

ВИСНОВОК

Електронний словник синонімів, створений у межах ЛІС, інтегровано в ній з електронним ТЛС за допомогою спеціальних програмних технологій. Внутрішня структура словникових статей забезпечує експліковане представлення інформації про синонімічні відношення в лексичній системі української мови. Все це дає підставу вважати побудований словник ефективним інструментом синонімічної параметризації ТЛС, одним з результатів якої є уточнення лінгвістичного опису значень полісемічних слів додатковою семантичною ознакою – синонімічним контекстом. Крім того, ЕСС є одним з основних засобів семантичної розмітки текстових корпусів.

Автори висловлюють щире подяку Олені Устимець за її внесок в укладання електронного словника синонімів.

Література

1. Широков В.А. Феноменологія лексикографічних систем : монографія / В.А. Широков. – К.: Наукова думка, 2004. – 327 с
2. Широков В. А. Семантичні стани мовних одиниць та їх застосування в когнітивній лексикографії / В.А. Широков // Мовознавство. – 2005. – №3 – 4.– С. 47 – 62.
3. Грязнухіна Т. Операційне визначення критерію семантичної подібності синонімів / Т.Грязнухіна, О.Устимець, В.Широков // Прикладна лінгвістика та лінгвістичні технології: MegaLing-2008: 36. наук.пр. / НАН України, укр. мовно-інформ.фонд та ін.; редкол. : Ю.Д. Апресян та ін.. – К.: Довіра, 2009. – С. 42 – 57.
4. Устимець О. В. Формування електронного словника синонімів української мови / Устимець О. В // Учёные записки Таврического национального университета им. В. И. Вернадского. Серия „Филология”. Том 20 (59).– Симферополь, 2007. – № 4.– С. 57 – 61.
5. Общее языкознание / под ред. А. Е. Супрун. – Минск, 1983. – 456 с.
6. Словник синонімів української мови: у 2 т.– К., 1999–2000.

Електронна лексикографічна система «Морфограф»: теоретичні засади та методика конструювання (проект)

Оксана Зубань

к. філол. н., доцент, доцент кафедри української мови та прикладної лінгвістики, Київський національний університет імені Т. Шевченка, Україна, E-mail: oxana.mell.zuban@gmail.com

The article describes theoretical issues, principles of construction and problems set in the project of creating the “Morphograph” electronic lexicographical system representing Ukrainian language morphemics. This system is created on the basis of linguistic morphemic data base containing 200,000 words which applies computer modelling technique of structural-functional connections of morphs in a word. The morphemic data base provides the following benefits: automatic classification of Ukrainian language lexicon according to different parameters of morphemic structure of a word (quantity of morphemes in a word, models of morphemic structures, specified affix or root morph); automatic compilation of root and affix lists on the basis of a dictionary of initial word forms. The “Morphograph” lexicographical system is based on theoretical principles of modern morphemology and sets a goal to systematize and describe basic units of the Ukrainian language morphemic system – morphs, morphemes and morphemic structures of words – in two electronic dictionaries: Dictionary of morphemic structures of words and Dictionary of morphemes. Each dictionary will be compiled in interactive mode according to the user’s specified search options and it will also provide the function of copying an automatically compiled list of units.

Ключові слова — комп’ютерна лексикографія, морфемна база даних, морф, морфема, морфна / морфемна структура слова, морфемна парадигма.

Вступ

Українська морфемна лексикографія, започаткована відомими традиційними словниками «Морфемний аналіз: Словник-довідник» І. Т. Яценка [31], «Морфемний словник» Л. М. Полюги [25], із розвитком теорії морфеміки та словотвору поповнилася лексикографічними працями, які ставлять нові лексикографічні завдання і є взірцем словникарства нової доби – комп’ютерної української лексикографії: «Словник афіксальних морфем української мови» [26], «Кореневий гніздовий словник української мови» [17] та «Шкільний словотвірний словник сучасної української мови» [29]. Ці словники нового типу уклалися за допомогою комп’ютера на матеріалі електронної словникорієнтованої бази даних морфемно-словотвірного фонду Інституту мовознавства ім. О. О. Потебні НАН України, вони існують у традиційному паперовому форматі, а також у форматі комп’ютерних копій паперових словників.

У сучасному комп’ютеризованому світі традиційний «паперовий» словник або його комп’ютерна копія не задовольняють потреб користувача-лінгвіста, який потребує швидкого й ефективного способу добору, оброблення та систематизації лінгвістичної інформації. Тому актуальним завданням у розвитку сучасної української комп’ютерної лексикографії є комп’ютерне моделювання електронних (комп’ютерних)

лінгвістичних інтерактивних лексикографічних систем, оснащених пошуково-класифікаційними опціями, що забезпечують лінгвістичний аналіз на базі великих лексичних масивів.

Метою статті є опис методологічних засад та структури лексикографічної системи з морфеміки української мови «Морфограф», яка створюється на матеріалі морфемної бази даних (МБД) лабораторії комп'ютерної лінгвістики Київського національного університету імені Тараса Шевченка з метою представлення в мережі Інтернет як інтерактивної автоматичної інформаційно-довідкової системи з морфеміки української мови.

Колектив лабораторії комп'ютерної лінгвістики впродовж багатьох років працює над створенням тексторієнтованих параметризованих баз даних різних мовних рівнів. Результатом роботи у цьому напрямку в галузі морфеміки та словотвору було створення в 2001 р. автоматизованої системи морфемно-словотвірної аналізу (АСМСА) української мови.

Система АСМСА створювалась на матеріалі МБД обсягом ≈ 170 тис. слів, поділених на морфи за принципами морфемного аналізу словника І. Т. Яценка. У створенні АСМСА можна визначити декілька етапів її укладання, що спрямовані на виконання різних завдань автоматичного морфемного аналізу та поповнення бази новою лексикою.

1 етап: 1) укладання бази даних морфемних структур слів, де в автоматизованому режимі було просегментовано на морфеми ≈ 170 тис. слів і приписано кожному слову модель морфемної структури; 2) укладання бази даних аломорфічних коренів (≈ 2500 коренів); 3) укладання бази даних омонімічних коренів (≈ 3100 коренів). Ці завдання виконувались із метою: 1) автоматичного групування слів у спільнокореневі вибірки з урахуванням омонімії та аломорфії коренів; 2) автоматичного сегментування текстових слововживань української мови на морфеми в Параметризованій базі даних поетичного мовлення. Опис автоматизації цих завдань і отримані результати описані в публікаціях авторів проекту [1; 2; 3; 7; 15].

У процесі створення Корпусу української мови (www.mova.info) МБД була використана для проведення морфемного аналізу початкових форм слів (лематизованих слововживань) і автоматичного укладання алфавітно-частотних морфемних словників за текстовими вибірками Корпусу [9; 10; 11; 12]. На сьогодні створено частотні словники на матеріалі 13 текстових вибірок: ЧС 9 авторів, ЧС публіцистики, ЧС художньої прози, ЧС фольклорних текстів та ЧС ендокринології. Із цими словниками можна ознайомитися у рубриці «Частотні словники» [28], де користувач за заданими параметрами може автоматично укласти такі типи морфемних словників: Частотний словник кореневих та афіксальних морфем, у якому представлено кореневі та афіксальні морфеми з реалізацією їх у спільнокореневій та спільноафіксальній лексиці; Частотний словник

морфемних структур слів, у якому представлено лексику, згруповану за кількісно-функціональними моделями морфемних структур.

2 етап: 1) поповнення МБД новою лексикою, яка автоматично добиралася як «необроблена» за текстами Корпусу української мови, на сьогодні реєстр МБД складає ≈ 200 тис. слів; 2) приписування кожній афіксальній морфемі слова значення та функції в морфемній структурі слова; 3) визначення мінімальної словотвірної пари, в якій аналізований афікс виконує словотвірну функцію; 4) визначення морфологічних явищ у морфемній структурі кожного походного слова. Ці завдання виконуються з метою автоматичного групування лексики у спільноафіксальні вибірки з урахуванням омонімії та аломорфії афіксів і укладання електронного тлумачного словника афіксальних морфем.

На сьогодні МБД складається із шести взаємопов'язаних баз даних: 1) бази морфемних структур слів ≈ 200 тис. слів; 2) бази аломорфічних коренів (≈ 2500 коренів); 3) бази омонімічних коренів (≈ 3100 коренів); 4) бази аломорфічних афіксів (у процесі укладання); 5) бази омонімічних афіксів (у процесі укладання); 6) бази морфологічних альтернатив (у процесі укладання). Укладання МБД відбувається автоматизовано в режимі on-line у мережі Інтернет за допомогою зручного комп'ютерного інструмента, роботу якого забезпечує система `morfem.exe`.

За типом джерел формування МБД визначається як словниковозорієнтована, оскільки лексичний реєстр бази (≈ 170 тис. слів) початково формувалася на матеріалі словників української мови, але використання бази як морфемного аналізатора в Корпусі української мови зумовило поповнення реєстру бази новою лексикою, дібраною із текстів Корпусу, тому МБД за джерелом формування можна визначити і як текстозорієнтовану базу даних.

Функціонально система АСМСА, яка представляє датологічний етап проектування МБД, може використовуватися як: 1) інформаційно-довідкова система з морфеміки та словотвору української мови; 2) автоматичний лінгвістичний конструктор частотних морфемних словників на базі лексики текстових вибірок у Корпусі української мови; 3) автоматичний морфемний сегментатор текстових слововживань. Таким чином, за класифікаційними ознаками комп'ютерних словників, визначених В.І.Перебийніс [24], МБД характеризується як: 1) комп'ютерний (електронний) морфемний словник, призначений для користувача в лінгвістичних дослідженнях; 2) автоматичний словник, що використовується в комп'ютерних системах автоматичного морфемного аналізу лексики українських текстів.

МБД як інформаційно-довідкова система з морфеміки та словотвору української мови активно використовується філологами КНУ імені Т. Шевченка у різноманітних наукових дослідженнях, проте ця база даних не представлена як лексикографічний продукт для широкого кола користувачів. Обсяг

лексичного реєстру, параметри комп'ютерного моделювання лінгвістичної інформації, класифікаційні можливості МБД уможливили формулювання нового лексикографічного завдання: створити на основі МБД загальнодоступний інформаційний ресурс – комп'ютерну лексикографічну систему «Морфограф» для користувачів-філологів.

I. Теоретичні засади створення лексикографічної системи "Морфограф"

Основною вимогою лінгвістичного опису, зокрема і лексикографічного, є повнота і цілісність, а в нашому дослідженні – створення повного і цілісного лексикографічного опису морфемної системи української мови. Тому одиниці і принципи лексикографічного опису повинні визначатися системно-структурними закономірностями рівневої організації мовної системи.

У мовознавчій теорії рівнем мовної системи називається «сукупність усіх відносно однорідних одиниць (одиниць однакової складності), які можуть вступати між собою у синтагматичні і парадигматичні відношення, але не можуть знаходитися в ієрархічних відношеннях (фонема не можуть складатися з фонем, морфема не можуть складатися з морфем)» [27:123]. Рівень мовної системи визначається синтагматичним дискретним сегментом мовлення, відповідно, базовою одиницею морфеміки виступає мінімальна знакова одиниця мови – морф.

Хоча у лінгвістичній традиції терміни «морф» і «морфема» часто вживаються як синонімічні, лексикографічний опис керується принципом чіткого розмежування цих двох понять як явищ синтагматики та парадигматики. Морф – мінімальна знакова синтагматична одиниця, яка є частиною слова і не має синтаксичної самостійності. Морф – одиниця мовлення, тобто одиниця, що характеризується як психолінгвістичною, так і субстанціальною (фонетичною // графемною) реальністю. Морфема – парадигматична одиниця мови, що реалізується у конкретних варіантах – морфах, і є інваріантом стосовно морфів, які її репрезентують у мовленні.

Актуалізація синтагматичних відношень між морфами на обох рівнях організації структури мовного знака (формальному і семантичному) приводить до утворення одиниці вищого рівня – словоформи і, відповідно, репрезентує лінгвістичний тип відношень за ступенем складності – ієрархічні відношення.

Враховуючи три типи відношень, в які вступає морф у системі мови, лексикографічний опис ставить завдання визначити одиниці морфемної системи української мови – морфи – і описати структуру цієї системи: синтагматичні, парадигматичні та ієрархічні відношення морфів як базових одиниць морфемного рівня мови. Керуючись традиційним визначенням морфа // морфема як одиниці, що визначається у складі слова, виконання поставлених завдань

передбачає насамперед визначення морфемної будови слів української мови – проведення морфемного аналізу. Саме це завдання є основним і воно виконано у МБД, яка представляє реєстр слів, просегментованих на морфи. Оскільки у сучасній морфемології, що відносно недавно визначилась як окрема лінгвістична дисципліна, не має чіткого розмежування теоретичних засад і методики морфемного та словотвірного аналізів, необхідно визначити принципи морфемного аналізу, представленого в МБД.

Активний розвиток теорії і практики словотвірного аналізу, порівняно з морфемним, зумовив визначення морфної структури слова у термінах словотвірного аналізу. Така методика використовує поняття морфної структури слова як результат послідовного здійснення актів словотворення – визначений морф повинен відображати один словотвірний такт, що суперечить традиційному визначенню морфема як неподільного мінімального мовного знака, оскільки призводить до узаконення комплексних морфем: біморфем (конфіксів) [30], тричленних морфем [4], поліморфем [22], а також спричинює неоднозначне морфне членування у різних словниках та граматиках.

Морфна структура слова не завжди пояснюється дериваційними процесами і за своїм кількісно-морфним складом не обов'язково відображає морфні межі, що актуалізуються в словотвірних структурах: 1) словотвірні форманти можуть складатися із декількох морфів; 2) непохідні основи слів також можуть поділятися на морфи. Наприклад, у словах *до-гляд-а-льн-иц-я*, *за-лиц-я-льн-иц-я* на морфемному рівні за аналогією до слів *регулювальниця*, *перевірляниця*, утворених від прикметників *регулювальний*, *перевірляльний*, визначається суфіксальна послідовність *-а-льн-иц-*. Оскільки в мові не зафіксовано таких слів, як *доглядальний* та *залицяльний*, словотвірний аналіз визначає афікс *-льниц-*, на базі словотвірних пар *доглядати* – *доглядальниця*, *залицятися* – *залицяльниця*.

Визначення афікса, що не характеризується словотвірною функцією, не заперечує його структурно-морфемної самостійності. Якщо суфікси (наприклад *-ок-*, що виражає значення 'предметності') знаходяться у відношенні додаткової дистрибуції із суміжними коренями, тобто сполучаються у препозиції з різними коренями, наприклад, *біл-* (*біл-ок* - 'хімічна сполука'), *буз-* (*буз-ок*), *-він-* (*він-ок*), то вони визначаються як синтагматичні знакові одиниці, що репрезентують морфему *-ок-*, незважаючи на те, що у синхронічному словотворі ці слова є непохідними.

Морфи *-льн-*, *-иц-* та *-ок-* визначаються не за словотвірними відношеннями, а на основі дистрибутивних та парадигматичних відношень цих морфів у системі морфеміки в цілому, що дозволяє використати метод аналогій у морфемній сегментації слів. Метод аналогій у морфемному аналізі

ґрунтується на загальносистемній мотивованості морфних структур. Морфна структура є константною і вона конденсує, відображає ті морфологічні процеси, які спрацьовують у системі мови в цілому. Тому морфна структура конкретно взятої словоформи може відображати не тільки актуальні морфологічні процеси, а й діахронічні і ті, що потенційно можливі. Отже, морфна структура словоформи може розглядатися як константна структура, що не конструюється, а наслідуються новоутворенням від свого аналогічного зразка.

«Морфемний різновид морфологічного аналізу є найбільш послідовною реалізацією принципу сегментації мовної форми – сегменти морфної структури словоформи можуть бути охарактеризовані як її кінцеві складові. Для здійснення такого членування необхідно задіяти для зіставлення максимально широке, практично необмежене коло словоформ. Це пов'язано з тим, що на рівні кінцевих складових морфологічна структура слова є наслідком не тільки процесів слово- і формотворення, а також і результатом аналогічного впливу вже існуючих морфологічних структур» [5:58].

МБД дозволяє послідовно і системно застосувати метод аналогій у морфемному аналізі, оскільки ця система автоматично класифікує лексику української мови у спільнокореневі та спільноафіксальні, а також спільноструктурні вибірки й уможливує зіставлення морфних структур слів на матеріалі великих лексичних масивів. Відповідно лексикографічна система «Морфограф», що будується на принципі розмежування морфемного і словотвірного аналізів, представлятиме морфні структури слів, визначені за методом аналогій з урахуванням дистрибутивних та опозиційних характеристик морфів// морфем у системі морфеміки української мови.

Морфні структури слів характеризуються чіткими схемами синтагматичної реалізації морфів і регулюються правилами комбінаторики та дистрибуції, що визначає семантичне і формальне об'єднання морфів у слово як знакову одиницю вищого рангу. "Здатність до комбінаторики – загальна і обов'язкова властивість одиниць мови, зумовлена загальносистемними і фундаментальними властивостями – дискретністю і неоднорідністю. Ієрархічність і лінійність також належать до фундаментальних властивостей мовних одиниць, які зумовлюють спосіб реалізації комбінаторики" [27:267].

Системний дистрибутивний опис морфів ставиться в центрі уваги сучасних лінгвістичних досліджень (Н. Ф. Клименко [19; 20; 21], Є. А. Карпіловської [18; 21; 26], Я. Горещького [33], Г. Ф. Єфремової [8], Р. Г. Зятковської [16], Х. Ковалик [32]), у яких описується валентність морфів, їх позиційні характеристики у різних морфних структурах слів, що зумовлює визначення комплексної одниниці морфемної системи мови – морфної // морфемної структури слова. Типовий характер морфемних структур,

регулярна повторюваність і відтворюваність у словах дозволяє визначити їх як окремий тип мовних одиниць – онтологічні одиниці структурного характеру морфемного рівня мовної системи. Опис морфемних структур слів прийнято описувати за допомогою моделей, в яких латинськими літерами позначається функціональний тип морфа, наприклад, *ви/чит/к/а* PRSF. В основу такого моделювання покладено фундаментальну лінгвістичну ідею: формула будови, що регулярно повторювана і відтворювана у мові, визначається як окрема одиниця мовної структури поряд із фонемою та морфемою [27]. Формули будови морфемних структур – інваріант класу морфемних структур слів, які утворюють парадигму цих структур. Цей методологічний принцип було покладено в основу комп'ютерного моделювання структурно-функціональних зв'язків морфем у слові при створенні МБД: базовою одиницею реєстру МБД є слово із програмною процедурою визначення морфемної будови слова (*задньоязиковий* RDSFIGRLSNFP), яка моделює морфну структуру цього слова (*зад-нь-о-язик-ов-ий* RSIRSF). Тому лінгвістична інформація про структурну організацію морфемної будови слова, закладена у реєстрі МБД, буде представлена у лексикографічному описі у вигляді кількісно-функціональних моделей морфних структур слів.

Формалізоване представлення інформації про морфну структуру слова в МБД забезпечує автоматичне формування реєстрів префіксів, коренів, суфіксів, інтерфіксів, постфіксів та флексій (система флексій формується обмежено, тільки на матеріалі початкових форм), що формують систему морфеміки української мови. У лексикографічній системі ставиться завдання описати дистрибуцію кожного морфа як елемента цієї системи, визначивши його валентність у морфних структурах усіх спільнокореневих чи спільноафіксальних слів.

Наступним завданням лексикографічного опису є визначення парадигматичних відношень морфів у системі морфеміки української мови [12; 13; 14:]. Виконання цього завдання базується на теоретичному визначенні морфемі як інваріантно-варіантної парадигматичної одиниці мови. Інваріантність і варіативність – дві характеристики мовної одиниці, що є взаємовизначальними. Парадигматична цілісність морфемі базується на цих двох ознаках, тому в лексикографічному описі морфема буде представлена як інваріантно-варіантна модель, що передбачає групування в один клас всіх морфів, у яких морфема реалізована у мовленні. Процедура ідентифікації морфів однієї морфемі враховує інваріантні ознаки морфемі: до однієї морфемі належать морфи, що мають тотожні значення та повну або часткову фонологічну // графемну подібність плану вираження. Інваріантно-варіантна модель морфемі також передбачає моделювання варіантних ознак морфемі: часткову фонологічну // графемну відмінність плану вираження морфів, що

зумовлює опис варіативності морфеми у термінах аломорфів (морфів, що знаходяться у відношенні додаткової дистрибуції) або варіативних морфів (морфів, що знаходяться у відношенні вільного варіювання). Таким чином модель морфемної парадигми може бути представлена: 1) фонологічно // графемно тотожними морфами; 2) аломорфами та варіоморфами.

В описі морфеми як цілісної парадигматичної одиниці найскладнішою процедурою є визначення значення морфеми як інваріантної ознаки. У сучасній лінгвістичній семантиці, яка ставить завдання описати значення основних мовних знаків: морфем, слів, речень, тестів, – опису семантики морфеми присвячено найменше досліджень. Морфема – це особливий мовний знак, оскільки він не виконує референції: не співвідносний безпосередньо з референтом (явищем дійсності). Морфема актуалізує своє значення тільки у складі слова, яке виконує номінативну або дейктичну референцію. Тому в мовознавстві побутує думка про те, що морфема має асоціативне значення (І. Бодуен де Куртене). Значенню морфеми не можна дати єдиного трактування, тому що кореневі морфеми, дериваційні афіксальні морфеми, реляційні афіксальні морфеми виражають різні типи значень. Найбільш глибоко в лінгвістиці описані значення дериваційних та реляційних афіксів, найменше – корневих морфеми. На першому етапі конструювання лексикографічної системи "Морфограф" ставиться завдання описати парадигматику корневих морфем, оскільки у МБД системно представлена аломорфія та омонімія коренів, а опис афіксів ще не закінчений. За особливістю семантики кореневі морфеми визначаються як такі, що виражають найбільш конкретне значення – елемент лексичного значення слова. Проте значення кореня також не має актуального референта, тому що встановлюється не в корені одного слова, а абстрагується як спільна семантика всіх спільнокорневих слів, які здійснюють референцію до різних явищ дійсності як предметних, так і абстрактних. Отже значення кореня також має узагальнений і категорійний характер.

У процесі укладання бази даних омонімічних коренів ставилося завдання не описати значення кореня, а протиставити за значенням корені-омоніми. Такий методологічний підхід ґрунтується на теоретичному постулаті Ф. Сосюра про поняття значущості елемента мовної системи. Розуміючи семантику морфеми, як таку, що має не значення, а значущість, О. С. Герд [6] визначає морфему напівзнаком. У нашому дослідженні це теоретичне положення набуло такого процедурного формулювання: якщо два морфи з однаковим планом вираження протиставляються у системі мови за своєю семантикою, то ці морфи належать до різних морфем. При цьому значення морфем не описувалось, але у семантичному протиставленні коренів бралась до уваги семантика непохідних слів, у яких ці кореневі

морфи реалізовані (до кожного слова у МБД подано його тлумачення за тлумачними словниками української мови). Безумовно, таке визначення значення кореня не може претендувати на семантичний опис, оскільки кореневий морф, який реалізовується у десятках або декількох сотнях слів у кожній новій морфемній дистрибуції може звужувати або розширювати своє значення. Проте на сьогодні в описі семантики мовних одиниць української мови лінгвістичне дослідження може посылатися тільки на тлумачні словники, тому приписування кореневій морфемі значення непохідного слова, в якому реалізований один із її морфів, на цьому етапі вивчення морфемної семантики є, на нашу думку, єдиним способом визначення семантики корневих морфем. Такий принцип опису семантики корневих морфем буде використано і в лексикографічній системі.

II. Формування лексичного реєстру лексикографічної системи «Морфограф»

Як уже зазначалося, система «Морфограф» укладатиметься автоматично на матеріалі МБД, яка на сьогодні складає ≈ 200 тис. слів української мови. У МБД кожному слову за методикою комп'ютерного моделювання структурно-функціональних зв'язків морфів у слові приписана модель, яка визначає межі та функціональний тип морфів і водночас є формулою програмної процедури морфемної сегментації слова. Наприклад: зад/нь/о/язик/ов/ий R3S5I6R10S12F14 – де латинськими літерами позначається функціональний тип морфа Р – префікс, R – корінь, S – суфікс, F – флексія, I – інтерфікс, X – постфікс, а цифрами – межі морфа через порядковий номер (із початку слова) кінцевої графеми кожного морфа. Формалізований опис морфемної структури слова у вигляді програмної процедури відображає функціональну закономірність організації слова на морфемному рівні – RSIRSF, а субстанціальне (графемне) вираження цієї структури представлено через кількісно-графемну модель (зад – 3; нь – 5; о – 6; язик – 10; ов – 12; ий – 14). Графемно-цифрові межі морфів у базі даних автоматично конвертовані латинською літерою: R3S5I6R10S12F14 = RDSFIGRLSNFP.

Систематизація лінгвістичної інформації про організацію морфемної будови слів української мови у МБД АСМСА за методикою комп'ютерного моделювання структурно-функціональних зв'язків морфів у слові [1; 2; 3; 7; 15] покладена в основу автоматичного конструювання лексикографічної системи "Морфограф". Програмне забезпечення МБД виконує операції автоматичної класифікації лексики української мови за різними параметрами організації морфемної будови слова: кількістю морфем у слові, моделями морфемних структур, заданим афіксальним чи корневим морфем, – а також автоматично укладає кореневі та афіксальні словники.

Лексикографічні завдання, які ставляться у створенні системи «Морфограф», зумовлюють необхідність редагування лексичного реєстру МБД. МБД сьогодні активно використовується в морфемному аналізі лексики українських текстів у Корпусі української мови, тому для системи «Морфограф» буде сформована автономна морфемна база даних, яка за класифікаційними параметрами лінгвістичної інформації та обсягом лексичного реєстру якісно і кількісно відрізнятиметься від МБД АСМСА. Необхідність формування нової бази даних зумовлена також системоорієнтованим характером створюваної лексикографічної системи, що передбачає представлення її в мережі Інтернет як автономного інформаційного лінгвістичного ресурсу. У формуванні лексичного реєстру планується: 1) вилучення власних назв та загальних назв із онімними коренями, оскільки оніми характеризуються особливою семантикою та морфною структурою і утворюють велику ономастичну систему, що заслуговує на окремий лексикографічний опис; 2) вилучення буквених та звукових абревіатур, оскільки семантика та морфемна будова цих слів також потребує окремого теоретичного підходу та принципів морфемного аналізу; 3) доповнення регулярними дериватами, які не ввійшли до реєстрів традиційних словників і не представлені у МБД: синтетичними формами ступенів порівняння якісних прикметників, відприкметниковими прислівниками, дієприкметниками, абстрактними іменниками-вербативами із суфіксом -нн-я; 4) поповнення словозмінними формами іменників, прикметників, займенників, числівників, дієслів, у кореневих та афіксальних морфемах яких відбуваються морфонологічні альтернативи. Всі завдання планується виконати автоматично, використовуючи ресурси морфемної бази даних і морфологічної бази даних системи АГАТ лабораторії комп'ютерної лінгвістики та алгоритмів автоматичного морфемного синтезу слів.

III. Структура та класифікаційні можливості лексикографічної системи «Морфограф»

Теоретичні засади та завдання, які ставляться в лексикографічному описі, визначають структуру лексикографічної системи «Морфограф». Ця система складатиметься з двох інтерактивних електронних словників: "Словника морфемних структур слів" та "Словника морфем". Кожен словник будуватиметься в інтерактивному режимі користувачем за заданими пошуковими опціями, а також передбачатиме опцію "скопіювати" до кожного автоматично сформованого реєстру чи списку.

"Словник морфемних структур слів" конструюється за такими зонами: 1) морфемна будова слова; 2) моделі морфемних структур.

У першій зоні "Морфемна будова слова" реєстровою одиницею виступає слово, поділене на морфеми, із приписаною морфемною моделлю,

наприклад: *ви/чит/к/а* PRSF; *ви/чит/а/ми* PRSF; *ви/чит/а/н/ий* PRSSF; *ви/чит/а/н/о* PRSSS. У цій зоні користувач може автоматично групувати реєстрові одиниці у різноманітні вибірки за такими пошуковими параметрами: 1) "кількість коренів": всі слова; прості слова; складні слова; 2) "граматична характеристика": всі слова; обрана у випадному списку частина мови; 3) «прямий алфавітний список реєстру»: пошук за буквою, частиною слова, цілим словом, морфом; 4) "інверсійний алфавітний список реєстру": пошук за буквою, частиною слова, морфом.

У другій зоні «Моделі морфемних структур» реєстровою одиницею є модель морфемної структури слова: PRSF, PRSSF, PRSSS. На першому рівні пошукових параметрів визначається обсяг лексичної вибірки за такими опціями: 1) «кількість коренів»: всі слова; прості слова; складні слова; 2) «граматична характеристика»: всі слова; обрана у випадному списку частина мови; 3) «кількість морфем у слові»: всі слова; обрана у випадному списку кількісно-морфемна характеристика. За сформованою лексичною вибіркою автоматично укладається список моделей морфемних структур слів із поданою інформацією про кількість слів, в яких кожна модель реалізована. На другому рівні класифікаційних опцій здійснюється: 1) автоматичне формування списку слів за обраною моделлю, наприклад: PRSF *в/тич/к/а*, *в/тих/оньк/а*, *в/тул/к/а*, *в/тиш/ниш/я*; 2) автоматичне формування списку морфів за обраним типом морфа у моделі, наприклад: PRSF Р (*без-*, *в-*, *ви-*, *во-*, *ді-*, *до-*, *з-*, *за-*, *зав*); R (*-бав-*, *-бавл-*, *-баат-*, *-баж-*, *-бач-*, *-бг-*, *-би-*, *-бир-*, *-біг-*); 3) автоматичне формування парадигми моделі із морфемним представленням докореневої та посткореневої зон, наприклад: PPRSSF: *ви-с-R-в-ува-ти*, *не-за-R-ну-т-ий*, *не-за-R-а-н-ий*, *не-на-R-ова-н-ий*.

«Словник морфем» також поділяється на дві зони: «Словник морфів» та «Словник корневих морфем». У «Словнику морфів» реєстровою одиницею виступає морф як знакова синтагма слова без врахування омонімії та аломорфії морфів. На першому етапі, як і в попередніх словниках, визначається обсяг лексичної вибірки за такими пошуковими опціями: 1) «кількість коренів»: всі слова; прості слова; складні слова; 2) «граматична характеристика»: всі слова; обрана у випадному списку частина мови. На матеріалі визначеної лексичної вибірки автоматично формується реєстр морфів за опцією «функціонально-позиційний тип морфа»: префікс, корінь, суфікс, інтерфікс, постфікс, флексія. На другому етапі користувач може автоматично здійснювати такі класифікаційні операції: 1) формувати реєстр слів, в яких реалізований обраний зі списку морф або заданий користувачем, наприклад, суфікс -ун-: *авіа/двиг/ун/0* RRSF, *бав/ун/я* RSF, *балак/ун/0* RSF, *балак/ун/ств/0* RSSF, *бельк/от/ун/0* RSSF; 2) формувати список морфемних моделей із визначенням реалізації обраного морфа, наприклад, RR-ун-F, R-ун-F, R-ун-SF, RS-ун-F.

«Словник кореневих морфем» укладатиметься на матеріалі трьох баз даних: бази даних морфемних структур, бази даних омонімічних коренів, бази даних аломорфічних коренів. Програмний взаємозв'язок цих баз забезпечує автоматичну класифікацію слів української мови у спільнокореневі вибірки з урахуванням кореневої омонімії та аломорфії. Реєстровою одиницею цього словника є коренева морфема, визначена за умовно вихідним морфом, тим, який реалізований у непохідному слові спільнокореневих слів, наприклад, вихідним морфом у спільнокореневих словах *берег/ти*, *береж/інн/я*, *енерг/о/з/беріг/а/ж/уч/ий* визначається дієслівний кореневий морф -берег-.

Словникова стаття представляє модель кореневої морфеми: парадигматичний клас морфів, визначений за інваріантно-варіантними ознаками морфеми як цілісної парадигматичної одиниці. Лексикографічне моделювання кореневої морфеми враховує спільне значення класу формально тотожних або частково тотожних за формою морфів і визначає п'ять зон словникової статті: 1) вихідний морф; 2) непохідне слово із визначеним вихідним морфом; 3) частина мови непохідного слова, яка визначає категорійну семантику морфеми; 4) тлумачення слова: лексичні значення всіх ЛСВ; 5) варіанти морфеми: аломорфи, варіоморфи для частково тотожних за формою морфів; один морф, що представляє реалізацію морфеми тільки в однакових за формою морфах. У першій і п'ятій зонах подається кількісна характеристика реалізації морфеми у словах української мови, наприклад:

-берег-
(71);

2. *берег-ти*;
3. дієслово: процесуальність;
4. 1) не давати витратитися, зберігати цілим; 2) тримати в доброму стані; 3) дбайливо ставитися до кого-, чого-небудь; 4) вживати заходів для захисту когось, чогось; 5) тримати в пам'яті;
5. -берег- (15),
-береж- (39),
-беріг- (17).

«Словник кореневих морфем» будуватиметься автоматично користувачем за такими пошуковими опціями: 1) всі корені; 2) корені обраної у випадному списку частини мови; 3) один корінь, заданий користувачем.

На другому рівні інтерактивного конструювання словника користувач працює із словниковою статтею і може поглиблювати лінгвістичну інформацію про морфему, використовуючи нові класифікаційно-пошукові опції. За вибором варіанта морфеми формуються списки спільнокореневих слів, в яких реалізований обраний морф, наприклад:

-берег-	-береж-	-беріг-
<i>берег/ти</i> RF	<i>береж/інн/я</i> RSF	<i>в/беріг/а/нн/я</i> PRSSF
<i>берег/ти/ся</i> RFX	<i>береж/к/ий</i> RSF	<i>енерг/о/з/беріг/а/ж/уч/ий</i> RIPRSSF
<i>з/берег/ти</i> PRF	<i>береж/лив/ий</i> RSF	<i>з/беріг/а/ти/ся</i> PRSFX

За опцією – морфемні моделі слів – автоматично формується список морфемних моделей, в яких реалізовані морфи кореневої морфеми, і список слів, в яких реалізована кожна модель (див. Табл.1):

ТАБЛИЦЯ 1.
МОДЕЛІ МОРФЕМНИХ СТРУКТУР
СПІЛЬНОКОРЕНЕВОЇ ЛЕКСИКИ

	-берег-	-береж-	-беріг-
PRF	<i>по/берег/ти</i> <i>з/берг/ти</i> <i>о/берег/ти</i>		
PRSF	<i>о/берег/0/0</i>	<i>з/береж/енн/я</i> <i>з/береж/ен/ий</i> <i>о/береж/н/ий</i>	<i>в/беріг/а/ти</i> <i>о/беріг/0/0</i> <i>о/беріг/а/т</i> <i>и</i>
PRSSF		<i>з/береж/ен/ість/0</i> <i>о/береж/н/ість/0</i> <i>о/береж/н/иш/ий</i>	<i>в/беріг/а/нн/я</i> <i>з/беріг/а/нн/я</i> <i>з/беріг/а/ч/0</i>

Висновок

Система «Морфограф», яку найближчим часом планується представити як лінгвістичний інформаційний ресурс у мережі Інтернет, представляє новий етап у розвитку сучасної української комп'ютерної лексикографії як в аспекті автоматичного конструювання цієї системи, так і в аспекті пошуково-класифікаційних можливостей користувача. Ця система має інтегрований лексикографічний характер, що визначається поєднанням і взаємодоповненням різних морфемних словників: словника морфемної будови слів, словника моделей морфемних структур слів, словника аіксальних морфів, словника кореневих морфів, словника кореневих морфем. Крім того, ці словники є інтерактивними, вони виконують функцію ефективного комп'ютерного інструмента, що автоматично формує за вибором користувача саме той лінгвістичний матеріал, який необхідний йому в доборі, обробленні та систематизації лінгвістичної інформації. Лексикографічний опис репрезентує великий лексичний матеріал і глибинний морфемний аналіз, що ґрунтується на узагальненні провідних теоретичних ідей сучасної морфотактики та морфонології, тому за критерієм лінгвістичної інформативності система «Морфограф» претендує на повноту і цілісність лінгвістичного опису морфемної системи української мови. Довідко-інформаційна

система з морфеміки української мови «Морфограф» може використовуватися у різноманітних наукових та навчальних завданнях сучасної лінгвістики.

Література

1. Алексеенко Л. А. Параметризованная база данных поэтической речи как источник и инструмент филологических студий / Л. А. Алексеенко, Н. П. Дарчук, О. М. Зубань, В. М. Сорокин // Международная конференция "Прикладная лингвистика без границ". Материалы конференции. – Санкт-Петербург, 2004. – С. 80 – 8
2. Алексієнко Л. А. Методика створення автоматизованої системи морфемно-словотвірного аналізу (АСМСА) слів української мови / Л. А. Алексієнко, Н. П. Дарчук, О. М. Зубань // Наукова спадщина професора С. В. Семчинського. Збірник наукових праць. – К., 2001, - Ч. 1. – С. 38 – 49.;
3. Алексієнко Л. А. Автоматизована система морфемно-словотвірного аналізу як інструмент лінгвістичних досліджень / Л. А. Алексієнко, Н. П. Дарчук, О. М. Зубань, В. М. Сорокін// Матеріали міжнародної науково-практичної конференції "Проблеми прикладної лінгвістики". // Мова. Науково-теоретичний часопис. – Одеса, 2004. –С. 19 – 26.;
4. Бирюкова Л. П. О функциональном подходе к морфемному анализу / Л. П. Бирюкова // Актуальные проблемы русского словообразования. Учен. зап. ташкент. ун-та., 1975. – Т. 143. – С. 295.
5. Богданов С. И. Форма слова и морфологическая форма / С. И. Богданов. – Санкт-Петербург: Изд. Санкт-Петербургского университета, 1997. – 245 с.
6. Герд А. С. Семантика морфемы: значение или значимость (valeur)? / А. С. Герд// Структурная и прикладная лингвистика. – Вып. 1. – Л., 1983. – С. 47 – 52.
7. Дарчук Н. П. Комп'ютерне ануотування тексту: результати і перспективи: монографія / Н. П. Дарчук. – К., 2013. – 543 с.
8. Ефремова Т. Ф. Опыт описания современного русского языка на уровне морфов: Автореферат дис. канд. филол. наук. Московский университет. / Т. Ф. Ефремова – М., 1970. – 18 с.
9. Зубань О. Електронні частотні морфемні словники в Корпусі української мови / О. Зубань// Науковий вісник Східноєвропейського національного університету імені Лесі Українки, Серія: Філологічні науки. - № 3 (304). — Луцьк, 2015. – С. 315 - 320.
10. Зубань О. Частотні морфемні словники в Корпусі української мови - джерело стилеметричних досліджень / О. Зубань // Acta Universitatis Palackianae Olomucensis Philologica 104 - 2016: UCRAINICA VII: Současná ukrajínistika Problémy jazyka, literatury a kultury. — Olomouc, 2016. – S. 22
11. Зубань О. М. Електронні словники у Корпусі української мови: параметри пошуку та систематизації мовних одиниць / О. М. Зубань // Мовні і концептуальні картини світу. – Вип. 54. – К., 2016.
12. Зубань О. М. Методологічні питання суфіксальної парадигматики українського дієслова / О. М. Зубань // Українське мовознавство.– Вип. 21., – К., 1997. – С. 16 – 25.
13. Зубань О. М. Морфонологічний аспект суфіксальної парадигматики українського дієслова / О. М. Зубань // Українське мовознавство.– Вип. 21. – К., 1997. – С. 73 – 81.
14. Зубань О. М. Створення морфемної бази даних: принципи опису морфем як інваріантної одиниці / О. М. Зубань // Українське мовознавство. – Вип. 23. –К., 2001. – С. 52 – 64.
15. Зубань О. Н. Задачи и методы автоматического морфемного анализа в Корпусе украинского языка / Зубань О. Н. // Актуальные проблемы современной прикладной лингвистики. – Минск, 2016. – С. 123 – 131.
16. Зятковская Р. Г. Суффиксальная система современного английского языка / Р. Г. Зятковская.– М., 1980. – 147 с.
17. Карпіловська Є. А. Кореневий гніздовий словник української мови / Є. А. Карпіловська. – К., 2002.
18. Карпіловська Є. А. Суфіксальна підсистема сучасної української мови: будова та реалізація: монографія / Є. А. Карпіловська. – К., 1999. – 297 с. 5.
19. Клименко Н. Ф. Морфеміка і словотворення як частини категоріальної граматики української мови / Н. Ф. Клименко // Праці 3-го Міжнародного конгресу україністів.– Харків, 1996.– С. 199.
20. Клименко Н. Ф. Основи морфеміки сучасної української мови: навч. посіб / Н. Ф. Клименко. – К.,1998. – 182 с.
21. Клименко Н. Ф. Словотвірна морфеміка сучасної української мови / Н. Ф. Клименко, Є. А. Карпіловська.– К., 1998.
22. Котова Н. В. О многообразии морфем в славянских языках / Н. В. Котова, М. О. Янакиев // Славянская филология. – Вып. X. – М., 1978. – С. 4 – 8.
23. Морфемна структура слова: монографія / Т. О. Грязнухіна, Н. Ф. Клименко, Л. І. Комарова та ін.. – К., 1979. – 327 с.
24. Перебийціс В. І. Традиційна та комп'ютерна лексикографія. Навч. посібник / В. І. Перебийніс, В. М. Сорокін. – К.: Вид. центр КНЛУ, 2009. –218 с.
25. Полюга Л. М. Морфемний словник / Л. М. Полюга. – К.: Радянська школа, 1983. – 462с.
26. Словник афіксальних морфем української мови / за ред. Н. Ф. Клименко; [уклали: Н. Ф. Клименко, Є. А. Карпіловська, В. С. Карпіловський, Т. І. Недозим]. – К., 1998. – 434 с.
27. Солнцев В. М. Язык как системно-структурное образование / В. М. Солнцев. – М.:Наука, 1971.– С. 123.
28. Частотні словники Корпусу. - Режим доступу від 15.03.2017: <http://www.mova.info/article.aspx?l1=210&DID=5215>.

29. Шкільний словотвірний словник сучасної української мови / Клименко Н. Ф., Карпілоська Є. А., Кислюк Л. В. – К., 2005.

30. Шуба П. П. О компонентах конфикса в русском языке / П. П. Шуба // Развитие современного русского языка. Словообразование. Членимость слова. – М.: Наука, 1975. – С. 249 – 253.

31. Яценко І. Т. Морфемний аналіз: Словник-довідник / І. Т. Яценко. – К., 1980 – 1981. Т. 1 - 2.

32. Gramatyka współczesnego języka polskiego. Morfologia. – Warszawa, 1984. – S. 513 – 545.

33. Horecki I. Morfematičká struktura slovensciny / I. Horecki. – Bratislava, 1964. – 423 s.

O projekcie internetowego słownika przysłów polskich

Robert Kościelniak

magister Instytutu Filologii Słowiańskiej, Uniwersytet Szczeciński, Polska, E-mail: koscielniakrobert@gmail.com

Jakub Olas

magister Instytutu Filologii Słowiańskiej, Uniwersytet Szczeciński, Polska, E-mail: jakub.olas@gmail.com

Proverbs are the wisdom of nations, however Polish lexicography does not pay them as much attention as the other language structures. First Polish proverbial collections are dated back to XVII century, however the proverbs themselves became the subject of interest of linguistics only in XX century when „A New Book of Polish Proverbs and Proverbial Expressions” was published by J. Krzyżanowski. Since then no one has published a dictionary that could be compared with Krzyżanowski monumental work. Nevertheless, the times and technology had changed and they had serious influence on paremiology. The project of an Internet dictionary of Polish proverbs could be a response to the challenges facing modern paremiologists focusing on anti-proverbs, proverbs’ translations issues or some paremiological units’ extinction. The article contains short information about the history of Polish paremiography, modern trends in paremiology and proposed structure of an Internet dictionary of Polish proverbs.

Słowa kluczowe — paremiologia, paremiografia, jednostka paremiczna, przysłowie, antyprzysłowie, słownik.

I. Wstęp

Przysłowia nazywa się mądrością narodu nieprzypadkowo. W przysłowiach zapisana jest jego historia, kultura i obyczaje. Te jednozdaniowe, obrazowe jednostki języka, dzięki swojej krótkiej, rymowanej formie przekazywane dawniej ustnie z pokolenia na pokolenie, mogły przetrwać wiele wieków do czasu, kiedy pojawiły się pierwsze drukowane słowniki zawierające przysłowia zbierane w różnych regionach kraju i wśród różnych środowisk. Polska tradycja paremiograficzna sięga pierwszej połowy XVII wieku, niemniej jednak przez długi czas miała ona głównie charakter czysto amatorski, etnograficzny, antykwaryczny, popularyzatorski. Dopiero w XX wieku zainteresowanie przysłowiami nabrało prawdziwie naukowego charakteru [9: 9]. Przysłowie ze sfery zainteresowań literaturoznawstwa przeszło wtedy do sfery zainteresowań językoznawstwa i m.in. zyskało status jednostki języka, chociaż niektórzy polscy badacze się z tym poglądem nie zgadzają (por. A. Bogusławski). Obecnie silnie zarysowuje się tendencja do

modyfikowania przysłów, co jest dowodem na to, że są one obecne nie tylko w samym języku, ale i w świadomości jego użytkowników. Takie nowe wersje kanonicznych przysłów bywają nazywane antyprzysłowiami. Termin ten jednak dla wielu badaczy jest niezbyt trafny i, szczególnie na gruncie językoznawstwa polskiego, na określenie takich jednostek używa się terminu *innowacja paremiczna*. XXI wiek, postęp technologiczny, a w szczególności rozwój Internetu dostarczyły nowych narzędzi w badaniach leksykograficznych, paremiograficznych i paremiologicznych. Tradycyjna, drukowana leksykografia powoli ustępuje miejsca leksykografii w formie elektronicznej. Ta druga, jest dostępna dla większego grona odbiorców, a przede wszystkim może być poszerzana w każdej chwili, bez konieczności czasochłonnego przedrukowywania słowników i publikowania nowych wydań.

Na chwilę obecną w polskiej przestrzeni internetowej na próżno szukać naukowego serwisu poświęconego stricte przysłowiom. Celem niniejszego artykułu jest omówienie projektu internetowego słownika przysłów polskich, który nawiązując bezpośrednio do tradycji, wypełni widoczną lukę w polskiej leksykografii, a właściwie paremiografii elektronicznej.

II. Krótko o historii paremiologii i paremiografii

„Wydawałoby się, że nie ma nic łatwiejszego, niż podanie precyzyjnej definicji przysłowia. Ale można chyba stwierdzić, że istnieje więcej prób definicji, niż samych przysłów” [7: 49]. Pogląd ten, wyrażony przez W. Miedera, trafnie ujmuje aktualny stan badań nad zdefiniowaniem przysłowia jako jednostki języka i przypisaniem pewnych wyróżniających go cech. Chociaż pierwsze próby podjął już w czasach starożytnych Arystoteles, a następnie w roku 1500 Erazm z Rotterdamu (we wstępie do swojego dzieła opisał istotę, budowę i funkcje przysłowia), to do dzisiaj nie powstała jedna, uniwersalna definicja. Podobnie zjawisko to