результати роботи модуля автоматичного морфологічного аналізу тексту. Вихідним для дії процедури лематизації (автоматичного лематизатора) є текст, всім словам якого присвоєно коди граматичних класів та граматичних підкласів. Наприклад, всім іменникам української мови приписані коди граматичних підкласів роду, числа та відмінка. Для кожного роду іменників створено свої алгоритмічні правила лематизації, оскільки та сама за виглядом флексія по-різному "поводить себе", скажімо, в складі імен-

[119] Див. про це докладніше там же – С.80-84.

[120] Див. про це докладніше: **Олексієнко Л., Дарчук Н.** Лематизація парадигм іменників української мови // Проблеми українізації комп'ютерів. – К., 1993. – С.62-65.

ників чоловічого та жіночого родів. Наприклад, функціонування флексії **-ю** в іменниках чоловічого та жіночого роду призводить до різних змін в їхніх основах, а отже, для опису процесу лематизації таких словоформ необхідні різні алгоритмічні правила[121], пор.:

настро-ю →	**настрій, о/і, ю → й**
корен-ю →	**корінь, е/і, ю → ь**
лікт-ю →	**лікоть, 0/о, ю → ь**
молебн-ю →	**молебень, 0/е, ю → ь**

з

доповідд-ю →	**доповідь, дю → ь**
пісн-ю →	**пісня, ю → я**
матір'-ю →	**матір, 'ю → 0**

Таким чином, створена для автоматичної процедури лематизації комп'ютерна морфологія оперує правилами двох різновидів – для незмінних (константних) та змінних (варіантних) основ слів. Для кожного типу словозміни іменників перелік таких алгоритмічних правил не перевищує 15 позицій. Одержавши операційний опис словозміни сучасної української мови, втілений у графах (деревах) алгоритмічних правил, автори обговорюваної комп'ютерної морфології дійшли таких важливих теоретичних висновків: "Автоматична обробка словоформ прекрасно демонструє структурний характер парадигм як міні-систем мови. Субстанціальність мовної одиниці сама по собі для системи байдужа, лише в структурних відношеннях з фіналями основ вона стає одиницею системи. Якщо б це було не так, всі словоформи з тією самою флексією оброблялися б одним алгоритмічним правилом"[122]. Завдяки своїй структурній стрункості й послідовності виконання певних операцій комп'ютерна граматика, зокрема такий її складник, як комп'ютерна морфологія, відкриває перспективи для нових теоретичних пошуків та узагальнень.

Методи комп'ютерного моделювання словозміни, апробовані на матеріалі українських іменників, С.Лук'янчук застосував до створення автоматизованої моделі синтезу змінюваних форм українського дієслова[123]. Створена ним модель містить алгоритми аналізу (розпізнавання) парадигматичного класу дієслова за певною квазіфлексією, яку автор назвав ***графемним ідентифікатором парадигматичного класу***, та синтезу (породження, генерування) парадигми дієслова на

[121] Більшість прикладів узято з поданої вище праці Л.А.Олексієнко та Н.П.Дарчук.

[122] **Олексієнко Л., Дарчук Н.** – С.65.

[123] **Лук'янчук С.** Комп'ютерна модель парадигматичних класів дієслів // Українське мовознавство. – 2000. – вип.22. – С.82-85.

основі встановленого парадигматичного класу, дія якого полягає у поєднанні визначених основ інфінітива та теперішнього часу з відповідними флексіями. Наприклад, графемним ідентифікатором парадигматичного класу для дієслова ***базувати*** є буквений ланцюжок **–увати**, а для дієслова ***атакувати*** – буквосполука **–ати**. Роботу цієї моделі забезпечує спеціальна база даних, до складу якої ввійшли таблиці графемних закінчень для форм різних граматичних категорій дієслова, шаблони утворення квазіоснов, тобто в цілому частин слів без флексій, списки дієслівних префіксів, які у випадках збігу графемних ідентифіторів дозволяють однозначно встановити парадигматичний клас дієслова. При виробленні правил алгоритмів враховано як морфонологічні, так і акцентні характеристики дієслівних форм. Специфіку пропонованої комп'ютерної моделі С.Лук'янчук цілком слушно вбачає в її спрямуванні "не на вирішення певних прикладних задач (як використовуються подібні моделі в системах машинного перекладу, автоматичної перевірки орфографії тощо)", а в її функціонуванні як дослідницького інструменту лінгвіста. За допомогою цієї моделі, на думку автора, можна представити в систематизованому вигляді інформацію про утворення дієслівних парадигматичних структур та забезпечити зручний спосіб використання цієї інформації, встановлення кореляцій, пошук винятків тощо[124]. Цьому сприяє і відкритий, доступний для використання в різних дослідницьких цілях характер представлення інформації в базі даних обговорюваної моделі. Більшість аналогічних моделей, на думку С.Лук'янчука, забезпечує користувачів лише остаточним результатом у вигляді синтезованої парадигми та номеру парадигматичного класу. Саме такою, наприклад, є модель синтезу словоформ, розроблена Г.Г.Бєлоноговим та його колегами. В моделі С.Лук'янчука "користувач має повний доступ до всієї внутрішньої структури класу, яка представлена в зручному для нього та зрозумілому вигляді"[125]. На вхід алгоритму розпізнавання парадигматичного класу надходить дієслово у формі інфінітива. За допомогою циклічного порівняння зі списком графемних ідентифікаторів парадигматичних класів комп'ютер встановлює номер такого класу в базі даних і передає розпізнане дієслово з таким приписаним йому номером на вхід наступного алгоритму синтезу парадигми дієслівних форм. За номером парадигматичного класу вхідного дієслова за правилами цього алгоритму комп'ютер визначає номер процедури утворення двох можливих квазіоснов такого дієслова – у формі інфінітива та теперішнього часу. Такі квазіоснови формуються з урахуванням морфонологічних та акцентних змін, можливих при створенні форм цього парадигматичного класу дієслова. На

[124] Там же. – С.84.
[125] Там же. – С.85.

наступному кроці роботи алгоритму комп'ютер до певної квазіоснови додає відповідні флексії. Наприклад, для дієслова ***атакувати*** квазіоснову теперішнього часу становитиме **атаку-**, до якої й додається відповідний набір квазіфлексій особових форм, тобто таких частин флексій, які дозволяють однозначно розпізнати ту чи іншу словоформу: **-ю, -єш, -є, -ємо, -єте, -ють**. Таким чином, алгоритмічні процедури комп'ютерної морфології побудовані за принципом рекурсивності: правила лематизації при зворотному порядку їхнього виконання обертаються на правила генерування текстових форм слів. Отже, процедури лематизації та генерування становлять дзеркальне відображення одне одного. Запропонована комп'ютерна модель аналізу та синтезу українських дієслів, як і відповідна модель розпізнавання й породження українських іменників, здатна виконувати крім дослідницької, навчальну та інформаційно-довідкову функцію.

🕮 Терміни

- **лінгвістична інтелектуальна комп'ютерна система** – організована сукупність процедур для моделювання розумової діяльності людини в процесі розв'язання теоретичних або практичних завдань опрацювання мовної інформації
  - **система автоматичного перероблення тексту (АПТ) (=автоматизована система опрацювання тексту (АСОТ) –** лінгвістична інтелектуальна комп'ютерна система для виконання процедур морфологічного, синтаксичного та логіко-семантичного аналізу тексту
    - **лінгвістичне забезпечення АПТ, або АСОТ –** сукупність баз даних або знань та лінгвістичних процесорів для їхнього опрацювання, які становлять основу та інструмент лінгвістичного аналізу тексту
      - **словникова стратегія створення лінгвістичного забезпечення АПТ, або АСОТ** – підхід, зорієнтований на представлення відомостей про будову, значення та функціонування мовних одиниць у допоміжних лінгвістичних базах даних (словниках, графах, деревах, зведеннях правил)
      - **безсловникова, або «незалежна» стратегія створення лінгвістичного забезпечення АПТ, або АСОТ** – підхід, зорієнтований на представлення відомостей про будову, значення та функціонування мовних одиниць у вигляді алгоритмічних правил
  - **комп'ютерна граматика** – сукупність алгоритмів аналізу тексту на морфологічному, синтаксичному та логіко-семантичному рі-

внях його будови, втілених у таблицях, графах, матрицях, зведеннях правил

- **комп'ютерна морфологія –** частина комп'ютерної граматики, яка моделює творення форм слова (словозміну) або окремих слів (словотворення)

- **модулі АПТ, або АСОТ** – компоненти лінгвістичного процесора, призначені для аналізу тексту на окремих рівнях його будови й розуміння
  - **автоматичний морфологічний аналіз тексту (АМА) –** аналіз граматичної будови тексту, під час якого для кожної одиниці тексту визначають її граматичний клас та підкласи
    - **граматичний клас одиниць тексту** – сукупність одиниць тексту, належних до однієї частини мови, одного лексико-граматичного розряду в межах частини мови або взагалі будь-яка сукупність однорідних за певними ознаками одиниць тексту
      - **граматичний підклас одиниць тексту –** сукупність слів у межах одного граматичного класу зі спільними словозмінними характеристиками
    - **доморфологічний аналіз** – етап, на якому в тексті визначають одиниці, для розпізнавання властивостей яких не потрібні процедури АМА
    - **флективний аналіз (ФА)** – етап АМА тексту за флексіями або квазіфлексіями слів
      - **квазіфлексія (=графемний ідентифікатор)** – кінцева буквосполука (флексія, флексія з частиною основи або слово в цілому), яка дозволяє визначити граматичний клас та підкласи слова
    - **контекстний аналіз (КА)** – етап АМА тексту за діагностичними контекстами слова, визначеними опорними точками
      - **опорна точка** – розділовий знак, слово або їхня певна комбінація, які дозволяють усунути омографію квазіфлексій і однозначно визначити граматичний підклас слова

§4. Автоматичний синтаксичний аналіз тексту (АСА)

- Стратегії аналізу синтаксичної будови тексту
- Графічні способи представлення результатів АСА
- Метод безпосередніх складників
- Граматика залежностей

Автоматичний синтаксичний аналіз (АСА) становить другий важливий модуль систем АПТ, або АСОТ, а його лінгвістичне забезпечення складає другий необхідний компонент комп'ютерної граматики – ***комп'ютерний синтаксис***. Він спрямований на виявлення в тексті синтаксичних структур та їхнє формалізоване представлення. У комп'ютерній лінгвістиці розрізняють кілька типів АСА залежно від сфери його застосування, вихідних елементів та способів виконання. За першою ознакою розрізняють ***універсальні***, або ***глобальні*** системи та системи ***часткові***, придатні для розв'язання окремих дослідницьких завдань, наприклад аналізу текстів певної структури та певної предметної галузі. За другою ознакою найбільш ефективними виявилися системи АСА, що встановлюють синтаксичні структури в тексті за частинами мови словоформ та за їхніми синтаксичними ролями, тобто за членами речення. Нарешті за третьою ознакою виділяють системи АСА з ***безперервним*** та ***циклічним (повторюваним)*** переглядом тексту. Перші орієнтовані на один перегляд тексту, під час якого для кожного слова встановлюються його синтаксичні зв'язки з іншими словами в тексті, другі передбачають під час одного перегляду встановлення тільки одного типу синтаксичних одиниць або одного різновиду синтаксичних зв'язків, наприклад, виділення лише підметів або лише слів з узгоджувальним зв'язком (атрибутивних словосполук). Оскільки ці типи систем АСА залежно від способу здійснення процедури перегляду тексту виділяють синтаксичні одиниці різної складності, то їх ще називають ***інтегральними*** та ***локальними***. Інтегральні системи як результат передбачають одержання всієї синтаксичної структури речення, локальні – лише якоїсь частини такої структури.

Розрізнення інтегральних та локальних систем АСА прямо пов'язане з різною стратегією здійснення в таких системах процедури розкладу тексту на мінімальні синтагми – пари слів, пов'язані певним типом синтаксичного зв'язку: координації (між членами предикативної пари – підметом та присудком), узгодження, керування або приляган-ня. Локальні системи при цьому застосовують процедури методу безпосередніх складників, або аналізу контактних слів у реченні, розроблені представниками американської дескриптивної лінгвістики. Інтегральні системи використовують процедури граматики залежностей, розроблені представниками генеративної лінгвістики і спрямовані на

виявлення в тексті головного й залежного слів безвідносно до їхньої позиції в реченні[126].

Кожен з названих методів виконання АСА має свої графічні способи представлення синтаксичних структур речення. У локальних системах результати АСА представляють у вигляді скобових (дужкових) записів пар безпосередніх складників або записів з допомогою стрілок. В інтегральних системах АСА також використовують записи зі стрілками, але найпоширенішим є представлення синтаксичних структур у вигляді орієнтованого графа – ***дерева залежностей*** між словами в реченні та між реченнями. Для розроблення систем АСА застосовувалися різні стратегії, серед яких найбільш ефективними виявилися чотири: 1) послідовний аналіз тексту; 2) передбачувальний аналіз; 3) методика опорних точок та 4) методика фільтрів. Всі ці методики, як довели діючі системи АСА, мають взаємодоповняльний характер. Так, методика послідовного аналізу тексту й виявлення синтаксичної структури представлених у ньому речень передбачає створення словника еталонів словосполук (синтагм), записаних у термінах граматичних класів слів. Методика передбачувального аналізу ґрунтується на наборах синтаксичних передбачень – гіпотетичних (ймовірних, можливих) у певних типах речень типів синтаксичних структур, синтаксичних функцій окремих слів. Її розвитком є методика опорних точок, яка для слів з певними характеристиками визначає типові контексти, що діагностують вживання слова з тією чи іншою синтаксичною функцією в разі його багатофункціональності. Методика фільтрів дозволяє завдяки встановлюваним обмежникам на вживання, сполучуваність або переміщення слів у реченні з усього набору інформації про певні слова виявити інформацію, релевантну саме для аналізованого тексту.

Принцип роботи системи АСА продемонструємо на прикладі відповідного модуля уже обговорюваної у зв'язку з проблемами АМА системи АГАТ. Цей модуль становить часткову, інтегральну систему з безперервним переглядом тексту. Основою для створення пробної версії модуля були тексти російських науково-технічних рефератів з програмування та прикладної математики. Результати аналізу синтаксичної структури речень представлено на виході роботи цього модуля у вигляді дерева залежностей. У системі АГАТ модуль АСА в своїй роботі спирається на результати роботи модуля АМА, тобто на його вході перебуває текст, в якому для кожного слова визначено граматичний

[126] Читачів, зацікавлених познайомитися з методами структурного аналізу синтаксичної структури речення, відсилаємо до оглядових праць з цієї проблематики: **Основные** направления структурализма. – М., 1964; **Апресян Ю.Д.** Идеи и методы структурной лингвистики. – М., 1966. – С.232-252; **Современная** американская лингвистика: Фундаментальные направления / Под ред. А.Е.Кибрика, И.М.Кобозевой и И.А.Секериной. – М., 2002. – Изд. 2-е, испр. и доп.

клас та граматичні підкласи. В процедурах виконання АСА для поділу тексту на речення, а речень – на їхні складники використовуються опорні точки – розділові знаки (крім тире) та сполучники. Крім того, під час поділу тексту на речення або їхні частини послідовно зліва направо (за напрямком розгортання тексту) нумеруються всі слова в ньому, що після АМА одержали коди граматичних класів та підкласів. Отже, алгоритмічні правила модуля АСА зорієнтовано на частини речень або речення, що розташовані між певними опорними точками. З допомогою окремих правил встановлюються такі типи зв'язків між членами речення: виділення 1) пари "підмет+присудок"; 2) пари з граматичним зв'язком узгодження; 3) пари з дієслівним керуванням; 4) пари з іменним керуванням. Хід здійснення цих процедур і способи представлення виявлених синтаксичних структур речення покажемо на прикладі аналізу двоскладного, розповідного речення ***Уперше (1) запропонована (2) модель (3) словозміни (4) українських (5) іменників (6)***. Здійснення цієї процедури можна розділити на 4 етапи залежно від встановлюваних типів синтаксичних зв'язків між членами цього речення:

- Виділення пари слів зі зв'язком узгодження:
  5 ← 6 (***українських іменників)***.
- Виділення пари з дієслівним керуванням:
  1 ← 2 (***уперше запропонована***).
- Виділення пари з іменним керуванням:
  3 → 4 (***модель словозміни***),
  4 → 6 (***словозміни іменників***)
- Виділення пари підмета з присудком, інакше кажучи, комбінації групи підмета з групою присудка:
  2 → 3 (***запропонована модель***).

Одержану синтаксичну структуру можна зобразити у вигляді такого дерева залежностей (див. рис. 19):

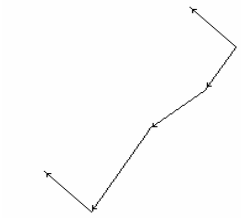


Рис. 19. Графічне представлення синтаксичної структури речення у вигляді дерева залежностей

Стрілковий запис складників синтаксичної структури цього речення подано вище в ілюстраціях до етапів виконання процедури АСА. Залежно від того, аналізуємо ми речення, тобто розкладаємо його на складники, чи синтезуємо його, тобто з окремих складників збираємо ціліс-

ну синтаксичну структуру, змінюється послідовність появи складників. Таку протилежність між аналізом та синтезом речення унаочнює скобовий (дужковий) запис його синтаксичної структури (див. рис. 20):

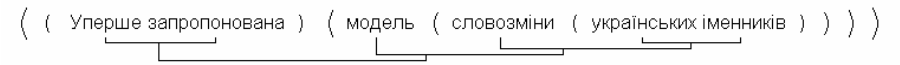


Рис. 20. Скобовий (дужковий) запис синтаксичної структури речення

В обговорюваній системі АГАТ модуль АСА на сьогодні здатен аналізувати всі типи синтаксичних структур, як прості, так і ускладнені та складні речення. Залежно від типу речення модуль працює в 4 етапи, на кожному з яких виконує такі процедури[127]:

І-ий етап – Сегментація тексту на відрізки, обмежені опорними точками: розділовими знаками, крім тире, та сполучниками. Всередині виділених сегментів за кодами граматичних класів та підкласів визначають членів предикативної пари. Виявлені в таких сегментах вставні слова та конструкції передають для аналізу спеціальному алгоритму.

ІІ-ий етап – Внутрішньосегментний аналіз. Він передбачає здійснення двох основних операцій: 1) формування предикативних синтагм зі зв'язком координації, або, інакше кажучи, встановлення в межах сегментів слів, пов'язаних безпосередніми синтаксичними зв'язками з членами предикативної пари, та 2) приєднання до сегментів з предикативними парами в межах речень сегментів, які таких пар не мають.

ІІІ-ій етап – Відокремлення в тексті простих речень, або речень з однією предикативною парою (предикативним центром) від складних речень, або речень з двома і більше предикативними парами (центрами). Ця процедура для аналізу складних речень передбачає виконання трьох операцій: 1) визначення структурного типу речення; 2) встановлення меж частин речення з предикативними парами (центрами), або його предикативних частин та їх нумерація; 3) визначення типу синтаксичного зв'язку між предикативними частинами складного речення.

ІV-ий етап – Аналіз всередині простого речення або предикативних частин складного речення. На цьому етапі роботи модуля АСА передбачено виконання чотирьох операцій: 1) виявлення прислівних зв'язків; 2) аналіз відокремлених зворотів – напівпредикативних синтаксичних одиниць; 3) з'ясування функціонального навантаження розділових знаків; 4) зняття омонімії граматичних форм, яка залишилася після роботи модуля АМА (ФА та КА) та перших трьох етапів роботи модуля АСА.

Вихідний продукт виконання всіх чотирьох етапів роботи модуля АСА в системі АГАТ і становить графічне зображення синтаксичної

[127] Читачів, зацікавлених ближче познайомитися з роботою цього модуля системи АГАТ, відсилаємо до його докладного опису в монографії: **Синтаксический** анализ научного текста на ЭВМ. – К., 2000.

структури речень тексту у вигляді орієнтованого графа – дерева залежностей між словами речень з визначеними для кожного слова синтаксичними ролями в реченні й типами синтаксичних зв'язків у ньому з іншими словами.

🕮 Терміни

- **автоматичний синтаксичний аналіз тексту (АСА) –** аналіз граматичної будови тексту, спрямований на встановлення типів синтаксичних структур речень, визначення синтаксичних ролей в них окремих слів та з'ясування типів синтаксичних зв'язків як між словами всередині речень, так і між реченнями
  - **універсальна (= глобальна) система АСА –** система АСА, зорієнтована на встановлення синтаксичних універсалій і призначена для аналізу будь-якого типу текстів певної мови
  - **часткова система АСА** – система АСА, яка розв'язує окремі дослідницькі завдання на матеріалі текстів певної підмови (мови окремої предметної галузі) або певної будови
  - **локальна система АСА** – система АСА, яка зорієнтована на "розпізнавання" певних типів синтаксичних одиниць або певних фрагментів тексту
  - **інтегральна система АСА –** система АСА, спрямована на "розпізнавання" синтаксичної будови тексту в повному обсязі
  - **система АСА з безперервним переглядом тексту** – система, яка за один перегляд тексту визначає його синтаксичну будову
  - **система АСА з циклічним переглядом тексту** – система, яка налаштована на цикл переглядів тексту, кожен з яких спрямовано на виділення окремого типу синтаксичних одиниць або структур тексту
  - **АСА за безпосередніми складниками речення –** метод встановлення синтаксичної структури речення за контактними словами, що пов'язані певним синтаксичним зв'язком
  - **АСА за зв'язком залежності між словами –** метод встановлення пов'язаних синтаксичним зв'язком слів незалежно від їхнього контактного чи дистантного розташування в тексті
  - **комп'ютерний синтаксис –** розділ комп'ютерної граматики, який моделює будову речень та складних синтаксичних цілих
  - **методики встановлення синтаксичної структури речення –** підходи до аналізу та графічного представлення синтаксичних зв'язків між членами речення та окремими реченнями
    - **методика послідовного аналізу –** аналіз речення з допомогою словника еталонів словосполук (синтагм), записаних як послідовності граматичних класів та підкласів слів

  - **методика передбачувального аналізу –** аналіз речення з допомогою наборів синтаксичних передбачень – гіпотетичних для певних типів речень синтаксичних структур
  - **методика опорних точок –** аналіз речення за формальними показниками його сегментів, в межах яких реалізовано певний тип синтаксичного зв'язку між словами
  - **методика фільтрів –** аналіз речення з допомогою наборів обмежень на вживання в синтаксичних структурах певних слів або на їхню сполучуваність чи переміщення в ньому
- **графічне представлення синтаксичних структур** – спосіб графічного зображення моделі синтаксичної структури
  - **скобовий (дужковий) запис** – запис пар слів без показу напрямку залежності між ними за допомогою скоб (дужок)
  - **стрілковий запис** – запис пар слів за допомогою стрілок, які вказують на напрямок залежності між словами
  - **орієнтований граф (= дерево залежностей)** – графічна модель у вигляді графа, вузли якого містять окремі слова речення, а ребра зі стрілками показують типи синтаксичних зв'язків між словами та напрямок залежності між ними

§ 5. Автоматичний логіко-семантичний аналіз тексту

- Методики визначення в тексті ключових слів (слів-концептів)
- Автоматичне індексування тексту
- Різновиди систем інформаційного пошуку (ІПС): документальні та фактографічні
- Інформаційно-пошукові мови (ІПМ): класифікаторні та дескрипторні
- Інформаційно-пошукові тезауруси (ІПТ)

Автоматичне опрацювання текстової інформації на всіх рівнях аналізу будови тексту – морфемному, морфологічному або синтаксичному – неминуче пов'язане з проблемами розпізнавання змісту одиниць тексту. Зрештою саме на унаочнення логіко-семантичної структури тексту, інакше кажучи, на "видобуття" змісту тексту й спрямовані всі системи його опрацювання. На думку Ю.М.Марчука та інших фахівців у галузі автоматичного опрацювання текстової інформації, сучасна технологія аналізу тексту все більше спирається на семантико-, а не синтаксико-зорієнтований підхід. "Тісний зв'язок синтаксиса з семантикою не дає можливості одержати ефективні системи та алгоритми синтаксичного аналізу у відриві від змісту висловлювання та тексту в цілому"[128]. Коло

[128] **Марчук Ю.Н.** Зазнач. праця. – С. 74.

теоретичних та практичних завдань, так чи інакше пов'язаних з проблемами встановлення змісту тексту, надзвичайно широке й різноманітне. Далі спробуємо коротко схарактеризувати ті напрямки розроблення проблеми автоматизації логіко-семантичного аналізу тексту, на яких на сьогодні досягнуто значного поступу й одержано результати, важливі не лише для практики суспільного життя, а й для розвитку самої лінгвістики, як комп'ютерної, так і традиційної. Результати опрацювання цієї проблеми формують третій необхідний компонент комп'ютерної граматики мови – ***комп'ютерну семасіологію***.

Розпізнавання змісту тексту становить важливу ділянку в системах так званого ***інформаційного пошуку***, якому, у свою чергу, передує процес ***індексування*** текстів, або їхнього розмічення за типами вміщеної в них інформації. Різновид таких систем становлять, наприклад, бібліотечні або архівні каталоги чи біобібліографічні інформаційно-довідкові системи різних установ та відомств, автоматизовані інформаційно-довідкові служби. Залежно від того, чи предмет пошуку становлять об'єкти дійсності (факти), чи описи таких об'єктів (фактів) – документи різної будови (здебільшого реферати або патенти), ***інформаційно-пошукові системи (ІПС)*** поділяють на ***фактографічні*** та ***документальні***. Кожна інформаційно-пошукова система має спеціальну мову доступу і роботи з нею – ***інформаційно-пошукову мову (ІПМ)***. Такі ІПМ можуть становити логічну класифікацію понять тієї предметної галузі, фактів або документів якої стосується інформаційний пошук. Прикладом ***ІПМ класифікаційного типу***, або ***ІПМ-класифікацій*** є відомі всім читачам універсальна десятична класифікація (УДК), бібліотечно-бібліографічна класифікація (ББК) або система міжнародних стандартних номерів книги (ISBN, International Standard Book Number). Спеціальні коди кожної з цих мов, які ми знаходимо на будь-якому різновиді друкарської продукції (книжці, брошурі, журналі, збірці чи газеті), і становлять інвентар її одиниць, використовуваних для індексування книжкових потоків за предметними галузями та дотичними до них поняттями.

Крім універсальних ІПМ-класифікацій, є ІПМ цього типу, зорієнтовані на роботу ІПС з текстами певної предметної галузі, тематики, тобто ІПМ-класифікації спеціального призначення. Таку оригінальну ІПМ класифікаційного типу розробили автори двотомного "Словаря славянской лингвистической терминологии", виданого у Празі в 1977 р[129]. Цей словник подає 2266 сучасних лінгвістичних термінів-понять всіма слов'янськими і трьома західноєвропейськими мовами (англійською, французькою та німецькою). Лінгвістичні терміни в ієрархічному дереві – основі цієї класифікації – розподілено за 9 предметними галузями: I.

[129] **Словарь** славянской лингвистической терминологии: В 2 тт. – Praha: Academia, 1977.

Загальні поняття; II. Звуковий бік мови; III. Графічний бік мови; IV.Словниковий склад; V. Частини мови; VI. Структура слова; VII. Синтаксис; VIII. Стиль та IX. Нові лінгвістичні напрями і методи. У межах кожної з цих галузей окремі терміни детально описано за додатковими, властивими їм ознаками. Подекуди така класифікаційна схема може містити 7 рівнів деталізації поняттєвої структури вихідного, базового для певної лінгвістичної галузі терміна, або 6 додаткових ознак, які уточнюють його зміст. Для позначення місця того чи іншого терміна в ієрархії понять вироблено систему спеціальних цифрових кодів. Ось, приміром, як представлені в цьому словнику-тезаурусі терміни на позначення різноманітних мовних засобів спілкування. Ці терміни перебувають в створеній укладачами словника ієрархії на 2-му рівні деталізації. Всі вони належать до галузі основних понять і тому містять цифровий код 1. Спільна для всіх них змістова ознака "засіб спілкування" в класифікації понять здобула цифровий код 5. Отже, коло всіх 23 вихідних понять цієї групи стоять тричленні цифрові коди: 1, 5 і порядковий номер терміна – назви конкретного мовного засобу спілкування. До порядкового номера, в свою чергу, можуть додаватися цифрові коди, які вказують на додаткові змістові ознаки, за якими його вміщено на відповідному місці в ієрархії лінгвістичних понять. Нижче подаємо перелік українських лінгвістичних термінів на позначення різновидів мовних засобів спілкування, представлених у "Словнику слов'янської лінгвістичної термінології", з відповідними цифровими кодами.

природна мова	1-5-1
штучна мова	1-5-2
міжнародна мова	1-5-3
світова мова	1-5-4
допоміжна мова	1-5-5
національна мова	1-5-6
племінна мова	1-5-7
одномовність, монолінгвізм[130]	1-5-8
двомовність, білінгвізм	1-5-9
тримовність	1-5-10
багатомовність, полілінгвізм	1-5-11
мовний контакт	1-5-12
інтерференція	1-5-13
<котериторіальні мови>[131]	1-5-14
усна мова	1-5-15
письмова мова	1-5-16
<друкована мова>	1-5-16-1
літературна мова	1-5-17
розмовна мова	1-5-17-1

[130] Через кому подано терміни-варіанти, або терміни-абсолютні синоніми.

[131] У ламаних дужках у цьому словнику подано вузькоспеціальні й малопоширені терміни.

мова художньої літератури	1-5-17-2
койне, спільна мова	1-5-18
старописемна мова[132]	1-5-19
молодописемна мова[133]	1-5-20
- (відсутній український еквівалент до поняття "мова культури")	1-5-21
<диглосія>	1-5-22
нелітературна форма мови, нелітературна мова, <позалітературна форма мови>	1-5-23
діалект 1, територіальний (місцевий) діалект	1-5-23-1
наріччя, діалектна група	1-5-23-1-1
діалект 2, говір 1	1-5-23-1-1-1
мікродіалект, говір 2, говірка	1-5-23-1-1-1-1
інтердіалект	1-5-23-2
мова міста, міська мова	1-5-23-3
жива розмовна мова, розмовно-побутова мова	1-5-23-4
просторіччя[134]	1-5-23-5
професійна мова	1-5-23-6
соціальний діалект	1-5-23-7
сленг	1-5-23-7-1
арго	1-5-23-7-2
дитяча мова	1-5-23-8

Максимальну (6 кроків) деталізацію лінгвістичного поняття (до 7-го рівня ієрархії) демонструють у цій групі терміни на позначення нелітературної форми мови:

1-ий рівень (код 1)	мова
2-ий рівень (код 5)	мовний засіб спілкування
3-ій рівень (код 23)	нелітературна форма мови
4-ий рівень (код 1)	діалект
5-ий рівень (код 1)	наріччя
6-ий рівень (код 1)	говір
7-ий рівень (код 1)	говірка

Іншим способом унаочнення змісту в системах інформаційного пошуку є виділення в текстах так званих ***ключових слів***, або ***слів-концептів***. Такі слова в конденсованій формі виражають основну інформацію про зміст тексту. Для їх позначення використовують спеціальні одиниці – дескриптори, а тому й самі ІПМ такого типу одержали назву ***дескрипторних***. ІПМ-класифікації та ІПМ дескрипторного типу не заперечують, а доповнюють одна одну. Мови дескрипторного типу більше прив'язані до текстів конкретної предметної галузі або тематики, а тому виявляють більшу гнучкість та ефективність у процесу авто-

[132] З поясненням "Мова з давньою писемною традицією".
[133] З поясненням "Мова народностей, що одержали писемність у новітній період".
[134] З поясненням "Широковживані (не діалектні) елементи, які не входять до літературно-мовної норми".

матичного аналізу їхнього змісту. З поняттям дескриптора ми вже стикалися з вами раніше при обговоренні підходів до укладання автоматичних семантичних словників (див. §5 розділу 1, присвячений проблемам комп'ютерної лексикографії, особливо обговорення принципів укладання "Російського семантичного словника" за редакцією Ю.М.Караулова). В ІПМ дескриптори можуть становити окремі слова, словосполуки або й частини слів, які виражають засадничі для окремої предметної галузі поняття. Для упорядкування інвентаря дескрипторів, а також уніфікації позначення понять у кожній ІПС створюється спеціальний ***інформаційно-пошуковий тезаурус (ІПТ)***, який складає лінгвістичне забезпечення (lingware) такої системи. Дескриптори в такому ІПТ упорядковують на основі не лише парадигматичних, а й синтагматичних відношень. Саме завдяки урахуванню останніх в ІПТ увиразнюють відношення так званої квазісинонімії, або контекстної синонімії, коли дескриптори на позначення певних понять зближуються лише в текстах, що стосуються окремої предметної галузі або певної проблемної ситуації в її межах. Так, наприклад, в лінгвістичних текстах з проблеми автоматичного синтаксичного аналізу синонімізуються дескриптори ***головне слово*** і ***"хазяїн"*** або ***залежне (підпорядковане) слово*** і ***"слуга"***, що в текстах, наприклад, присвячених устрою суспільного життя або відношенням між представниками різних суспільних верств не виявляють змістової близькості. Крім того, до ІПТ потрапляють і так звані ***асоціативні дескриптори***, тобто слова, що можуть виявляти лише опосередковану семантичну близькість у певних комунікативних ситуаціях. Скажімо, дескриптор ***детство*** в уже згаданому вище "Російському семантичному словнику" за редакцією Ю.М.Караулова перебуває в опосередкованих, асоціативних зв'язках з дескрипторами ***наив*** та ***неразум***, що позначають поняття ***наивность*** та ***неразумность***, пор. такі визначення, як ***дитячий погляд на речі (=наївний)*** та ***поводитися по-дитячому, як дитина (=нерозумно, нерозсудливо)***. Здебільшого розробники ІПТ відбивають в них такі відношення між дескрипторами: 1) "рід-вид" (***відмінок-номінатив (називний в.), генетив (родовий в.)*** тощо); 2) "частина-ціле" (***граматика – морфологія, синтаксис***); 3) "причина-наслідок" (***послаблення семантичних зв'язків у гнізді – розпад (деетимологізація) гнізда, формування кількох нових гнізд***); 4) "об'єкт-його типова функція" (***текстовий редактор – орфографічний контроль тексту***).

Процедура пошуку інформації в ІПС здійснюється в режимі "запит-відповідь". "Запит" на пошук інформації містить спеціальний ***пошуковий образ документа (ПОД)***, який створюють вручну або за допомогою комп'ютера, індексуючи (розмічаючи) текст з допомогою одиниць певної ІПМ, зокрема дескрипторів. "Відповідь", або ***пошуковий припис (ПП)*** на такий "запит" становить певним чином упорядкована суку-

пність дескрипторів, які описують певну проблемну ситуацію або предметну галузь у цілому в ІПТ системи. Після порівняння ПОД та ПП користувач ІПС одержує всі документи певної бібліотеки, архіву або взагалі будь-якого інформаційного масиву, зміст яких відповідає вміщеним у ПОД та ПП дескрипторам або одиницям (наприклад, кодам) мовкласифікацій. При цьому основними вимогами до ПОД та ПП є ***повнота*** та ***точність видачі інформації***. Першу обчислюють як відношення кількості спільних одиниць у ПОД та ПП до загальної кількості одиниць у ПП. Друга вимірюється в цілому відповідністю ПОД і ПП (кількості їхніх одиниць, їхнього вираження, характеру зв'язків між ними). Чим вищі параметри повноти й точності інформаційного пошуку, тим менший у такій системі показник ***інформаційного шуму***, або неправильно виданої у відповідь на запит інформації.

Для усунення інформаційного шуму застосовують методики індексування тексту, які враховують комунікативну значущість та функціональне навантаження слів у ньому. Одну з таких методик виділення в тексті ключових слів на основі процедур сіткового моделювання лексики розробив український дослідник Е.Ф.Скороходько[135]. Нагадаємо читачам, що в семантичній сітці слова впорядковуються залежно від того, які вони мають семантичні складники (компоненти) або деривати якого іншого слова вони виступають (див. про це докладніше у §2 цього розділу). Отже, можна припустити, що найбільше функціональне навантаження в тексті матимуть слова, що містять найбільшу кількість семантичних складників або з них можна вивести найбільшу кількість семантичних дериватів, Таким словам у тексті під час індексування приписують найбільшу вагу, або ранг. Наприклад, слово-родова назва, або гіперонім, одержить більшу вагу (вищий ранг), ніж слово-видова назва (гіпонім), пор. гіперонім ***приголосний (консонант)*** і гіпоніми ***африката, вібрант, сонант.*** Проте залежно від завдань інформаційного пошуку вимоги до змістового ранжування одиниць тексту можуть змінюватися. Е.Ф.Скороходько так ілюструє цей висновок: "Наприклад, для фахівців, які обговорюють особливості таксі порівняно з автобусом, значущість слів ***таксі*** та ***автобус*** вища, ніж слова ***автомобіль***"[136]. Найефективнішими для потреб індексування тексту виявилися гнучкі методики встановлення ключових слів, які поєднують різні функціональні властивості слів: їхню частоту (абсолютну, середню й відносну), комунікативну значущість, силу зв'язків з іншими словами в тексті (словотвірних, синтаксичних, асоціативних) тощо.

---

[135] Див. докладніше про це у праці: **Скороходько Э.Ф.** Семантические сети и автоматическая обработка текста. – К., 1983, а також у: **Сетевое** моделирование лексики // Использование ЭВМ в лингвистических исследованиях. – С.144-156.

[136] **Сетевое** моделирование лексики. – С.153.

Цікавий підхід до індексування лінгвістичних текстів реалізували в своїй праці польські дослідниці З. Руднік-Карватова та Х.Карпінська. Проіндексувавши тексти авторефератів мовознавчих дисертацій, вони уклали "Словник ключових слів славістичного мовознавства"[137]. У цьому словнику близько 2500 термінів з різних галузей сучасної славістики впорядковано за абеткою, між ними встановлено парадигматичні відношення (переважно родо-видові та відношення за ознакою "частина-ціле") та синтагматичні (синонімічні) відношення. Наприклад, для родового терміна – ключового слова лінгвістичних текстів ***мова*** (пол. ***język***) укладачі словника встановили 90 видових термінів – ключових слів текстів, присвячених окремим лінгвістичним проблемам. Серед таких видових ключових слів назви різновидів мови за походженням або належністю до певної групи чи родини за генеалогічною класифікацією мов, напр.: ***білоруська, верхньолужицька, давньоруська, українська, хорватська, церковнослов'янська мова*** (пол. ***białoruski, górnołużycki, staroruski, ukraiński, chorwacki, cerkiewnosłowiański język***), за типом будови мови: ***аналітична, силабічна, синтетична, флективна мова*** (пол. ***analityczny, sylabiczny, syntetyczny, fleksyjny język***), за сферою суспільного життя, яку обслуговує мова, напр.: ***наукова, сакральна, розмовна мова, мова реклами, мова політики*** (пол. ***naukowy, sakralny, potoczny język, język polityki, reklamy***) тощо. Синонімічні відношення, або відношення рівноправності в ієрархії ключових слів виявили такі пари лінгвістичних термінів, як ***мова засобів масової інформації*** (пол. ***język środków masowego przekazu***) – ***мова медіа, мова масмедіа*** (пол. ***język mediów***) або ***мова етнічна*** (пол. ***język etniczny***) – ***етнолект*** (пол. ***etnolekt***). Дібрані польськими дослідницями ключові слова досить детально й повно структурують інформаційне поле сучасної лінгвістики. Спирання на них в процесі пошуку повинно забезпечити високий ступінь повноти й точності одержуваної інформації.

Великого поширення набули сьогодні й різноманітні методики так званого ***контент-аналізу***, або аналізу змісту тексту (від англ. **content** "зміст") за певними ***концептуальними змінними***, що позначають центральне поняття аналізованого тексту. Активно такі методики логіко-семантичного аналізу тексту застосовують останнім часом в дослідженнях з політичної лінгвістики, напрямку мовознавчих досліджень, що вивчає мовні механізми формування громадської свідомості, впливу на громадську думку. Кожна концептуальна змінна вивчається у різноманітних зв'язках зі своїми текстовими ***мовними корелятами***. Ось, приміром, які мовні кореляти до концептуальної змінної ***держава (Україна)*** на основі аналізу українських газет різного політичного

[137] **Rudnik-Karwatowa Z., Karpińska H.** Słownik słów kluczowych językoznawstwa slawistycznego. – Warszawa, 1999.

спрямування та різного часу видання виявила молода українська дослідниця І.І.Брага[138]: ***наша держава, наша країна, Вітчизна, велика європейська держава, незалежна Українська держава, банкрот, зла мачуха, "Титанік", зменшена копія "Титаніка"*** тощо. Такі номінації формуються й розвиваються у межах певних моделей найменування, зокрема таких, як **ВІЙНА, РОДИНА, СПОРТ, ПРИРОДА, МИСТЕЦТВО, ТРАНСПОРТ, МЕДИЦИНА.** І.І.Брага ілюструє функціонування згаданих мовних корелятів у мові преси сучасної України такими прикладами, як: ***війна за владу, політична війна без правил, війна компроматів, таємна війна технологій із застосуванням зброї з політичного арсеналу; Хвороба держави, яка почалася раніше, загострилася і прогресує, а "медична картка" України поповнюється новими записами: тромби транспортних артерій, судомні конвульсії властей, метастази мафіозності і корумпованості, передвиборна пропагандистська сверблячка, передвиборна епідемія, економічний колапс***[139]***.*** Якщо концептуальні змінні мають універсальний характер і завдяки цьому можуть відігравати роль еталонів у різноманітних типологічних дослідженнях, то в їхніх реалізаціях у системах окремих мов – у їхніх мовних корелятах – яскраво відбито національно-культурні, історичні, етнопсихічні та соціополітичні конотації. Наприклад, досліджуючи функціонування концептуальних змінних – найменувань вищих державних правителів у творах О.С.Пушкіна, українська дослідниця Л.М.Захарова виявила такі мовні кореляти концепту "державний правитель" – найменування осіб Петра І та Наполеона Бонапарта: Петро І – ***суровый царь, государь, «на троне вечный был работник», герой Полтавы, великий человек, самодержавный великан, кумир, горделивый истукан – священный истукан, медный всадник – бронзовый всадник, мощный властелин судьбы – муж судьбы***; Наполеон Бонапарт – ***посланник провиденья, чудный муж, «в могучей дерзости венчанный исполин», «всадник, пред кем склонилися цари», «тяготеющий над царствами кумир», великий человек***[140].

Саме нові ефективні методики встановлення у тексті ключових слів дали можливість застосовувати в системах ІПС на противагу пошуку потрібної інформації за допомогою ІПТ метод так званого ***безтезаурусного пошуку***. Він передбачає роботу з масивами документів в

---

[138] Див. докладніше: **Брага І.І.** Мовна репрезентація образу держави у пресі України (кінець 70-х – початок 2000-х років). – Автореф. дис. ... канд.філол.наук. – К., 2002.

[139] **Брага І.І.** Зазнач. праця. – С. 10-11.

[140] **Захарова Л.М.** Національно-культурні конотації іменувань найвищих державних правителів (на матеріалі творів О.С.Пушкіна). – Автореф. дис. ... канд.філол.наук. – Сімферополь, 2000.

інтерактивному режимі з допомогою спеціальної діалогової системи, що дозволяє користувачеві створювати потрібні пошукові образи документів під час безпосереднього перегляду того чи іншого інформаційного масиву й залежно від типу опрацьовуваних документів вносити в такі пошукові образи необхідні корективи.

🕮 Терміни

- **автоматичний логіко-семантичний аналіз тексту –** модуль систем автоматичного аналізу тексту, результатом роботи якого є логіко-семантична структура тексту, певна модель його змісту
- **комп'ютерна семасіологія –** розділ комп'ютерної граматики, який моделює зміст окремих мовних одиниць та тексту в цілому
  - **інформаційно-пошукова система (ІПС) –** система опрацювання тексту з метою одержання з нього інформації про певні об'єкти або предметні галузі
    - **параметри ефективності інформаційно-пошукової системи (ІПС)** – характеристики якості роботи такої системи
      - **повнота видачі інформації** – ступінь відповідності обсягу виданої ІПС інформації про об'єкти або предметні галузі обсягу інформації про них, представленої в оброблюваному інформаційному масиві
      - **точність видачі інформації** – ступінь відповідності одержаної інформації запиту користувача
    - **автоматичне індексування тексту –** виділення в тексті різних типів інформації або інформації потрібного типу з допомогою спеціальних кодів
      - **пошуковий образ документа (ПОД)** – запит на пошук в інформаційному масиві потрібної інформації
      - **пошуковий припис (ПП)** – відповідь ІПС на запит користувача – певні документи або відомості, які містять інформацію потрібного користувачеві типу
      - **інформаційний шум** – неправильно видані на запит користувача документи або відомості
    - **інформаційно-пошукова мова (ІПМ)** – мова для спілкування з ІПС
      - **мова-класифікація** – мова, в основі якої лежить класифікація об'єктів або понять певної предметної галузі
      - **дескрипторна мова** – мова, яка спирається на дескриптори та логічні відношення між ними
        - **дескриптор (= ключове слово, слово-концепт)** – ключове поняття, концепт певної предметної галузі

- **тезаурусний метод** – метод пошуку інформації за допомогою тезаурусу
  - **інформаційно-пошуковий тезаурус (ІПТ) (=дескрипторний словник)** – словник з упорядкуванням дескрипторів за логічними (ієрархічними або асоціативними) відношеннями
- **безтезаурусний метод –** метод пошуку інформації в інтерактивному режимі роботи з комп'ютером без тезаурусу

- **контент-аналіз –** концептуальний аналіз логіко-семантичної структури тексту
  - **концептуальна змінна** – родове поняття, загальне ключове слово тексту
    - **мовний корелят –** видове поняття, ключове слово тексту – конкретизатор концептуальної змінної

§6. Системи машинного перекладу (МП)

- Сучасні стратегії створення систем МП
- Типи систем МП
- Мова-посередник (interlingua) та трансфер в системах МП
- Автоматизоване робоче місце (АРМ) перекладача

Моделювання інтелекту людини передбачає створення комп'ютерних моделей для будь-яких різновидів її мисленнєвої діяльності, що набувають суспільного значення. Серед цих завдань розроблення моделей перекладацької діяльності людини, відтворення продуктів її розумової діяльності, оформлених засобами однієї мови, засобами іншої мови посідає чільне місце в нинішній ситуації активної міжмовної та міжкультурної комунікації у світі. Особливого значення створення таких систем з україномовною компонентою набуває для піднесення престижу української мови у світі, повноцінного входження української держави і її мови у міжнародну комунікаційну мережу. Оскільки процедури розв'язання мовних проблем закладають теоретичну й практичну основу створення систем ***машинного перекладу***, останні з повним правом можна вважати різновидом лінгвістичних інтелектуальних систем. ***Машинний переклад (machine translation) (МТ)*** – це процес перетворення комп'ютером тексту, оформленого засобами однієї природної мови, в текст, оформлений засобами іншої природної мови. Моделює і автоматично здійснює цей процес комп'ютерний аналог такого різновиду розумової діяльності людини – система машинного перекладу.

Один з піонерів у галузі розроблення систем машинного перекладу в колишньому Радянському Союзі, керівник колективу учених м. Санкт-

Петербург (колишній Ленінград), який створив промислову систему китайсько-російського перекладу, що 1982 р. здобула Державну премію СРСР – професор Санкт-Петербурзького педагогічного університету імені О.І.Герцена Р.Г.Піотровський, оглядаючи шлях, пройдений цією галуззю комп'ютерної лінгвістики, поділив історію досліджень з МП на два періоди. Він назвав їх романтичною ерою та прозаїчним часом МП[141]. Очевидно, це закономірний шлях для реалізації будь-якої наукової ідеї: від захоплення нею, ейфорії від швидкого одержання за її допомогою нових знань про досліджуваний об'єкт або явище до планомірної, вдумливої практичної роботи над втіленням такої ідеї, критичної оцінки проблем і перешкод, що виникають на цьому шляху. Саму ідею комп'ютерного перекладу впродовж багатьох років піддавали нищівній критиці, адже навіть більш-менш прийнятні продукти дії таких систем неминуче вимагають людського втручання, постредагування різного ступеня складності: від поверхневого стилістичного шліфування до докорінного перероблення структури тексту та його лексичного наповнення. Проте системи МП продовжують розроблювати, більше того, вони на очах одного покоління (60–80-і роки минулого століття) перетворилися на комерційний продукт, тобто продукт купівлі-продажу. В чому секрет цього парадоксу? Слушну й вичерпну відповідь на це питання дає американський мовознавець Дж.Слокум у своєму огляді розробок з МП у США, Західній Європі та Японії: "Тим, хто користується системами машинного та людино-машинного перекладу (відповідно, МП та ЛМП), немає діла до академічних дебатів з приводу того, що слід розуміти під "високоякісним" та "повністю автоматичним перекладом". Для них суттєвими є лише два моменти: чи здатна система давати на виході результати, прийнятні за своєю якістю для тих цілей, для яких вони призначені (наприклад, для наступної редакторської правки), та чи є вся робота в цілому економічно рентабельною або, рідше, чи виправдана вона якимись іншими міркуваннями, скажімо, міркуваннями швидкості її виконання"[142]. Удосконалення систем МП здійснювалося саме в напрямку глибшого проникнення в суть співвідношення мислення і його вербалізації, мовного оформлення, пошуків гнучких моделей для відтворення цих процесів формально-логічними методами і засобами, приступними для комп'ютерного опрацювання.

---

[141] **Piotrovskij R.G.** MT in the former USSR and in the Newly Independent States (NIS). Prehistory, romantic era, prosaic time // Early Years in Machine Translation. Memoirs and Biographies of Pioneers. Ed. by W.John Hutchins. – Amsterdam/Philadelphia, 1995. - P.233-242.

[142] **Слокум Дж.** Обзор разработок по машинному переводу: история вопроса, современное состояние и перспективы развития // Новое в зарубежной лингвистике. Компьютерная лингвистика. Вып. XXIV. – М., 1989. – С. 360.

Романтична ера в історії МП пройшла під знаком розуміння мови як одного з різновидів математичного числення, однозначного й несуперечливого виведення одних одиниць з інших, створення жорстких і детальних моделей перетворення структур однієї мовної системи в структури іншої мовної системи. Таке однобічне і спрощене розуміння мовних процесів, породження тексту та його перетворення й завело в глухий кут дослідження, здійснювані в 60-і роки минулого століття, зокрема, колективом учених – лінгвістів і математиків – Ленінградського державного університету, очолюваного М.Д.Андреєвим, та колективом московських учених з Інституту математики ім. Стеклова та Інституту мовознавства АН СРСР, до складу якого входили О.Кулагіна та І.Мельчук. Жодна з цих систем не здатна була на виході дати більшменш прийнятні для редакторської правки результати. Крім того, завдяки деталізації правил, що їм підлягали процеси перетворення одних мовних структур в інші, такі системи відзначалися громіздкістю і повільністю, а отже, незручністю в роботі з ними, нерентабельністю. Разом з тим обидва ці колективи в процесі роботи над системами МП накопичили величезний дослідницький матеріал, який містив якісно нову важливу інформацію про структуру мовних об'єктів, відношення між ними та закономірності їхнього функціонування в різноманітних текстах. Наприклад, на основі багаторічної роботи над створенням мови-посередника в системах МП, так званої ***інтерлінгви*** **(interlingua)** М.Д.Андреєв виробив особливий метод вивчення текстової інформації, який ґрунтувався на врахуванні статистичних та комбінаторних властивостей морфологічних, лексичних та синтаксичних одиниць. Він назвав його ***статистико-комбінаторним методом***. За допомогою цього методу були успішно розв'язані конкретні дослідницькі завдання як теоретичного, так і практичного характеру в різних галузях мовознавства: словотворенні, граматиці, лінгвістиці тексту, дешифруванні давніх писемностей[143]. Такого ж важливого теоретичного значення набула і модель "Зміст↔Текст" ("Смысл↔Текст"), розроблена для систем МП І.Мельчуком, О.Жолковським та Ю.Апресяном[144].

Ця модель була призначена для створення правил переходу від глибинної (семантичної) синтаксичної структури повідомлення до її оформлення засобами конкретної мови – поверхневої синтаксичної структури. Її правила детально описували парадигматичні та синтагматичні відношення між компонентами синтаксичної структури – словами чи словосполуками. Для кожного з них встановлювався певний набір

[143] Див. докладніше про принципи і результати застосування цього методу у праці: **Андреев Н.Д.** Статистико-комбинаторные методы в теоретическом и прикладном языковедении. – Ленинград, 1967.

[144] Зацікавлені читачі одержать про неї докладнішу інформацію з праці: **Мельчук И.А.** Опыт теории лингвистических моделей "Смысл↔Текст". – М., 1974.

функцій, ролей, які вони здатні були виконувати в синтаксичній структурі. Розрізнялися два типи функцій залежно від відношень, в які між собою вступали слова: для відношень парадигматичних це була функція лексичної заміни, а для відношень синтагматичних – функція параметрів. Наприклад, функція **SYN** (синонімія) вказує на можливість заміни слова ***комп'ютер*** словами ***машина, автомат, робот, ЕОМ (електронно-обчислювальна машина), персональний комп'ютер (персоналка)***. А функція ***MAGN*** (інтенсивність) поєднує зі словом ***перемога*** такі прикметники, як ***остаточна, повна, беззаперечна, важка, піррова.*** Створена модель виконує в системі МП роль спеціального механізму перекодування мовних повідомлень, причому перекодування двоступеневого: спочатку в межах вхідної мови, або мови з якої перекладають, перекодування компонентів поверхневої структури тексту в компоненти її глибинної семантичної структури, а потім уже в межах мови вихідної, мови, на яку перекладають, перекодування компонентів глибинної структури в компоненти поверхневої структури цієї мови. Для співвіднесення двох мов – вхідної та вихідної – компоненти глибинної структури записують максимально формалізовано й відсторонено від засобів конкретної мови, а отже, цей модуль системи виконує роль своєрідної мови-посередника, мови-трансфера. Так, приміром, глибинну семантичну структуру дієслова ***дарувати*** в термінах моделі "Зміст↔Текст" можна записати як "А експліцитно спричинює наявність у В С", пор. з дефініцією цього дієслова в СУМі: "Передавати що-небудь у власність як подарунок (В)". Відчуженість метамови тлумачного словника від конкретного адресанта спричинила відсутність у цій дефініції реалізатора компонента А (адресант, відправник дії). Об'єкт дії залишився у СУМі поза межами дефініції, його позначено граматичною ремаркою *кому*. Кожен з компонентів цієї глибинної структури виступає у функції параметра дієслова ***дарувати***, або актанта цього предиката. Наприклад, компонент А – суб'єкт дії – може бути виражений узагальнено іменником ***даритель*** (пор. "Той, хто дарує") або назвами конкретних осіб, що дарують: ***гість, батьки, друзі, колеги, рідні, доля*** (у переносному значенні, пор. *доля подарувала щастя, талант*). Актант В реалізують лексеми з узагальненим значенням ***обдарений, обдарований*** та їхні конкретизатори ***іменинник, дитина, донька, син, людина, спільнота.*** І, нарешті актант С також реалізують лексеми з узагальненим значенням ***дар, дарунок, подарунок, обдарування, презент, сувенір*** та їхні конкретизатори-назви об'єктів, що виступають у цій ролі: ***квіти, парфуми, одяг, прикраси, гроші, дім, машина, маєток*** і метафоричні об'єкти на зразок уже згаданих вище ***щастя, талант*** і т.ін. Наявність еталонної глибинної структури дозволила з'ясувати спектр різних поверхневих структур, в яких її реалізують вхідна чи вихідна мова, або спектр можливих трансформа-

цій такої глибинної структури. Наприклад, подана вище глибинна структура "А експліцитно спричинює наявність у В С" може трансформуватися у такі різні за своєю будовою поверхневі структури: ***Друзі подарували іменинникуу магнітофон. Магнітофон був іменинникуу подарований друзями. Іменинник одержав від друзів магнітофон у подарунок.*** Тонка класифікація семантичних відношень між одиницями, правила переходів від глибинних семантичних структур до поверхневих (графічних або фонетичних) оформлень окремих мовних одиниць дали в руки лінгвістів надзвичайно важливі відомості про механізми вербалізації ("ословлення") розумовомовленнєвих процесів, відомості, вкрай потрібні для створення гнучких моделей штучного інтелекту. Крім того, ці відомості заклали надійне підґрунтя для розбудови функціональних та комунікативних граматик окремих мов, дали новий поштовх розвитку таких мовознавчих дисциплін, як когнітивна лінгвістика або лінгвістика тексту.

У КЛ вироблено класифікацію систем МП за такими ознаками:

- підхід до опрацювання тексту: ***прямий***, зорієнтований на певні вхідну та вихідну мови і в їхніх межах – на підмови, мови певних предметних галузей, та ***непрямий***, при якому зміст тексту вхідною мовою "розпізнається" засобами лінгвістичного аналізу незалежно від того, якою буде вихідна мова перекладу;
- спосіб зіставлення текстів вхідною та вихідною мовами: наявність ***мови-посередника (interlingua)*** – спеціальної формально-логічної мови, яка в експліцитному вигляді подає опис змісту мовних одиниць та відношень між ними у вхідному тексті, чи ***трансфера*** – спеціального модуля міжмовних перетворень, перезапису вхідної інформації, переносу (від значення етимона терміна – лат. **trānsfero** "переносити, перетворювати") кодів структури та лексичного наповнення вхідного тексту в коди тексту вихідної мови;
- спосіб опрацювання тексту вхідною мовою: ***локальний***, при якому точку відліку в роботі системи МП становить аналіз окремої одиниці – слова чи словосполуки, та ***глобальний –*** при якому точку відліку в роботі системи МП становить текст або його складові: абзаци, надфразні єдності, речення, в структурі яких аналізують окремі слова чи словосполуки;
- роль людини у здійсненні МП: ***системи машинного перекладу (МП)***, в яких людина виконує роль пре- чи постредактора, а самий процес перекладу здійснює комп'ютер, та ***системи людино-машинного перекладу (ЛМП) (machine-aided translation) (MAT)***, де людина втручається в самий процес перекладу в режимі "онлайн (on-line)";

- спосіб представлення інформації вхідною та вихідною мовами: у вигляді корпусів текстів та термінологічних банків даних (ТБД) на зразок чотиримовної бази даних СЛОВО, описаної в §5 розділу 1.

На думку Р.Г.Піотровського, історія створення систем МП в колишньому Радянському Союзі бере свій початок з середини 30-х рр. минулого століття , коли російський винахідник П.П.Смирнов-Троянський розробив метод механічного перекладу, який, утім, не знайшов тоді підтримки ані серед математиків, ані серед лінгвістів[145]. Робота над однією з перших діючих систем МП розпочалася в 1952 р. в Джорджтаунському університеті (США). У 1964 р. систему GAT (Georgetown Automatic Translation – Джорджтаунський Автоматичний Переклад) було передано в розпорядження Комісії з атомної енергетики США та в Євроатом – аналогічну наукову установу в Європі (в м. Іспра, Італія). Система GAT здійснювала переклад англійською мовою російських текстів з фізики. За своєю стратегією це була система з прямим локальним підходом до опрацювання вхідного тексту: "передбачалася просто послівна заміна вхідних слів на їхні перекладні еквіваленти, після якої здійснювалися нечисленні позиційні перестановки, що дозволяли одержати дещо віддалено подібне до тексту англійською мовою. Дуже скоро під "словом стали розуміти будь-яке окреме слово або послідовність слів, яка утворювала "ідіому"[146]. Незважаючи на примітивність реалізованого підходу до опрацювання текстової інформації, відсутність якісного лінгвістичного й програмного забезпечення систему GAT експлуатували в Євроатомі до 1976 р. Як видно, вона задовольняла потреби користувачів в оперативному ознайомленні із загальним змістом оброблюваних російських текстів з фізики.

Однією з перших радянських діючих систем МП була система ЕТАП (**Е**лектро**Т**ехнічний **А**втоматичний **П**ереклад) з версіями ЕТАП-1 (французько-російський переклад) та ЕТАП-2 (англо-російський переклад). Робота над її створенням розпочалася в 1974 р. в ІНФОРМЕЛЕКТРО і була продовжена в Інституті проблем передачі інформації Російської Академії Наук. Версію ЕТАП-1 було здано в експлуатацію в 1980 р.; версію ЕТАП-2 – у 1985 р. Відмітною рисою цієї системи є її потужна теоретична лінгвістична основа: стратегія її створення базується на поняттєвому і процедурному апараті моделі "Зміст↔Текст", оскільки одним із керівників цієї розробки був Ю.Д.Апресян, один з авторів моделі "Зміст↔Текст". У системі ЕТАП реалізовано непрямий, як локальний (послівний), так і глобальний (оснований на повному синта-

[145] **Piotrovskij Raimund G.** MT in the former USSR and in the Newly Independent States (NIS). Prehistory, romantic era, prosaic time. – P.233.
[146] **Слокум Дж.** Обзор разработок по машинному переводу: история вопроса, современное состояние и перспективы развития. – С.367.

ксичному аналізі тексту) підхід до опрацювання тексту. За способом зіставлення текстів вхідною та вихідною мовами – це система з трансфером – модулем міжмовних перетворень, який виконує інтегральна модель опису й представлення морфології, синтаксису та словника, реалізованих в текстах вхідної та вихідної мов. Систему ЕТАП, що й засвідчено в її назві, призначено для перекладу електротехнічних текстів та патентів[147].

Проблеми, що виникають при створенні лінгвістичного забезпечення систем МП, продемонструємо на прикладі промислової системи українсько-російського та російсько-українського перекладу ПЛАЙ, у створенні лінгвістичної компоненти якої активну участь взяли співробітники відділу структурно-математичної лінгвістики Інституту мовознавства ім. О.О.Потебні НАН України (Т.О.Грязнухіна, Н.П.Дарчук, Л.В.Орлова, В.І.Критська, Т.К.Пуздирєва та Т.І.Недозим). Система ПЛАЙ належить до різновиду людино-машинних систем МП, оскільки передбачає втручання користувача в разі потреби в процес перекладу тексту. Наприклад, система надає можливість користувачеві доповнювати представлені в ній словники новими словами, а також здійснювати попередню підготовку тексту, пропускаючи і не позначаючи відповідно в ньому, зокрема, слова з недопустимими для такого роду текстів символами або невідомі слова, видавати або не видавати можливі граматичні варіанти слів або варіанти їхнього перекладу (властиві їм еквіваленти). На сьогодні словникове оснащення системи ПЛАЙ дозволяє за її допомогою, попередньо налаштувавшись на відповідний словник, перекладати тексти із загальномовною лексикою, а також з лексикою юридичної, лінгвістичної, залізничної, технічної, хімічної та фізичної терміносистем, а отже, система ПЛАЙ реалізує обидва можливі підходи до машинного перекладу тексту: локальний, розрахований на мову певної предметної галузі, або підмову, та глобальний, орієнтований на мову в цілому без обмеження певною предметною галуззю. За способом опрацювання тексту ця система також подає широкі можливості, оскільки в режимі "Translate (перекладати)" дозволяє перекладати весь текст без виділення його складових, а в режимі "Word translation (переклад слова)" дає можливість перекладати окремі слова, а в разі потреби і доповнювати словники системи відсутніми в них словами і в такий спосіб усувати з тексту лакуни – не розпізнані і не перекладені слова. Система ПЛАЙ реалізує різні стратегії до опрацювання текстової інформації: прямого і непрямого, локального й глобального підходів. Завдяки морфологічному й синтаксичному аналізу вхідних текстів створюється спеціальний трансфер для міжмовних перетворень.

[147] Докладніший опис системи ЕТАП див. у праці: **Апресян Ю.Д. и др.** Лингвистическое обеспечение системы ЭТАП-2. – М., 1988.

Система ПЛАЙ дає на виході в цілому задовільні за якістю й повнотою переклади. Крім того, завдяки гнучким модулям морфологічного й синтаксичного аналізу тексту, в основу яких покладені відповідні алгоритми, створені цим же колективом лінгвістів для системи АГАТ і описані нами вище у §§3 і 4 цього розділу, а також зручним процедурам доступу до словникових баз даних та довідково-інформаційних баз система ПЛАЙ зручна в користуванні і має високу рентабельність. Проте лінгвістичне забезпечення системи ПЛАЙ ще потребує вдосконалення як у плані поповнення перекладних еквівалентів, врахування багатозначності слів, так і в плані уточнення моделей граматичної та лексичної сполучуваності слів. Для кращого розуміння цих проблем подамо читачам кілька прикладів перекладу текстів системою ПЛАЙ, що потребують редакторської правки:

вхідний текст: *Предметом обговорення в доповіді є комплекс завдань, посталих перед сучасним українським мовознавством у зв'язку з набуттям українською мовою в 1991 р. статусу мови нової незалежної багатонаціональної держави. Прихід української мови в нові для неї сфери економічного та політичного життя суспільства, раніше обслуговувані російською мовою, спричинив необхідність вироблення нового літературного стандарту та субстандартів української мови з огляду на зміни в умовах її функціонування, розширення кола користувачів, якісні зрушення в складі останнього. Успішне виконання поставлених завдань вимагає передусім нормалізації та кодифікації корпусу наявних мовних засобів, здійснюваних з принципово інших позицій (робота над новою редакцією українського правопису), створення серій різнотипних словників (тлумачних, перекладних, навчальних) та навчальної, інформаційно-довідкової літератури, орієнтованих на різномовну й різнопрофільну аудиторію, а також вироблення нової концепції гармонійного й гнучкого мовного планування.*

переклад російською: *Предметом обсуждения в доклада есть комплекс задач, **посталих** перед современным украинским языковедением в связи с обретением на украинском языке в 1991 г. статуса языка нового независимого многонационального государства. Приход украинского языка в новые для нее сферы экономической и политической жизни общества, ранее обслуженные на русском языке, послужил причиной необходимости изготовления нового литературного стандарта и **субстандартів** украинского языка, учитывая изменения в условиях его функционирования, расширения кола пользователей, качественные сдвиги в составе последнего. Успешное выполнение поставленных задач требует **передусім** нормализации и кодификации корпуса имеющихся языковых средств, осуществляемых из принципиально других позиций (работа над новой редакцией украинского правописания), создания серий разнотипных словарей (толковых, переводных, учебных) и учебной, информационно-справочной литературы, ориентированных на разноязычную и **різнопрофільну** аудиторию, а также изготовления новой концепции гармонического и гибкого языкового планирования.*

Різними шрифтами та підкресленням у тексті виділено одиниці, які з тих чи інших причин викликали в системи ПЛАЙ труднощі при перекладі. Жирним курсивом позначено слова, відсутні в базовому словнику системи, а отже, незнайомі їй. Підкреслено слова, вжиті в даному контексті в неправильному значенні (напр. укр. ***вироблення концепції*** – рос. ***выработка (разработка) концепции***, але ***изготовление продукции***, зміна об'єкта дії – фізичної чи розумової – вимагає й зміни російського еквівалента єдиного українського слова). Підкреслено в тексті конструкції з порушеннями в лексичній чи синтаксичній сполучуваності слів, пор.: укр. ***доти обслуговувані російською мовою*** – рос. ***ранее обслуженные на русском языке*** замість ***ранее обслуживаемые русским языком***, а також слова, перекладені неправильно, як-от слова-омографи укр. ***кола*** (від ***коло*** «круг») (користувачів, читачів, друзів) та рос. ***кола*** (від ***кол*** – укр. ***кіл*** «загострена палиця»). Інші помилки в поданому вище перекладі пропонуємо читачам проаналізувати самостійно.

Ми спробували перевірити якість зворотного шляху перекладу – переклали одержаний російський текст українською мовою з тим, щоб пересвідчитися, наскільки вхідний український текст буде відповідати вихідному тексту. І тут також виявилися проблеми, що спонукають до вдосконалення відповідних модулів системи ПЛАЙ. Надаємо читачам можливість проаналізувати текст цього перекладу самостійно:

переклад українською: *Предметом обговорення в доповіді є комплекс задач, посталих перед сучасним українським мовознавством у зв'язку зі знаходженням українською мовою в 1991 р. статусу мови нової незалежної багатонаціональної держави. Прихід української мови в нові для неї сфери економічного і політичного життя суспільства, доти обслуговані російською мовою, послужив причиною необхідності виготовлення нового літературного стандарту і субстандартів української мови, з огляду на зміни в умовах його функціонування, розширення кола користувачів, якісні зрушення в складі останнього. Успішне виконання поставлених задач вимагає передусім нормалізації і кодифікації корпуса наявних мовних засобів, здійснюваних із принципово інших позицій (робота над новою редакцією українського правопису), створення серій різнотипних словників (розумних, перекладних, навчальних) і навчальної, інформаційно-довідкової літератури, орієнтованих на різномовну і різнопрофільну аудиторію, а також виготовлення нової концепції гармонійного і гнучкого мовного планування*.

Пор. також переклади інших текстів системою ПЛАЙ:

вхідний текст: *Глубокоуважаемая г-жа N!*

Искренне тронута Вашим письмом, вниманием ко мне и моей работе, а также тем, что Вы сочли возможным включить ее в действующий фонд Российской Национальной библиотеки. Сердечно признательна Вам и Вашим коллегам за помощь в представлении моей книги читательской аудитории Санкт-Петербурга и рада, что она вызвала интерес читателей-специалистов.

Пользуясь случаем, наш авторский коллектив передает в Российскую Национальную библиотеку работу 1998 года – «Словарь аффиксальных морфем украинского языка», составленный с помощью компьютера (есть и его компьютерная версия). Надеемся, что Словарь заинтересует специалистов и всех, кто занимается украинским языком, поскольку на сегодня это наиболее полный инвентарь приставок и суффиксов современного украинского языка с разнообразной информацией об их формальных, содержательных и функциональных свойствах. Заранее глубоко признательны за Вашу помощь в ознакомлении российского читателя с нашей работой.

С неизменной готовностью к сотрудничеству и глубокой благодарностью за внимание и помощь

переклад українською: *Високошановна **г-жа** N!*

Щиро торкнута Вашим листом, увагою до мене і моїй роботі, а також тим, що Ви рахували можливим уключити її в діючий фонд Російської Національної бібліотеки. Сердечно вдячна Вам і Вашим колегам за допомогу в представленні моєї книги читацької аудиторії Санкт-Петербурга і рада, що вона викликала інтерес читачів-фахівців.

*Користуючись випадком, наш авторський колектив передає в Російську Національну бібліотеку роботу 1998 року – «Словник афіксальних морфем української мови», складений за допомогою комп'ютера (є і його комп'ютерна версія). Сподіваємося, що Словник зацікавить фахівців і усіх, хто займається українською мовою, оскільки на сьогодні це найбільш повний інвентар **приставок** і суфіксів сучасної української мови з різноманітною інформацією про їх формальні, **змістовні** і функціональні властивості. Заздалегідь глибоко вдячні за Вашу допомогу в ознайомленні російського читача з нашою роботою.*

З незмінною готовністю до співробітництва і глибокою вдячністю за увагу і допомогу

переклад російською: *Глубокоуважаемая **г-жа** N!*

*Искренне тронутая Вашим письмом, вниманием ко мне и моей работе, а также тем, что Вы **считали** возможным **уключити** ее в действующий фонд Русской Национальной библиотеки. Сердечно признательная Вам и Вашим коллегам за помощь в представлении моей книги читательской аудитории Санкт-Петербурга и **совет**, что она вызвал интерес читателей-специалистов.*

*Пользуясь случаем, наш авторский коллектив передает в Русскую Национальную библиотеку работу 1998 года – «Словарь аффиксальных морфем украинского языка», составленный с помощью компьютера (есть и его компьютерная версия). Надеемся, что Словарь заинтересует специалистов и всех, кто занимается на украинском языке, поскольку на сегодня это наиболее полный инвентарь приставок и суффиксов современного украинского языка с разнообразной информацией про их формальные, содержательные и функциональные свойства. Заранее глубоко **признательные** за Вашу помощь в ознакомлении русского читателя с нашей работой.*

С неизменной готовностью к сотрудничеству и глубокой благодарностью за внимание и помощь

Спершу останній текст було перекладено з налаштуванням на словник загальної лексики, підключення словника лінгвістичних термінів дозволило правильно перекласти російське ***приставка*** українським ***префікс***.

Система МП з її словниковими базами даних та знань, лексикографічними й текстовими процесорами для роботи з ними утворюють так зване ***автоматизоване робоче місце (АРМ) перекладача*** – різновид автоматизованого робочого місця будь-якого фахівця-користувача лінгвістичних інтелектуальних комп'ютерних систем. Подані приклади перекладів, здійснених системою ПЛАЙ, засвідчують досить високу якість її роботи, незважаючи на певні огріхи. А чинник часу, що пішов на одержання цілком прийнятного результату, ще більше переконує в доцільності її використання і підтверджує слушність згаданого вище міркування Дж.Слокума. Недоліки перекладу вказують на певні лакуни в лінгвістичному забезпеченні системи: неврахування різних еквівалентів слова вихідної мови до різних значень слова вхідної мови (пор. рос. ***толковый словарь*** та ***толковый человек*** з укр. ***тлумачний словник*** і ***розумна, тямуща людина*** або ***содержательные свойства*** (тобто ***свойства содержания***) та ***содержательная беседа*** (тобто ***беседа, наполненная содержанием***) з укр. ***змістові властивості*** та ***змістовна бесіда***); вибір неправильного відповідника у вихідній мові (пор. укр. ***набуття статусу, вироблення стандарту*** з рос. ***нахождение статуса, изготовление стандарта***), а також нерозрізнення так званих міжмовних омографів на зразок укр. ***кола користувачів*** (форма род.відм.одн. слова ***коло***) та рос. ***кола*** (форма род. відм.одн. слова ***кол*****,** пор. з такою ж формою його укр. відповідника ***кіл***). Інші виділені в тексті недоліки й помилки перекладів читачі при бажанні можуть проаналізувати без нашої допомоги.

Проте, без сумніву, подані вище переклади можна вважати задовільними, якщо прийняти загальну робочу оцінку якості будь-якого перекладу (здійсненого людиною чи комп'ютером). Дж.Слокум сформулював її так: "З погляду постредагування "добрим" первинним перекладом вважають той, який є сенс виправляти, тобто той, який редактор готовий спробувати в чомусь змінити, проте не збирається цілком його заперечувати або замінювати своїм власним первинним перекладом"[148]. Незаперечні переваги цієї системи пояснюються її спиранням на розвинені процедури автоматичного морфологічного та синтаксичного аналізу текстів вхідної та вихідної мов, описані в §§3 та 4 цього розділу, створенням деталізованого набору правил взаємопереходів, розгалужених, налаштованих як на загальнонародну мову, так і на мови окремих предметних галузей словників еквівалентів слів, словоспо-

[148] Там же. – С.399.

лучень, синтаксичних конструкцій зіставлюваних української та російської мов. Досвід якісних систем МП для флективних мов доводить, що майбутнє саме за такими системами, які завдяки граматичному й логіко-семантичному аналізу тексту окреслюють моделі функціонування тих чи інших мовних одиниць у текстах певної структури та тематики. Ширше проблема вдосконалення таких систем як різновиду систем штучного інтелекту полягає в створенні ефективних і достатньо повних та глибоких структур представлення знань про ту чи іншу предметну галузь та способів їхньої вербалізації засобами окремих природних мов, а в перспективі – створення комп'ютерних моделей "картини світу" для зіставлюваних природних мов. Інший аспект проблеми вдосконалення таких систем становить вивчення специфіки процесу перекладу як різновиду комунікації – процесу обміну інформацією та знаннями. У цій царині залишається без відповіді ще ціла низка важливих питань, зокрема, чи дійсно для якісного перекладу потрібні повне розуміння вхідного тексту, глибока обізнаність з предметом перекладу, чи становить процес перекладу просту послідовність "двох процедур: розуміння вхідного тексту і наступного відтворення його вихідною мовою"[149]. Ефективні сучасні системи МП довели важливість для якості перекладу механізму міжмовних операцій, правил перемикання одного мовного коду на інший, створенням якого і займаються лінгвісти.

🕮 Терміни

- **машинний переклад (МП) (machine translation) (MT) –** процес перетворення комп'ютером тексту, створеного засобами однієї природної мови, в текст, оформлений засобами іншої природної мови
  - **прямий підхід до МП –** підхід до перекладу тексту, зорієнтований на певні вхідну та вихідну мови і в їхніх межах – на підмови, мови певних предметних галузей
  - **непрямий підхід до МП –** підхід до перекладу тексту, при якому комп'ютер "розпізнає" зміст тексту вхідною мовою засобами лінгвістичного аналізу незалежно від того, якою буде вихідна мова перекладу
  - **мова-посередник (interlingua) –** спеціальна формально-логічна мова, яка в експліцитному вигляді подає зміст мовних одиниць та відношення між ними у вхідному тексті
  - **трансфер (=міжмовне перетворення, міжмовна операція) –** спеціальний модуль міжмовних перетворень, перезапису вхідної інформації, переносу кодів структури та лексичного наповнення вхідного тексту в коди тексту вихідної мови

[149] Там же. – С.405.

  - **локальний підхід до МП** – підхід, при якому точку відліку в роботі системи МП становить аналіз окремої одиниці – слова чи словосполуки
  - **глобальний підхід до МП –** підхід, при якому точку відліку в роботі системи МП становить аналіз структури тексту в цілому або його складників: абзаців, надфразних єдностей, речень
- **людино-машинний переклад (ЛМП) (machine-aided translation) (MAT) –** система машинного перекладу, в якій людина втручається в самий процес перекладу в режимі "он-лайн (on-line)"
- **термінологічний банк даних (ТБД) –** спосіб представлення термінологічної інформації вхідною та вихідною мовами у вигляді словникової бази даних
- **автоматизоване робоче місце (АРМ) перекладача** – різновид автоматизованого робочого місця будь-якого фахівця-користувача лінгвістичних інтелектуальних комп'ютерних систем, який складає система МП, її словникові бази даних та знань з лексикографічними та текстовими процесорами для роботи з ними

§7. Моделювання мовленнєвої діяльності в комп'ютерних діалогових системах

- Інтерфейс та його різновиди: штучно- і природномовні інтерфейси
- Діалогові (=питально-відповідні) системи
- Жорстка та м'яка форма діалогу з комп'ютером
- Графічні та звукові інтерфейси

Створення комп'ютера радикально змінило розуміння процесів комунікації, оскільки з'явився новий учасник таких процесів. Учасник, що не має людського інтелекту, свідомості, знань, манери вести бесіду у прийнятому в людському суспільстві трактуванні цих понять. Дослідження, спрямовані на розроблення штучного інтелекту, передбачали не лише моделювання розумової діяльності людини, а й створення комп'ютерних аналогів людської поведінки, зокрема, її можливих реакцій на ті чи інші повідомлення, стратегії людського спілкування як такого. Цей напрямок досліджень у межах комп'ютерної лінгвістики й, ширше, інформатики пов'язаний зі створенням різноманітних інтерфейсів – засобів спілкування людини з комп'ютером. На сьогодні виразно окреслилися два напрямки у розбудові таких засобів – ***штучномовні інтерфейси***, зорієнтовані на штучні мови (машинні мови, мови програмування) та ***природномовні інтерфейси***, зорієнтовані на використання засобів певної природної мови (англійської, російської, української тощо) з певною уніфікацією та стандартизацією лексичних та

граматичних елементів її системи. Такі обмеження наборів засобів природної мови полягають в уніфікованому доборі слів або словосполук, максимальному спрощенні синтаксичних конструкцій, обмеженні вживаних морфологічних форм слів. Усім користувачам, які послугуються, наприклад, операційною системою сучасних комп'ютерів Windows («Вікна»), відомі ті засоби ведення діалогу, які вона пропонує для виконання певних операцій. Це наприклад, різноманітні повідомлення про хід виконання того чи іншого завдання. Зокрема система Windows («Вікна») повідомляє користувачеві про хід запуску тієї чи іншої програми опрацювання інформації такими повідомленнями англійською, первинною мовою для цього інтерфейса, або іншими мовами, мовами його модифікацій (зокрема, російською чи українською): ***Windows is starting up*** (Windows стартує), ***Please wait*** (Будьте ласкаві, зачекайте), якщо виконання певної операції потребує часу, ***What do you want the computer to do?*** (Що ви хочете, щоб зробив комп'ютер?). В останньому випадку користувачеві пропонують перелік можливих операцій, з якого слід вибрати потрібне. Наприклад, у меню системи Windows («Вікна») подані операції ***Shut down*** (Вимкнути)***, Restart*** (Перезапустити)***, Stand by*** (Підтримувати, допомагати), кожна з яких передбачає виконання спеціальних процедур, забезпечуваних операційною системою Windows.

Серед таких засобів спілкування з комп'ютером є і спеціальні повідомлення, спрямовані на захист користувача і комп'ютера, повідомлення, що змушують людину ще раз перевірити правильність поставленого комп'ютеру завдання і пересвідчитися в його доцільності. Програмісти, жартуючи, запровадили навіть спеціальний термін на позначення систем з такими засобами перевірки правильності й доцільності виконуваних операцій: ***дурнестійкі системи*** (англ. ***fool-proof systems***). Сучасні операційні системи комп'ютерів максимально наближені до потреб користувачів-непрограмістів. Їхню взаємодію з комп'ютером полегшують різноманітні графічні, зображальні засоби, картинки-символи окремих операцій (так звані "іконки" від англ. ***icon*** "зображення, ілюстрація"), системи підказок або довідок (так звані "хелпи" від англ. ***to help*** "допомагати"). Істотним спрощенням таких засобів спілкування з комп'ютером є і створення національномовних модифікацій операційних систем. Наприклад, уже згадувана вище операційна система Windows, створена для персональних комп'ютерів американською корпорацією «Microsoft», уже не вимагає від користувачів знання англійської мови. На сьогодні вже працюють версії цієї операційної системи, налаштовані на понад 100 національних мов. Для таких світових мов, як англійська, французька або іспанська, враховано навіть їхні різні регіональні варіанти. Наприклад, для англійської мови таких варіантів в останній версії системи представлено 13, серед них не лише варіанти англійської мови, поширені у США, Австралії, Канаді, Новій

Зеландії, Ірландії та Південній Африці, а й такі екзотичні різновиди "піджин інглиш", як англійська мова Карибських островів, Зімбабве або острова Тринідад. Французька мова представлена 6 своїми регіональними варіантами, крім мови Франції, це варіанти, поширені у князівстві Монако, Бельгії, Швейцарії, Люксембурзі та Канаді. Для деяких мов враховано й варіанти з різними системами письма. Наприклад, для азербайджанської та узбецької мов представлені варіанти із записом слів засобами кириличної або латинської графіки. Є модифікації системи Windows («Вікна») для мов з ієрогліфічною писемністю (японської та китайської), а також для роботи з семітськими мовами (арабською, іврит), писемність яких передбачає розгортання тексту не зліва направо, як у мовах індоєвропейської родини, а, навпаки, справа наліво.

Однак користувачі комп'ютера стикаються з необхідністю вміти послугуватися й штучномовними інтерфейсами, що вимагають знань про роботу самої операційної системи. Математики-програмісти використовують у роботі з комп'ютером спеціальні мови програмування, що крім елементів природних мов (слів, що позначають команди-оператори на окремих кроках алгоритму виконання того чи іншого завдання), містять спеціальні символи, які також вказують комп'ютеру хід виконання певної процедури опрацювання інформації. Процес програмування вимагає від користувача певних знань про розміщення інформації в пам'яті комп'ютера, а також про ті ділянки робочої пам'яті комп'ютера, які обслуговують виконання певних заданих операцій. Проте й у цьому напрямку іде процес наближення засобів спілкування з комп'ютером до користувача, що не має спеціальних знань про будову або функціонування машини та її окремих блоків. Спостерігаємо розвиток і вдосконалення так званих ***мов програмування високого рівня*** (Сі$^{++}$, ПАСКАЛЬ), які на відміну від ***мов програмування низького рівня*** (наприклад, Асемблер, Бейсик), зосереджені на логіці виконання процедур завдання, а не на їхньому представленні у вигляді окремих машинних кодів або на детальному розписі операцій блоків комп'ютера. Інтерфейси, наближені до потреб певних класів користувачів, а також інтерфейси, налаштовані на взаємодію з користувачем, що не має спеціальних знань про апаратне, математичне чи програмне забезпечення комп'ютера, називають ***дружніми***.

Найяскравіше комп'ютерні моделі вербалізованої комунікативної діяльності людини втілені у так званих ***діалогових,*** або ***питально-відповідних системах***. Кожна з таких систем становить модель процесу спілкування, обміну інформацією, яка містить кілька блоків-аналогів складників такого процесу: блок аналізу повідомлення відправника інформації – адресанта, блок інтерпретації такого повідомлення одержувачем інформації – адресатом, блок формування змісту відповіді адресата та блок його представлення у формі, зрозумілій для адресанта. Кожен з цих блоків обслуговують спеціальні бази даних та

знань, а також алгоритми аналізу й синтезу інформації. Їхня структура, склад і деталізація процедур опрацювання інформації залежать від тих завдань, на виконання яких спрямовано такі діалогові системи. Виділяють два типи структури діалогових систем, або ***систем людино-машинної взаємодії***, ***систем спілкування людини з комп'ютером***. Перші мають стандартизовані, жорсткі правила побудови повідомлень (=запитів) і відповідей на них, обмежений набір засобів природної мови: лексичних (ключові слова та слова-зв'язки) та граматичних (усталені форми слів, типи будови словосполук та речень), що можуть вживатися під час діалогу, спеціальний формалізований запис інформації (символи, коди, умовні позначки тощо). Такі системи дістали назву ***систем з жорсткою структурою діалогу***. Саме таку будову мають, наприклад, системи інформаційного пошуку, діалогові системи роботи з базами даних або знань. Наприклад, діалогова система "МОРФОЛОГ", створена для спілкування з базою даних комп'ютерного морфемно-словотвірного фонду української мови Інституту мовознавства ім.О.О.Потебні НАН України (див. про неї у §3 розділу 1), передбачає жорстку структуру повідомлень для процедур коригування, доповнення або вилучення інформації з бази. Крім того, робота з цією системою вимагає від користувача знання правил формалізованого запису слів у базі даних.

Натомість системи, що здатні залежно від характеру одержуваної від відправника інформації змінювати структуру діалогу, доповнювати її новими даними або й взагалі змінювати форму ведення діалогу, називають ***системами з м'якою структурою діалогу***. До таких належать передусім експертні системи, в яких передбачені спеціальні блоки генерування нових сценаріїв для опису тих чи інших ситуацій спілкування або процесів у певних предметних галузях, а також процедури зміни й доповнення вже існуючих у пам'яті комп'ютера сценаріїв.

Створення діалогових систем стало сьогодні чи не найважливішим завданням комп'ютерної лінгвістики і інформатики в цілому, адже саме вони уможливлюють успішне розв'язання завдань з допомогою комп'ютера. Розроблення комп'ютерних моделей спілкування сприяє підвищенню ефективності не лише систем зі штучним інтелектом, а й набуває важливого значення для вивчення розумовомовленнєвої діяльності людини в різних її проявах. На думку учених, функціонування комп'ютерних моделей спілкування поставило перед фахівцями в галузі теорії та практики мовної комунікації цілу низку раніше не опрацьовуваних проблем. "До центру уваги, – стверджує А.М.Баранов, – потрапили питання, які раніше перебували на периферії теорії діалогу, дискурс-аналізу та теорії комунікації. Що забезпечує природність спілкування? Які умови зв'язності бесіди? Коли спілкування виявляється успішним? У яких випадках виникають комунікативні невдачі і чи можна

їх уникнути? Які стратегії спілкування використовують учасники комунікативної взаємодії для досягнення своїх комунікативних цілей?"[150] І цей список можна продовжити.

Моделювання різноманітних комунікативних ситуацій виявляє механізми побудови повідомлень і реакцій на них, дозволяє виявити певні комунікативні стереотипи, шаблони, усталені комунікативні формули, виявити в них загальнолюдські, загальномовні й етнокультурні, національномовні компоненти. Який потужний поштовх дослідженням з теорії і практики мовної комунікації дають комп'ютерні діалогові системи продемонструвала свого часу найвідоміша з таких систем "Еліза", створена Джозефом Вейценбаумом у 1966 р. Назва програми промовиста – Еліза Дулітл – героїня п'єси Бернарда Шоу "Пігмаліон", вчилася літературної англійської мови, як комп'ютер вчиться спілкуватися з людиною. Програма моделювала поведінку людини, що підтримує бесіду. Вона спиралася на стереотипи певних комунікативних ситуацій, алгоритми перетворення повідомлень на загальні відповіді, на певні процедури трансформації попереднього висловлення, а також на визначені в попередньому повідомленні ключові слова. В одному з варіантів програма "Еліза" моделює бесіду лікаря-психіатра з пацієнтом. Ось уривок з такого діалогу у нашому перекладі[151]:

Лікар. Кажіть.
Пацієнт. Мене хвилює погода.
Лікар. Як давно вас хвилює погода?
Пацієнт. Відтоді, як померла моя мати.
Лікар. Розкажіть більше про вашу родину.
Пацієнт. Вони всі в галузі комп'ютерів.
Лікар. Машини лякають вас.
Пацієнт. Ні.
Лікар. Будь ласка, не будьте таким лаконічним.
Пацієнт. Вибачте.
Лікар. Ви говорили про свою матір.

Навіть побіжний аналіз наведеного вище уривка діалогу дає можливість виявити недоречності або неточності у відповідях комп'ютера, що виступає в ролі пацієнта. Він не вміє розвинути задану тему діалогу через відсутність у його пам'яті відповідного сценарію такої комунікативної ситуації або неповноту такого сценарію, "неналаштування" на дане конкретне повідомлення або й комп'ютер взагалі відповідає "не до ладу", як хоча б у випадку відповіді на прохання ЛІКАРЯ (людини –

[150] **Баранов А.Н.** Зазнач. праця. – С. 23.

[151] Приклад подано у згаданій праці А.М.Баранова. – С.22 з посиланням на працю: **П.Уинстон.** Искусственный интеллект. – М., 1980. – С.391, 393. Подаємо його в українському перекладі.

учасника діалогу) розповісти докладніше про свою родину. Разом з тим такі недоречності можливі і при контакті з людиною, яка неуважно слухає співрозмовника або необізнана з темою бесіди, а також в ситуації, коли обом чи одному зі співбесідників нецікава розмова і вони "тягнуть час", обмінюючись беззмістовними репліками, як кажуть, "переливають з пустого в порожнє". У багатьох випадках ті, хто спілкувався з програмою "Еліза" не могли визначити, що їх співрозмовником є комп'ютер, а не людина. Наприклад, з 24 учасників експерименту, який проводив М.Майгайр та його група, 15 були переконані, що їм відповідала людина, 5 – вагалися і лише 4 були цілком певні, що спілкувалися з комп'ютером. На думку А.М.Баранова, програма "Еліза" "функціонувала настільки успішно, що фактично спростувала відомий тест Тьюринга на створення штучного інтелекта"[152]. Нагадаємо, що за умовами цього тесту, якщо людина, спілкуючись з машиною, не зможе помітити, що її співрозмовник – автомат, то можна визнати, що така машина мислить, тобто має інтелект, подібний до інтелекту людини.

За аналогією з програмою "Еліза" група працівників Київського національного університету імені Тараса Шевченка створила навчальну програму спілкування з комп'ютером "Елочка". Подаємо уривок з діалогу, сконструйованого комп'ютером за цією програмою[153]:

Л. (Людина). Що таке алгоритм?
К. (Комп'ютер). Ви вважаєте, цей алгоритм – найоптимальніший?
Л. Ні.
К. Ваша відповідь занадто однозначна – це більше личить машині. А Ви хто?
Л. Машина.
К. Я, здається, не зовсім Вас розумію. Що Ви цим хочете сказати?
Л. Нічого.
К. Ви занадто категоричні!
Л. Ви впевнені?
К. Що це Вам дає?
Л. Нічого.
К. Здається, ми втомили одне одного.
Л. Ні.
К. Ви впевнені?
Л. Ні.

Аналіз цього тексту також дає можливість виявити логічні непослідовності у побудові цього діалогу, наявні в ньому штампи, стеротипи відповідей, наприклад, однозначні відповіді на зразок "Ні" або "Нічого".

---

[152] **Баранов А.Н.** Зазнач. праця. – С. 21.

[153] Текст діалогу запозичено з праці А.В.Анісімова. Див.: **Анисимов А.В.** Компьютерная лингвистика для всех: Мифы. Алгоритмы. Язык. – К., 1991. – С.168. Подаємо текст у нашому перекладі з російської мови.

Явними стають і ті "скріпи", речення-відмовки, які об'єднують окремі питання й відповіді у змістове ціле – діалог на певну задану тему.

Подальше вдосконалення систем спілкування людини з комп'ютером полягає в поєднанні великого обсягу різноманітної інформації про різні предметні галузі, комунікативні ситуації, деталізації й специфікації таких відомостей і знань з максимальним спрощенням самої структури комунікативних актів (=ходів алгоритмів комунікації), її відкритістю до можливої зміни стратегії діалогу залежно від одержуваної комп'ютером нової інформації. Великою проблемою залишається й форма представлення повідомлень і відповідей на них, форма, приступна для користувача, що не має спеціальної математичної або програмістської підготовки. Безперечно, неабияк посприяють успішному розвитку таких діалогових систем детально випрацьовані логічні моделі знань (сценарії, фрейми, скрипти, плани тощо), налаштовані на різні типи комунікативних ситуацій і різні типи виконуваних у них комунікативних завдань. Разом з тим розробники лінгвістичного забезпечення комп'ютерних систем моделювання мовленнєвої діяльності повинні знайти зручну й придатну для ефективного застосування форму організації баз знань.

Мовленнєва діяльність, як відомо, здійснюється в двох формах: звуковій (усній) та графічній (письмовій). Досі при обговоренні проблем комп'ютерного моделювання процесів мовної комунікації йшлося про графічний (письмовий) різновид такої діяльності. Натомість з розвитком комп'ютерної техніки і появою так званих звукових карт (спеціальних пристроїв у комп'ютерах та програмного забезпечення, що їх обслуговує) з'явилися можливості для моделювання й усної (звукової) форми спілкування людини з комп'ютером. Великий і важливий внесок у розроблення цієї проблеми зробили українські вчені. В Україні над проблемою усного спілкування з комп'ютером почали працювати з середини 60-х років минулого століття. Перші дослідні теоретичні розробки здійснив колектив співробітників Інституту кібернетики АН Української РСР у складі Т.К.Вінцюка, Б.Б.Тимофеєва, В.Г.Зайцева та ін. Перші програми та пристрої розпізнавали кілька десятків окремо вимовлюваних слів. Згодом сформувалося кілька наукових шкіл українських учених, які активно працювали і продовжують працювати над розв'язанням проблеми комп'ютерного аналізу (розпізнавання) усного мовлення людини та його синтезу, або озвучування письмового тексту. У Києві в Інституті кібернетики ім.В.М.Глушкова НАН України цим займається колектив науковців під керівництвом Т.К.Вінцюка. Крім нього, цю проблему також опрацьовують у Львівському національному університеті ім.Івана Франка (під керівництвом М.П.Деркача), Харківському інституті радіоелектроніки (під керівництвом М.Ф.Бондаренка), Дніпропетровському університеті (під керівництвом О.М.Карпова) та Одеському

державному університеті ім. І.І.Мечникова (під керівництвом Т.О.Бровченко та Е.О.Нушікян)[154].

Найбільш розвиненим із запропонованих методів розв'язання цієї проблеми визнаний так званий ***ІКДП-метод***, запропонований колективом кібернетиків під керівництвом Т.К.Вінцюка. Його суть полягає у створенні ієрархічної структури породження певних моделей (прототипів) окремих мовних сигналів (слів, синтагм – фонемних слів, ритмогруп звуків, об'єднаних наголосом певного типу) і вибору з них того, що найбільше відповідає мовному сигналу, який треба розпізнати. Прототипи сигналів зорієнтовані на розпізнавання звукотипів (фонем) у трьох можливих для них ситуацій побутування в мовному потоці: на їхню якість у позиціях перед і після іншої фонеми та в позиції під наголосом. Для процедур синтезу мовлення в письмовій формі (озвучування письмового тексту) здійснюється автоматичний перезапис слів в орфографічній формі у форму фонемну, що відповідає системі фонемної транскрипції українських слів, обраній розробниками методу. На основі існуючих класифікацій українських фонем (М.А.Жовтобрюха, Н.І.Тоцької, В.С.Перебийніс та ін.) Т.К.Вінцюк та його колеги виробили свою максимально деталізовану за якісними й кількісними характеристиками систему з 70 одиниць. Враховуючи їхнє можливе функціонування у трьох вищеназваних типах внутрішньо- або міжслівних позицій, встановлено прототипи (моделі) для 70^3 їхніх реалізацій у потоці мовлення, окремих мовних сигналів. Завдяки запровадженій процедурі оптимізації пошуку потрібного прототипу мовного сигналу описуваний метод і здобув назву методу ***ієрархічної (І)*** *структури* ***композиції (К)*** *складних модельних сигналів мовлення та їхнього порівняння з розпізнаваним сигналом за допомогою* ***динамічного програмування (ДП)***, або ***ІКДП-методу.*** Наприклад, створена на основі цього методу система усного спілкування з персональним комп'ютером "Мова-4" здатна була розпізнавати в реальному часі (в режимі інтерактивного спілкування з комп'ютером) 300 усних слів-команд та озвучувати (синтезувати) будь-який текст українською та російською мовами[155].

Нині колектив, очолюваний Т.К.Вінцюком, успішно працює над розширенням словника розпізнаваних мовних сигналів до 100 тис. слів, створенням так званих дикторонезалежних моделей розпізнавання,

[154] Читачів, зацікавлених детальніше познайомитися зі станом розроблення цієї проблеми, відсилаємо до праць: **Вінцюк Т.** Комп'ютерні автоматичні системи розпізнавання та синтезу українського мовлення // Проблеми українізації комп'ютерів. – К., 1993. – С.21-32; **Вінцюк Т.К.** Анализ, распознавание и интерпретация речевых сигналов. – К., 1987; **Allen T., Hunnicutt M.S., Klatt D.** From Text to Speech. The MITTalk system. – Cambridge etc., 1987.

[155] **Вінцюк Т.** Комп'ютерні автоматичні системи розпізнавання та синтезу українського мовлення. – С.24-25.

тобто моделей, здатних розпізнавати сигнали різного тембру, темпу вимовляння або сили звучання тощо. Цікавим і важливим як з теоретичного, так і з практичного погляду є опрацьовуване завдання створення "автоматичної друкарки", або комп'ютерної системи, здатної друкувати та редагувати тексти з голосу, під диктовку. Системи комп'ютерного аналізу та синтезу мовлення уможливлюють у звуковій формі в цілому різновиди роботи людини з комп'ютером: здійснення машинного перекладу, навчання мови, одержання з баз даних чи знань потрібної довідкової інформації тощо. Окрему важливу ділянку розбудови й функціонування подібних систем становить створення засобів діалогу з комп'ютером для користувачів з вадами зору або слуху. "Окрім "банального" автоматичного перетворення в текст, – зауважує Т.К.Вінцюк, – який читається-сприймається глухими людьми, актуальним є перетворення мовного сигналу в зображення, які "читаються людьми, або які перетворюють текст (автоматично читають текст) в мовлення та "портрет-мовець", що сприймаються людьми, які погано чують"[156]. Таким чином, сьогодні проблема створення засобів взаємодії людини з комп'ютером розв'язується в повному обсязі проблем моделювання штучного інтелекту, які при цьому виникають, а саме: в комплексі завдань аналізу (розпізнавання) та синтезу (породження) як зорових (графічних), так і слухових (акустичних, звукових) образів.

🕮 Терміни

- **діалогова, або питально-відповідна система (= інтерфейс, =система спілкування (діалогу) з комп'ютером, =система людино-машинної взаємодії) –** система мовних засобів для спілкування людини з комп'ютером: надання комп'ютеру завдань (питань, запитів) та одержання відповідей на них
  - **діалогова, або питально-відповідна система (= інтерфейс, =система спілкування з комп'ютером, =система людино-машинної взаємодії) природномовна –** система засобів спілкування людини з комп'ютером, яка використовує лексикон та граматику певної природної мови
  - **діалогова, або питально-відповідна система (=(= інтерфейс, =система спілкування з комп'ютером, =система людино-машинної взаємодії) штучномовна –** система засобів спілкування людини з комп'ютером, яка використовує спеціальну штучну систему символів та засобів їхнього комбінування
  - **діалогова, або питально-відповідна система з жорсткою структурою –** система засобів спілкування людини з комп'юте-

[156] **Вінцюк Т.** Зазнач. праця. – С.30-31.

ром, яка використовує стандартизовані, жорсткі правила побудови повідомлень (=запитів) і відповідей на них, обмежений набір засобів природної мови, що можуть вживатися під час діалогу, а також спеціальний формалізований запис інформації

- **діалогова, або питально-відповідна система з м'якою структурою –** система засобів спілкування людини з комп'ютером, яка залежно від характеру одержуваної від відправника інформації здатна змінювати структуру діалогу адресанта (людини) і адресата (комп'ютера), доповнювати її новими даними або й взагалі змінювати форму їхнього спілкування
  - **блок аналізу повідомлення відправника (= адресанта, людини) –** частина (модуль) діалогової системи аналізу (розпізнавання) форми й змісту повідомлення
  - **блок інтерпретації повідомлення відправника (=адресанта, людини) одержувачем (= адресатом, комп'ютером) –** частина (модуль) діалогової системи, в якому здійснюється переведення змісту розпізнаного повідомлення у форму, вироблену для роботи з ним комп'ютера
  - **блок породження змісту відповіді одержувача (=адресата, комп'ютера) –** частина (модуль) діалогової системи, в якому здійснюється формування змістової структури повідомлення – відповіді адресата (комп'ютера)
  - **блок синтезу відповіді одержувача (= адресата, комп'ютера) –** частина (модуль) діалогової системи, в якому здійснюється втілення змістової структури відповіді комп'ютера у форму, вироблену для його сприйняття відправником (адресантом, людиною)
  - **дружній інтерфейс –** інтерфейс для взаємодії з користувачем, що не має спеціальних знань про апаратне, математичне чи програмне забезпечення комп'ютера
- **форма здійснення діалогу (= взаємодії) людини з комп'ютером** – спеціальна система засобів для оформлення змісту повідомлень
  - **письмова (= графічна) –** оформлення змісту повідомлень з допомогою засобів певної системи письма (графічної системи)
  - **усна (= звукова, акустична) –** оформлення змісту повідомлень з допомогою засобів певної системи звукових сигналів (акустичної системи)
    - **аналіз (= розпізнавання) мовлення –** виділення в мовному потоці окремих сигналів та їхня змістова інтерпретація
    - **синтез мовлення (= породження мовлення, =озвучування письмового тексту) –** комбінування окремих мовних сигналів у змістові блоки повідомлення (слова або синтагми)

- **ІКДП-метод (= Ієрархічна Композиція** складних модельних сигналів мовлення та їхнього порівняння з розпізнаваним сигналом за допомогою **Динамічного Програмування)** – метод аналізу (=розпізнавання) та синтезу (породження) мовлення, який ґрунтується на ієрархічній системі прототипів (моделей) мовних сигналів, що встановлюються з допомогою процедур динамічного програмування
  - **прототип (= модель) мовного сигналу –** певний узагальнений стандартизований зоровий (графічний) або слуховий (звуковий, акустичний) образ мовного сигналу

Л І Т Е Р А Т У Р А:

Праці загального характеру: підручники, посібники, проблемні огляди

1. **Апресян Ю.Д.** Идеи и методы современной структурной лингвистики. – М., 1966.
2. **Баранов А.Н.** Введение в прикладную лингвистику. – М., 2001.
3. **Білецький А.О.** Про мову і мовознавство. – К., 1996. Особливо розд.: "Перекладає машина", "Семіотика – наука про знаки", "Мовознавство серед інших наук", "Мова – особлива знакова система".
4. **Височанський В.С., Кардаш А.І., Костів О.В., Черняхівський В.В.** Елементи інформатики. – Львів, 1990.
5. **Городецкий Б.Ю.** Компьютерная лингвистика: моделирование языкового общения // Компьютерная лингвистика. Вып. XXIV. – Новое в зарубеж. лингвистике. – М., 1989.
6. **Гутер Р.С., Полунов Ю.Л.** От абака до компьютера. – М., 1981.
7. **Ершов А.П.** Машинный фонд русского языка: Внешняя постановка // Машинный фонд русского языка: идеи и суждения. – М., 1986. – С.7-12.
8. **Использование** ЭВМ в лингвистических исследованиях. – К., 1990.
9. **Казакевич О.А.** Автоматизация лексикографических работ. Автоматические словари (обзор зарубежных публикаций) // Научн.-техн. информация. - сер.2. – 1985. - № 9. – С.25-29.
10. **Казакевич О.А.** Машинные фонды языков народов СССР // Научн.-техн. информация. - Сер.2. – 1989. - № 10. – С.25-32.
11. **Караулов Ю.Н.** Общая и русская идеография. – М., 1976.
12. **Клименко Н.Ф.** Нові підходи до укладання комп'ютерних словників // Мовознавство. – 1996. - № 4-5. – С.11-15.
13. **Клименко Н.Ф., Карпіловська Є.А., Комарова Л.І. та ін.** Морфемно-словотвірний фонд української мови як дослідницька та інформаційно-довідкова система // Мовознавство. – 1990. - № 6. – С.41-50.
14. **Клименко Н.Ф., Русанівський В.М.** Від універсальної бази лінгвістичних знань до комп'ютерного укладання словників // Мовознавство. – 1995. - № 4-5. – С.3-10.
15. **Лосев А.Ф.** Введение в общую теорию языковых моделей. – М.: Едиториал УРСС, 2004. – Изд 2-е, стереотип.
16. **Марчук Ю.Н.** Основы компьютерной лингвистики. – М., 2000.
17. **Математические** аспекты структуры языка // Новое в лингвистике. – М., 1965. – вып.IV.
18. **Нелюбин Л.Л.** Компьютерная лингвистика и машинный перевод. – М., 1991.
19. **Основные** направления структурализма. – М., 1964.
20. **Перебийніс В.І.** Статистичні методи для лінгвістів: Навч. посібник. – Вінниця, 2002.
21. **Перебийніс В.С., Муравицька М.П., Дарчук Н.П.** Частотні словники та їх використання. – К., 1985.
22. **Пещак М.М.** Нариси з комп'ютерної лінгвістики. – Ужгород, 1999.

23. **Плат У.** Математическая лингвистика // Новое в лингвистике. – М., 1965. – вып.IV.
24. **Попова Р.К.** Новые информационные технологии и лингвистика. – М., 2005.
25. **Поспелов Г.С.** Искусственный интеллект – основа новой информационной технологии. – М., 1988.
26. **Прикладная** лингвистика // Новое в зарубежной лингвистике. – М., 1983. – вып. XII.
27. **Прикладное** языкознание / Под ред. А.С.Герда. – Санкт-Петербург, 1996.
28. **Рождественский Ю.В., Волков А.А., Марчук Ю.Н.** Введение в прикладную филологию. – М., 1988.
29. **Слама-Казаку Т.** Место прикладной лингвистики в системе наук: отношение ПЛ к "лингвистике" // Новое в зарубеж. лингвистике. – М., 1983. – Вып. XII. Прикладная лингвистика. – С.23-34.
30. **Слокум Дж.** Обзор разработок по машинному переводу: история вопроса, современное состояние и перспективы развития // Новое в зарубеж. лингвистике. Компьютерная лингвистика. Вып. XXIV. – М., 1989. – С.357-406.
31. **Современная** американская лингвистика: Фундаментальные направления / Под ред. А.Е.Кибрика, И.М.Кобозевой и И.А.Секериной. – М., 2002. – Изд. 2-е, испр. и доп.
32. **Сосюр Ф. де.** Курс загальної лінгвістики. – К., 1998. Особливо Вступ: розд.II. "Дослідний матеріал і завдання лінгвістики; її зв'язок із суміжними дисциплінами"; розд.III "Предмет мовознавства"; розд. IV "Лінгвістика мови та лінгвістика мовлення"; розд. V "Внутрішні та зовнішні елементи мови" та частина I. "Загальні принципи": розд. III. Статична лінгвістика та еволюційна лінгвістика".
33. **Українська** мова: Енциклопедія. – К., 2000. – Статті: Автоматизація лінгвістичних досліджень. Автоматичне оброблення тексту. Автоматичний аналіз. Автоматичний словник. Аналіз за безпосередніми складниками. Аналіз у термінах залежностей. Граматики формальні. Конкорданс. Математична лінгвістика. Машинна мова. Машинний переклад. Машинний фонд української мови. Модель. Прикладна лінгвістика. Прогнозування лінгвістичне. Формалізація в лінгвістиці. Частотний словник
34. **Язык** компьютера. – М., 1989.

Праці з окремих проблем комп'ютерної лінгвістики

1. **Анализ** метаязыка словаря с помощью ЭВМ. – Ю.Н.Караулов, В.И.Молчанов, В.А.Афанасьев, Н.В.Михалев. – М., 1982.
2. **Андреев Н.Д.** Статистико-комбинаторные методы в теоретическом и прикладном языковедении. – Ленинград, 1967.
3. **Андрющенко В.М.** Концепция и архитектура Машинного фонда русского языка. – М., 1989.
4. **Анисимов А.В.** Компьютерная лингвистика для всех: Мифы. Алгоритмы. Язык. – К., 1991.
5. **Апресян Ю.Д. и др.** Лингвистическое обеспечение системы ЭТАП-2. – М., 1988.
6. **Баранов А.Н.** Категории искусственного интеллекта в лингвистической семантике. Фреймы и сценарии. – М., 1987.

7. **Белецкий А.А.** Семиотический аспект языковой системы // // Структурная и математическая лингвистика. - К., 1979. – вып.7. – С.11-18.
8. **Белоногов Г.Г., Новоселов А.П.** Автоматизация процессов накопления, поиска и обобщения информации. – М., 1979.
9. **Белоногов Г.Г., Калинин Ю.П., Поздняк М.Ф., Яфаева Г.М.** Алгоритм многоступенчатого морфологического анализа русских слов // Научн.-техн. информация. – Сер 2. - 1983. - № 1. —С.6-10.
10. **Білевич Т.Л.** Принципи ствоорення машинної версії "Словника староукраїнської мови XIV-XV ст." // Проблеми українізації комп'ютерів. – Львів, 1993. – С.14-15.
11. **Брага І.І.** Мовна репрезентація образу держави у пресі України (кінець 70-х – початок 2000-х років). – Автореф. дис. ... канд.філол.наук.–К, 2002.
12. **Вінцюк Т.К.** Анализ, распознавание и интерпретация речевых сигналов. – К., 1987.
13. **Войскунский А.Е.** Я говорю, мы говорим...: Очерки о человеческом общении. – М., 1982.
14. **Вінцюк Т.** Комп'ютерні автоматичні системи розпізнавання та синтезу українського мовлення // Проблеми українізації комп'ютерів. – К., 1993. – С.21-32.
15. **Герд А.С.** Типы русских текстов и организация Машинного фонда русского языка // Машинный фонд русского языка: идеи и суждения. – М., 1986. – С. 67-75.
16. **Герд А.С., Богданов В.В., Азарова И.С., Аверина С.А., Зубова Л.В.** Автоматизация в лексикографии и словари-конкордансы // Филологические науки. – 1981. - № 1. – С.72-78.
17. **Глушков В.М.** Гносеологические основы математизации науки // Диалектика и логика научного познания. – М., 1966. - С. 406-412.
18. **Голубовская И.А.** Этнические особенности языковых картин мира. – К., 2002. – 293 с.
19. **Грязнухіна Т.О.** Лінгвістичне забезпечення автоматизованих систем управління // Мовознавство. – 1983. - № 5. – С.39-43.
20. **Грязнухіна Т.О., Нікула М.В.** Система автоматичного морфологічного аналізу українського наукового тексту // Проблеми українізації комп'ютерів. – К., 1993. – С.42-46.
21. **Дарчук Н.П., Грязнухіна Т.О.** Частотний словник сучасної української публіцистики // Мовознавство. – 1996. - № 4-5. – С.15-18.
22. **Дерягин В.** Учить ли ЭВМ писать с ошибками? // Правда. – 1983. – 23 сент.
23. **Ершов А.П.** Методологические предпосылки продуктивного диалога с ЭВМ на естественном языке // Вопросы философии. – 1981. - № 8. – С.115.
24. **Захарова Л.М.** Національно-культурні конотації іменувань найвищих державних правителів (на матеріалі творів О.С.Пушкіна). – Автореф. дис. ... канд.філол.наук. – Сімферополь, 2000.
25. **Звегинцев В.А.** Язык как фактор компьютерной революции // Научн.-техн. информация. - Сер.2. – 1985. - № 9. – С.1-7.
26. **Интегральные** роботы. – М., 1973.
27. **Караулов Ю.Н.** Лингвистическое конструирование и тезаурус литературного языка. - М., 1981.

28. **Карпіловська Є.А.** Конструювання складних словотворчих одиниць. – К., 1990.
29. **Карпіловська Є.А.** Машинні версії традиційних словників як основа для укладання комп'ютерних словників та тезаурусів // Мовознавство. – 1996. - № 4-5. – С.19-22.
30. **Карпіловська Є.А.** Морфемна сітка як інструмент дослідження будови слова // Українське мовознавство. – 1992. – вип.19. – С.100-110.
31. **Карпіловська Є.А.** Суфіксальна підсистема сучасної української літературної мови: будова та реалізація. - К., 1999.
32. **Карпіловська Є.А.** Термінологічний підфонд у складі морфемно-словотвірного фонду української мови /принципи формування та можливості використання/ // Україномовне програмне забезпечення. Матеріали 4-ої та 5-ої Міжнарод. науково-практ. конф. "УкрСофт". - Львів,1995. - С.161-162.
33. **Клименко Н.Ф.** Основи морфеміки сучасної української мови. -К., 1998, 1-е вид., 2000, 2-е вид.
34. **Клименко Н.Ф.** Построение тезауруса с помощью ЭВМ // Украинский семантический словарь. Проспект. – К., 1990. – С.81-89.
35. **Клименко Н.Ф.** Система афіксального словотворення сучасної української мови. - К., 1973.
36. **Клименко Н.Ф., Карпіловська Є.А.** Морфеміка слов'янських мов як об'єкт типологічного вивчення // Мовознавство. – 1998. - № 2-3. – С.117-135.
37. **Клименко Н.Ф., Карпіловська Є.А.** Морфемні структури слів у сучасній українській літературній мові // Мовознавство. – 1991. - № 4. – С.10-21.
38. **Клименко Н.Ф., Карпіловська Є.А.** Словотвірна морфеміка сучасної української літературної мови. - К., 1998.
39. **Клименко Н.Ф., Карпиловская Е.А., Комарова Л.И.** Машинные словари морфемно-словообразовательного фонда украинского языка // Актуальные проблемы компьютерной лингвистики. - Тарту,1990. – С.54-57.
40. **Коваль М.Д., Багдасар'ян Г.М.** Україномовні комп'ютерні навчальні системи // Проблеми українізації комп'ютерів. – К., 1993.
41. **Ковтуненко Л.С.** Комп'ютерні аспекти лексикографічних систем // Мовознавство. – 1996. - № 4-5. – С.28-34.
42. **Козачков Л.С.** Прикладная логика информатики. – К., 1990.
43. **Колодяжная Л.И.** Структура словарного текста в аспекте машинной лексикографии. – Автореф. дис. ... канд. філол.наук. - М., 1986.
44. **Компьютерная** лингвистика // Новое в зарубежной лингвистике. – М., 1989. - Вып.XXIV.
45. **Коссак О., Коруд О., Хвищун Л.** Система комп'ютерної підтримки словникових баз даних СЛОВО // Проблеми українізації комп'ютерів. – К., 1993. – С.73-76.
46. **Кузнецова Э.В.** Лексикология русского языка. - М., 1982.
47. **Кузнецова Э.В.** Ступенчатая идентификация как средство описания семантических связей слов // Вопросы металингвистики. - Ленинград, 1973. - С.84-94.
48. **Лексична** семантика в системі "людина-машина". – К., 1986.
49. **Лингвистическая** прагматика и общение с ЭВМ / Отв. ред. Ю.Н.Марчук. – М., 1989.

50. **Лингвистические** проблемы автоматизации редакционно-издательских процессов. – К., 1986.
51. **Лук'янчук С.** Комп'ютерна модель парадигматичних класів дієслів // Українське мовознавство. – 2000. – вип.22. – С.82-85.
52. **Лурия А.Р.** Язык и сознание. – М., 1981.
53. **Мельчук И.А.** Опыт теории лингвистических моделей "Смысл↔Текст". – М., 1974.
54. **Минский М.** Структура для представления знаний // Психология машинного зрения. – М., 1978.
55. **Минский М.** Фреймы для представления знаний. – М., 1979.
56. **Моделирование** языковой деятельности в интеллектуальных системах. – М., 1987.
57. **Морфологический** анализ научного текста на ЭВМ. – К., 1990.
58. **Нагорна Л.І., Терзян Т.К., Цейтлін Г.О.** Засоби автоматизованого озвученого конструювання текстів // // Проблеми українізації комп'ютерів. – К., 1993.
59. **Олексієнко Л., Дарчук Н.** Лематизація парадигм іменників української мови // Проблеми українізації комп'ютерів. – К., 1993. – С.62-65.
60. **Осуга С.** Обработка знаний. – М., 1989.
61. **Пещак М.М.** [Рецензія] // Мовознавство. – 1984. - № 2. – С.74-75. – Рец. на кн.: Русский семантический словарь.
62. **ПиотровскийР.Г.** Лингвистический автомат и его речемыслительное обоснование. – Минск, 1999.
63. **Проблеми** українізації комп'ютерів. – К., 1993.
64. **Рогожникова Р.П., Чернышева Л.В.** Организация словарной картотеки на базе автоматизированной системы // Теория и практика современной лексикографии. - Ленинград, 1984. – С.20-27.
65. **Розенцвейг В.Ю.** Опыт создания национальных лексикографических служб за рубежом // Машинный фонд русского языка: идеи и суждения. – М.: Наука, 1986. – С.75-84.
66. **Русанівський В.М., Тараненко О.О., Широков В.А.** Теоретико-лінгвістичні засади та інформаційно-комп'ютерне забезпечення україномовних лінгвістичних інтелектуальних систем // Мовознавство.– 1996.- № 4-5.
67. **Середницька А.Я.** Ідеографічний поділ дієслівної лексики в сучасній українській мові. – Автореф.дис. ... канд.філол.наук. – К., 2001.
68. **Синтаксический** анализ научного текста на ЭВМ. – К., 2000.
69. **Скороходько Э.Ф.** Семантические сети и автоматическая обработка текста. – Киев, 1983.
70. **Сніжко Н.В.** Ідеографічний тезаурус як модель лексико-семантичної системи (за наслідками автоматизованого аналізу українських іменників) // Мовознавство. – 1995. - № 6. – С.28-35.
71. **Сніжко М.Д., Сніжко Н.В.** Автоматизована система класифікації лінгвістичних об'єктів // Проблеми українізації комп'ютерів. – К., 1993. – С.69-72.
72. **Сніжко Н.В., Сніжко М.Д.** "Ідеографічний тезаурус" як інформаційно-довідкова система при вивченні закономірностей структурно-функціональної організації лексики // Мовознавство. - 1996. - № 4-5. – С.23-28.

73. **Соссюр Ф.де.** Мемуар о первоначальной системе гласных в индоевропейском языке // Соссюр Ф. де. Труды по языкознанию - М., 1977. – С.302-562.
74. **Структуры** представления знаний в языке. – М., 1994.
75. **Тер-Минасова С.Г.** Язык и межкультурная коммуникация. – М., 2000.
76. **Тулдава Ю.А.** О машинных фондах русского языка за рубежом // Машинный фонд русского языка: идеи и суждения. – М., 1986. – С.141-142.
77. **Уинстон П.** Искусственный интеллект. – М., 1980.
78. **Фитиалов С.Я.** О построении формальной морфологии в связи с машинным переводом // Тезисы конф. по обработке информации, машинному переводу и автоматическому чтению текста. – М., 1961.
79. **Формалізовані** основи семантичної класифікації лексики. – К.: Наук.думка, 1982.
80. **Хант Э.** Искусственный интеллект. – М., 1978.
81. **Шайкевич А.Я.** Об автоматическом построении тезауруса на основе толковых словарей // Науч.-техн. информация. - Сер.2.– 1985. - № 4.– С.12-19.
82. **Шайкевич А.Я.** Дистрибутивно-статистический анализ текстов. – Ленинград, 1982.
83. **Шайкевич А.Я.** О статистическом словаре языка Достоевского // Русский язык в научном освещении. – М., 2001. - № 2. – С. 122-149.
84. **Шаумян С.К., Соболева П.А.** Основания порождающей грамматики русского языка: Введение в генотипические структуры. - М., 1968.
85. **Шевченко І.В.** Алгоритмічна словозмінна класифікація української лексики // Мовознавство. – 1996. - № 4-5. – С.40-44.
86. **Шенк Р.** Обработка концептуальной информации. – М., 1980.
87. **Широков В.А.** Інформаційна теорія лексикографічних систем. – К., 1998.
88. **Язык** и структуры представления знаний. – М., 1992.
89. **Allen T., Hunnicutt M.S., Klatt D.** From Text to Speech. The MITTalk system. – Cambridge etc., 1987.
90. **Klymenko N.F., Karpilovs'ka E.A.** Computer Morpheme-Word-Formative Database of the Ukrainian Language and Its Applications // Journal of Quantitative Linguistics. – 1994. - Vol.1. - No.2. - P.113-131.
91. **Menzerath P.** Die Architektonik des deutschen Wortschatzes // Phonetische Studien. - 1954. - № 3.
92. **Piotrovskij Raimund G.** MT in the former USSR and in the Newly Independent States (NIS). Prehistory, romantic era, prosaic time // Early Years in Machine Translation. Memoirs and Biographies of Pioneers. Ed. by W.John Hutchins. – Amsterdam/Philadelphia, 1995.

Словники

1. **Англо-**русский словарь по вычислительной технике. - М., 1974.
2. **Баранов О.С.** Идеографический словарь глаголов русского языка. М., 1995.
3. **Караванський Святослав**. Словник рим української мови, укладений як лінгвографічна модель формального нагромадження звукових сигналів мовним центром людини. – Львів: БаК, 2004.
4. **Караулов Ю.Н.** Частотный словарь семантических множителей русского языка. – М., 1980.

5. **Карпіловська Є.А.** Кореневий гніздовий словник української мови: Гнізда слів з вершинами – омографічними коренями. – К., 2002.
6. **Клименко Н.Ф., Карпіловська Є.А., Карпіловський В.С., Недозим Т.І.** Словник афіксальних морфем української мови. – К., 1998.
7. **Конкорданція** поетичних творів Тараса Шевченка / A Concordance to the Poetic Works of Taras Shevchenko. – I-IV тт. – Edmonton-Toronto, 2001.
8. **Коссак О.М., Маньковський С.Л.** Англо-українсько-російський словник з інформатики та обчислювальної техніки. – Л., 1991.
9. **Кубрякова Е.С., Демьянков В.З., Панкрац Ю.Г., Лузина Л.Г.** Краткий словарь когнитивных терминов. – М., 1996.
10. **Лексическая** основа русского языка / Под ред. В.В.Морковкина. - М., 1984.
11. **Обернений** частотний словник сучасної української художньої прози. – К., 1998.
12. **Русский** семантический словарь: Опыт автоматического построения тезауруса: от понятия к слову. - М., 1982.
13. **Русский** семантический словарь. Толковый словарь, систематизированный по классам слов и значений. Т.1.: Слова указующие (местоимения). Слова именующие: имена существительные (Все живое. Земля. Космос) / Под ред. Н.Ю.Шведовой. - М., 1998.
14. **Словарь** славянской лингвистической терминологии: В 2 тт.– Praha, 1977.
15. **Толковый** словарь по вычислительным системам. – М., 1989.
16. **Толковый** словарь русских глаголов: Идеографическое описание. Английские эквиваленты. Синонимы. Антонимы / Под ред. Л.Г.Бабенко. – М., 1999.
17. **Украинский** семантический словарь: Проспект. – К., 1990.
18. **Computer-Konkordanz** zum Novum Testamentum Graece. – Berlin, New York, 1980.
19. **Rudnik-Karwatowa Z., Karpińska H.** Słownik słów kluczowych językoznawstwa slawistycznego. – Warszawa, 1999.

60

СЛОВНИК ТЕРМІНІВ КОМП'ЮТЕРНОЇ ЛІНГВІСТИКИ

1. **автоматизоване робоче місце (АРМ) перекладача**
2. **алгоритм**
   - лінгвістичний
   - машинний
   - ***алгоритму*** блок-схема (= граф-схема)
   - ***алгоритму*** крок
   - ***алгоритму*** мова представлення
   - ***алгоритму*** правило виконання
   - ***алгоритму*** програмна реалізація
3. **база даних (database)**
   - ієрархічна
   - реляційна
   - словникозорієнтована (= словникова, лексикографічна)
   - текстозорієнтована (= текстова)
     - повнотекстова (= корпус текстів)
       - фундаментальна
       - дослідницька (пошукова)
   - ***бази даних датологічний*** етап проектування
     - ***датологічного етапу*** логічна стадія
     - ***датологічного етапу*** фізична стадія
   - ***бази даних*** інфологічний етап проектування
   - зонний принцип запису мовної інформації
   - комп'ютерна підтримка
4. **база знань (knowledge base), інтелектуальна база даних (intelligent database)**
5. **діалогова, або питально-відповідна система (= інтерфейс, = система спілкування (діалогу) з комп'ютером, = система людино-машинної взаємодії)**
   - блок аналізу повідомлення відправника (= адресанта, людини)
   - блок інтерпретації повідомлення відправника (= адресанта, людини) одержувачем (= адресатом, комп'ютером)
   - блок породження змісту відповіді одержувача (= адресата, комп'ютера)
   - блок синтезу відповіді одержувача (= адресата, комп'ютера
   - з жорсткою структурою
   - з м'якою структурою
   - природномовна
   - форма здійснення діалогу (= взаємодії) людини з комп'ютером
     - письмова (= графічна
     - усна (=звукова, акустична)
       - аналіз (= розпізнавання) мовлення
       - синтез мовлення (= породження мовлення, = озвучування письмового тексту)
       - ІКДП-метод (= **Ієрархічна Композиція** складних модельних сигналів мовлення та їхнього порівняння з

розпізнаваним сигналом за допомогою **Динамічного Програмування**)
прототип (= модель) мовного сигналу
штучномовна

6. **експертна система**
7. **електронна картотека (= ілюстративна база даних, база цитат)**
8. **знання**
   декларативне (= ***знання*** "що")
   логічні моделі (= ***знання*** структури представлення)
   динамічні
   сценарій
   сцена
   скрипт
   статичні
   реляційні
   семантична сітка
   семантичний дериват
   семантична похідна
   семантичний компонент
   семантичний складник
   рекурсивність
   табличні (=спискові)
   тезаурусні (=ієрархічні)
   фрейм
   підфрейм (= ***фрейм*** вкладений)
   терм (= об'єкт, термінальний вузол)
   слот
   слота зміст
   процедурне (= ***знання*** "як")
9. **інтелект**
   коефіцієнт інтелектуальності (Intelligence quotient, IQ)
   "машина Тьюринга"
   природний (= ***інтелект*** людини)
   "тест Тьюринга"
   штучний (artificial intelligence) (= розум штучний) (artificial mind)
10. **інтерфейс (= засоби доступу до бази даних та її ведення)**
    дружній
    природномовний
    штучномовний
    меню
    режим оброблення бази даних або бази знань
    інтерактивний (= діалоговий, онлайновий (on-line), роботи в масштабі реального часу)
    пакетний
11. **комп'ютерна версія традиційного словника**
    комп'ютерної версії традиційного словника граматика
12. **комп'ютерна граматика**
    комп'ютерна морфологія

комп'ютерна семасіологія
комп'ютерний синтаксис

13. **комп'ютерна копія традиційного словника**
14. **комп'ютерне середовище для лінгвіста**
    забезпечення роботи комп'ютера
        алгоритмічне та програмне (= software)
        інструментальне (= апаратне, hardware)
        лінгвістичне (= lingware)
15. **комп'ютерний (= автоматичний) словн**ик
    комп'ютерного (автоматичного) словника формат статті
16. **конкорданс (= словники контекстів вживання слова, його лексичних оточень)**
17. **конструювання лінгвістичних об'єктів**
18. **корпус**
19. **корпусна лінгвістика**
20. **лінгвостатистика**
    статистична лексикографія
        частотний словник
            генеральна сукупність
                вибірка
                    випадкова
                    механічна
                    типова (зональна)
                        підвибірка
            ранг слова
            список алфавітно-частотний
            список ранговий
            частота
                абсолютна
                відносна
                середня
                    частоти середньої показник міри коливання
                частоти поріг
21. **машинний переклад (МП) (machine translation) (MT)**
    глобальний підхід до МП
    локальний підхід до МП
    людино-машинний переклад (ЛМП) (machine-aided translation) (MAT)
    мова-посередник (interlingua)
    непрямий підхід до МП
    прямий підхід до МП
    трансфер (= міжмовне перетворення, міжмовна операція)
22. **машинний продукт (= продукт оброблення бази даних, = продукт обчислення)**
23. **модель**
    моделі гіпотетична функція
    дедуктивна (= синтезу, = породжувальна)
    індуктивна (= аналізу)

динамічна (= функціональна, = процесуальна)
статична (= структурна, = класифікаторна, = таксономічна)
моделі оригінал (= натурний об'єкт, = прообраз, = прототип)
моделі пояснювальна функція

24. прикладна лінгвістика (applied linguistics)
дані
знання
інформатика
інформаційна модель
інформація
комп'ютерна лінгвістика (computational linguistics)
предметна галузь

25. процесор
лінгвістичний (= мовний)
корпусний (= текстовий)
лексикографічний (= словниковий)
комп'ютерні засоби конструювання нових словників
постпроцесор
препроцесор
центральний

26. лінгвістична інтелектуальна комп'ютерна система
система автоматичного перероблення тексту (АПТ) (= автоматизована система опрацювання тексту (АСОТ)
АПТ, або АСОТ лінгвістичне забезпечення
лінгвістичного забезпечення АПТ, або АСОТ стратегія створення безсловникова, або «незалежна»
лінгвістичного забезпечення АПТ, або АСОТ стратегія створення словникова
АПТ, або АСОТ модулі
автоматичний морфологічний аналіз тексту (АМА)
граматичний клас одиниць тексту
граматичний підклас одиниць тексту
доморфологічний аналіз
контекстний аналіз (КА)
опорна точка
флективний аналіз (ФА)
квазіфлексія (= графемний ідентифікатор)
автоматичний синтаксичний аналіз тексту (АСА)
АСА за безпосередніми складниками речення
АСА за зв'язком залежності між словами
графічне представлення синтаксичних структур
орієнтований граф (= дерево залежностей)
скобовий (дужковий) запис
стрілковий запис
методики встановлення синтаксичної структури речення
опорних точок
передбачувального аналізу

послідовного аналізу
фільтрів
АСА з безперервним переглядом тексту
АСА з циклічним переглядом тексту
АСА інтегральна
АСА локальна
АСА часткова
АСА універсальна (= глобальна)
автоматичний логіко-семантичний аналіз тексту
інформаційно-пошукова система (ІПС)
параметри ефективності інформаційно-пошукової системи (ІПС)
повнота видачі інформації
точність видачі інформації
автоматичне індексування тексту
пошуковий образ документа (ПОД)
пошуковий припис (ПП)
інформаційний шум
інформаційно-пошукова мова (ІПМ)
мова-класифікація
дескрипторна мова
дескриптор (= ключове слово, слово-концепт)
тезаурусний метод
інформаційно-пошуковий тезаурус (ІПТ) (= дескрипторний словник)
безтезаурусний метод
контент-аналіз
концептуальна змінна
мовний корелят

27. словопокажчик (= індекс)
28. термінологічний банк даних (ТБД)
29. формалізація
внутрішня
зовнішня

Навчальне видання

Карпіловська Євгенія Анатоліївна

ВСТУП ДО ПРИКЛАДНОЇ ЛІНГВІСТИКИ: КОМП'ЮТЕРНА ЛІНГВІСТИКА

Підручник

Комп'ютерна верстка *В.С. Карпіловський*
Технічний редактор *Ю.М. Федюшкіна*

Підписано до друку 30.01.2006 р.
Формат 60x84/16. Папір офсетний.
Гарнітура «Times». Друк — різографія.
Ум. друк. арк. 10,93. Обл.-вид. арк. 10,38.
Наклад 300 прим. Зам. № 007.

Видавництво та друк ТОВ «Юго-Восток, Лтд».
83055, Донецьк, вул. Щорса, 17.
Тел./факс: (062) 305-50-13. E-mail: vostok@skif.net
Свідоцтво про держреєстрацію:
серія ДК №1224 від 10.02.2003 р.