КОМП'ЮТЕРНА ЛІНГВІСТИКА

КОМП'ЮТЕРНЕ ЛЕКСИКОГРАФІЧНЕ МОДЕЛЮВАННЯ МОРФЕМНОЇ СИСТЕМИ УКРАЇНСЬКОЇ МОВИ

ОКСАНА ЗУБАНЬ

Міністерство освіти і науки України
Київський національний університет імені Тараса Шевченка

Оксана Зубань

КОМП'ЮТЕРНЕ ЛЕКСИКОГРАФІЧНЕ МОДЕЛЮВАННЯ МОРФЕМНОЇ СИСТЕМИ УКРАЇНСЬКОЇ МОВИ

Монографія

КИЇВСЬКИЙ ВТД УНІВЕРСИТЕТ 2020

УДК 811.161.2'32'374
ББК 81.2.411.4–4
391

Рецензенти:

д-р філол. наук, проф. **Н.П. Дарчук**
(Інститут філології Київський національний університет імені Тараса Шевченка);
д-р філол наук, проф. **О.П. Левченко**
(Національний університет "Львівська політехніка");
д-р філол. наук, старш. наук. співроб. **Л.П. Кислюк**
(Інститут української мови НАН України).

Рекомендовано до друку Вченою радою Інституту філології Київського національного університету імені Тараса Шевченка (протокол № 7 від 28 січня 2020 року)

Зубань О.М.

391 Комп'ютерне лексикографічне моделювання морфемної системи української мови : монографія. – К. : Видавничо-поліграфічний центр "Київський університет", 2020. – 259 с.

ISBN 978-966-439-819-7

У розглянуто лінгвістичні теоретико-методологічні засади комп'ютерного моделювання знакових одиниць морфемної системи, процесів морфемного і словотвірного аналізів слів української мови та принципи проектування лінгвістичних баз даних автоматизованої системи морфемно-словотвірного аналізу (АСМСА) української мови. Обґрунтовано інфологічну інтегральну лексикографічну модель текстоорієнтованого інтерактивного електронного морфемного словника української мови та проаналізовано даталогічний етап проектування монографії лексикографічної системи частотних морфемних словників, автоматично укладених за текстовими вибірками Корпусу української мови. Проведено ілюстративний аналіз пошукових та класифікаційних можливостей інтерактивних електронних частотних морфемних словників, за статистичними даними яких проведено стилеметричне дослідження ідіостилів українських поетів.

Для науковців, викладачів, аспірантів, студентів філологічних спеціальностей.

УДК 811.161.2'32'374
ББК 81.2.411.4–4

ISBN 978-966-439-819-7

З М І С Т

РОЗДІЛ 2
АВТОМАТИЗОВАНА СИСТЕМА МОРФЕМНО-СЛОВОТВІРНОГО АНАЛІЗУ СЛІВ УКРАЇНСЬКОЇ МОВИ – БАЗА ЗНАНЬ УКРАЇНСЬКОЇ МОРФЕМОЛОГІЇ ТА СЛОВОТВОРУ

РОЗДІЛ 3
КОМП'ЮТЕРНА ПАРАМЕТРИЗАЦІЯ УКРАЇНСЬКОМОВНОГО ТЕКСТУ НА МОРФЕМНОМУ РІВНІ СТРУКТУРИ

РОЗДІЛ 4
ЕЛЕКТРОНІ ЧАСТОТНІ МОРФЕМНІ СЛОВНИКИ – КОМП'ЮТЕРНИЙ ІНСТРУМЕНТ СТИЛЕМЕТРИЧНИХ ДОСЛІДЖЕНЬ

П Е Р Е Д М О В А

Сучасна практика укладання словників та новітня теоретична лексикографія із розвитком комп'ютерних технологій набули стрімкого зростання й сприяли створенню великого фонду лексикографічних праць, що описують мову за різними аспектами, на основі різних теоретичних підходів. Українська лексикографія виконує свою основну функцію: відповідає інтересам суспільства, вирішуючи наукові, навчальні, пізнавальні та культурні завдання, що зумовлює визначення в межах цього напряму прикладної лінгвістики навчальної, довідкової та наукової лексикографії. Тенденція розвитку наукової лінгвістичної лексикографії з другої половини XX ст. характеризується проникненням у всі галузі сучасного мовознавства й виходом за межі лексикології, тому об'єктами лексикографічних описів стають одиниці і явища всіх мовних рівнів. «Між лексикографією та іншими мовознавчими дисциплінами існує тісний зв'язок. З одного боку, лексикографія становить, так би мовити, матеріальну базу лінгвістичних досліджень, постачаючи для них перевірений і відповідним чином систематизований матеріал, без якого неможливе будь-яке серйозне науково-лінгвістичне узагальнення, а з іншого, – своїми успіхами лексикографія завдячує розвиткові суміжних з нею галузей науки про мову» [СУЛМ 1973: 292]

"Лексикографічний ефект" [1] наукового лінгвістичного пізнання визначає необхідність лексикографічної систематизації матеріалу дослідження в наукових розвідках. «На початку XXI століття стало очевидним, що практично будь-яка лінгвістична ідея може бути оформлена й представлена лексикографічно, і ця обставина породила феноменальне наукове явище, яке, здається, не знає меж у кількості й різноманітності словників, що відображають практично всі аспекти існування, розвитку й функціонування мови» [Скляревская 2013: 588 – 589].

Різноманітність сучасних словників як за об'єктами опису, так і за параметрами та концепціями відображає, з одного боку, традиційність лексикографічної практики, а з іншого, пошук нових підходів і принципів в укладанні словників, що втілюються насамперед у нові комп'ютерні лексикографічні проєкти. Перевага комп'ютерних (електронних) словників над традиційними паперовими словниками є незаперечним фактом, тому що електронний словник «...відкриває цілу низку можливостей, недоступних у паперовій формі: динамічний пошук і добір лексики за будь-яким критерієм, динамічне сортування й групування, перегляд словника в різних видах, швидкий перехід із словника до корпусу текстів, видача інформації в різних форматах. Крім того, електронна форма не має обмежень за обсягом, що суттєво впливає на форму представлення інформації в традиційних друкованих словниках» [Поляков 2008: 11].

[1] Поняття "лексикографічний ефект" використано у трактуванні В. Широкова [Широков 1998].

Сучасна українська комп’ютерна лексикографія має у своєму доробку багато різногалузевих електронних словників, зокрема досвід у конструюванні електронних словників колективу лабораторії комп’ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка засвідчують лексикографічні продукти, частина з яких представлена на лінгвістичному порталі mova.info: Відкритий словник новітніх термінів [ВСНТ 2018]; Українсько-італійський граматичний словник дієслів [УІГС 2001]; серія частотних словників (Частотний словник художньої прози [ЧСХП 2017], Частотний словник сучасної української публіцистики [ЧССУП 2017], Частотний словник наукового стилю [ЧСНС 2017], Частотний словник сучасної української поетичної мови [ЧССУПМ 2017] та ін.); Електронний граматичний словник української літературної мови (словозміна) [ЕГСУЛМ 2018]; Чотиримовний словник термінів із хімії та фізики [ЧСТХФ 2017]; Система електронних навчальних словників "Глоса" (англо-український, українсько-англійський навчальний словник) [ГЛОСА 2003]; Труднощі англійського слововживання для українців [ТАС 2018]; Таблиця окремих термінів, вживаних у диференційній геометрії (багатомовний перекладний словник) [Мацюк 2018]; Тезаурус комп’ютерної лексикографії [Сірук 2018]; Короткий українсько-сербський словник сполучуваності слів [КУСС 2005]; Українсько-російсько-англійський тезаурус із лінгвістичної термінології [УРАТ 2018]; Семантичний словник української мови [ССУМ 2005] та ін..

Поняття словника, представленого в електронній формі, визначається в лінгвістичній наукометрії різними термінами: "машинний словник", "електронний словник", "автоматичний словник", "автоматизований словник", "комп’ютерний словник". Диференціювання термінів зумовлене відмінностями в структурі словників, їх функціональним призначенням та програмно-технічними інструментами, проте існує ряд спільних ознак, які лежать в основі визначення поняття "електронний словник". У монографії "електронний словник" трактується у широкому розумінні за дефініцією Л. Нелюбіна: «електронний словник – це будь-який упорядкований, відносно обмежений масив лінгвістичної інформації, представлений у вигляді списку, таблиці або переліку, зручному для розміщення в пам’яті ЕВМ і оснащеному програмами автоматичного оброблення та поповнення» [Нелюбин 1983].

Створенням електронних словників займається нова лінгвістична галузь – комп’ютерна лексикографія. Єдиного термінологічно визначення цієї галузі також немає, науковці використовують терміни "електронна лексикографія", "обчислювальна лексикографія", "кібернетична лексикографія", "машинна лексикографія", "автоматична лексикографія", що зумовлено невизначеністю й самого терміна "комп’ютерна лінгвістика". Термін "комп’ютерна лексикографія" є найбільш усталеним у мовознавстві, проте в нових лінгвістичних дослідженнях, зокрема в русистиці, пропонується використовувати термін "електронна лексикографія", що є достатньо мотивованим. «На наш погляд, більш доречним є використання

терміна "електронна лексикографія", тому що в цьому випадку відбувається зміна акценту з технічного компонента на формат даних, що дозволяє розширити поняття й відійти від обмеження процесу розроблення та використання словника тільки за допомогою комп'ютера, оскільки зараз багато лексикографічних ресурсів установлюються не тільки на комп'ютери, але й на інші електронні пристрої (смартфони, планшети тощо)» [Палкова 2015: 89]. Термін "електронна лексикографія" використовується і в зарубіжному науковому дискурсі, зокрема в назві відомої міжнародної наукової конференції в цій галузі, яка проводиться із 2009 р., використовується термін електронна лексикографія (Conference on electronic lexicography «eLex») [eLex 2019].

Дискусійним є також питання про статус комп'ютерної лексикографії в системі наук про мову. Більшість науковців схильні визначати комп'ютерну лексикографію в широкому розумінні терміна – розділ прикладної (комп'ютерної) лінгвістики, що ставить лексикографічні завдання [Чепик 2006], [Палкова 2015], [Карпіловська 2006], [Голубкова 2014] та ін., проте існує вузьке трактування цієї галузі, як розділу лексикографії [Демська 2009], [Селегей 2008] та ін. «Пропонуючи виділяти у структурі лексикографії класичну та комп'ютерну лексикографію, ми схильні все-таки зараховувати останню до лексикографії, а не прикладної лінгвістики, хоча й з прикладною і, особливо, з корпусною, лінгвістикою комп'ютерна лексикографія має багато спільного» [Демська 2009: 22].

В. Селегей, фахівець із лінгвістичних досліджень компанії ABBYY, визначає галузь, що займається конструюванням електронних лексикографічних продуктів, частиною лексикографії й обмежує її завдання тільки укладанням електронних словників "з нуля". «Зазвичай розуміється, що словник на комп'ютері – це введений у нього паперовий словник, оснащений зручними інструментами пошуку й відображення. Тобто, укладачі електронних словників переливають старе добре лексикографічне вино у нові електронні міхи. Комп'ютерна лексикографія як галузь прикладної лінгвістики, що продукує такі словники, позбавлена власного мовного предмета. На її долю залишається тільки ефектна демонстрація канонічного змісту. Ми б хотіли запропонувати інший погляд, за яким комп'ютерна лексикографія є особливим напрямом практичної лексикографії зі своїми власними підходами не тільки до відображення, але й до змісту словника. Ми вважаємо, що електронний словник – це особливий лексикографічний об'єкт, у якому можуть бути реалізовані [...] численні продуктивні ідей, не затребувані з різних причин у паперових словниках» [Селегей 2008].

На наше переконання, у комп'ютерній лексикографії виконуються обидва завдання (і конвертація паперових словників, і укладання нових словників), тому у монографії приймається широке розуміння комп'ютерної лексикографії як розділу прикладної (комп'ютерної) лінгвістики в такому визначенні: «Комп'ютерна лексикографія – це прикладна наукова дисципліна в мовознавстві, що вивчає методи, технологію й окремі прийоми

використання комп'ютерної техніки в теорії і практиці укладання словників. Спеціальні програми – бази даних, комп'ютерні картотеки, програми оброблення текстів – дозволяють в автоматичному режимі формувати словникові статті, зберігати словникову інформацію й обробляти її» [Чепик 2006: 276]. Відповідно, комп'ютерна лексикографія ставить і нові вимоги до лексикографів, які повинні мати міждисциплінарні знання і уміння, що виходять далеко за межі компетентності традиційних лексикографів. «Нова інформаційна парадигма потребувала від укладачів серйозної перекваліфікації, переоцінки стратегій укладання лексикографічних джерел, змісту інформації й способів взаємодії з «новим» типом користувача «digital native» – «уродженцем Інтернету». Укладач втілює в собі дві іпостасі: вченого-лінгвіста і розробника програмного забезпечення (tool-builder)» [Голубкова 2014: 75 – 76].

На основі концепції широкого трактування комп'ютерної лексикографії можна визначити чотири основні завдання цієї прикладної галузі: 1) використання комп'ютера в укладанні традиційних паперових словників; 2) автоматична конвертація паперових словників у електронні лексикографічні системи; 3) автоматичне / автоматизоване укладання електронних (комп'ютерних) словників; 4) автоматичне / автоматизоване укладання автоматичних словників.

На підставі визначених завдань ми розмежовуємо поняття "електронний словник" та "автоматичний словник", як електронні ресурси, що адресовані користувачеві-людині ("електронний словник") та машині ("автоматичний словник"). Автоматичні словники призначені для лінгвістичних процесорів автоматичного оброблення тексту (інформаційного пошуку, машинного перекладу, реферування тощо) і за своєю структурою, одиницями-об'єктами та іншими аспектами відрізняються від електронних словників, для яких обов'язковим є людино-машинний інтерфейс. У сучасній комп'ютерній лінгвістиці спектр лексикографічних завдань набагато ширший, ніж укладання електронного словника: кожне прикладне завдання, спрямоване на створення лінгвістичного процесора, є проявом «лексикографічного ефекту в інформаційних системах» [Широков 1998], тому що будь-яка лінгвістична електронна система працює на основі бази даних, яка є автоматичним резидентним словником.

Комп'ютерна лексикографія, як і комп'ютерна лінгвістика загалом, у виконанні своїх завдань базується на трьох основних атрибутах: технічні пристрої і засоби; програмне забезпечення; лінгвістичне забезпечення. Третій аспект міждисциплінарного синтезу знань, представленого в концепції комп'ютерної лінгвістики, належить до компетенції комп'ютерного лінгвіста, що передбачає: 1) лінгвістичні знання (фундаментальної, структурної, математичної лінгвістики) про об'єкт дослідження; 2) розроблення адекватної концептуальної моделі об'єкта на базі формальних ознак; 3) розроблення алгоритмів автоматичного лінгвістичного аналізу та синтезу.

Аналізуючи сучасний стан розвитку комп'ютерної лінгвістики, відома українська лінгвістка Є. Карпіловська наголошує про незаперечну важливість у парадигмі комп'ютерної лінгвістики теоретичного складника – структурної лінгвістики, що виявляється у моделюванні мовних об'єктів та процесів. «З появою комп'ютера структурна, математична та прикладна лінгвістика дедалі активніше перетворюються на лінгвістику комп'ютерну. Остання з удосконаленням, з одного боку, систем опрацювання мовної інформації, ускладненням моделей штучного інтелекту, а з іншого, – з розвитком обчислювальної техніки та її програмного забезпечення не обмежується розв'язанням суто практичних завдань мовознавства, а дедалі глибше просувається в царину теоретичної, фундаментальної лінгвістики та лінгвосеміотики. Про це свідчать сьогодні й комп'ютерні словники з відповідними моделями опису семантики мовних одиниць, і багаторівнева анотація змісту текстів у корпусах національних мов, і розгалужені можливості видобування інформації у сучасних пошукових машинах та інші здобутки комп'ютерної лінгвістики. Усі ці можливості комп'ютерного опрацювання інформації, знань у мовній формі ґрунтовано на моделях мовних об'єктів і процесів, створених представниками структурної і математичної лінгвістики. Вони і складають те, що в сучасній інформатиці прийнято називати лінгвістичним забезпеченням комп'ютера – *лінгвером* поруч із його апаратним, інструментальним (*хардвером*) та математичним (алгоритмічним і програмним) забезпеченням (*софтвером*)» [Карпіловська 2019: 19].

Автоматичний лінгвістичний аналіз, спрямований на будь-яку прикладну задачу – лексикографічну (у вузькому розумінні) чи в NLP-галузі[2], ґрунтований на комп'ютерному моделюванні об'єктів і процесів різних мовних рівнів. Н. Дарчук називає таку комп'ютерну модель мови комп'ютерною граматикою: «Для автоматичного аналізу українського тексту нами створено комп'ютерну граматику, яка є ієрархічним комплексом комп'ютерних моделей: морфемно-словотвірної, морфологічної, синтаксичної моделі, побудованих на основі формальних, точних і однозначних правил. Ці моделі можна вважати дослідницькими, тому що закладені у граматики алгоритмічні правила призводять до виявлення того чи іншого мовного явища (морфів, словоформ з їх частиномовними і категорійними характеристиками, словосполучень, дерев залежностей речень тощо). [...]. Розроблені моделі є моделями аналізу, індуктивними, несемантичними і детерміністськими (структурними)» [Дарчук 2013, 28].

Ґрунтуючись на постулатах структурної лінгвістики у вивченні мови-об'єкта, комп'ютерна лінгвістика керується концепцією рівневої організації мовної системи, що продукує й відповідну гносеологічну стратифікацію в межах комп'ютерної лінгвістики: комп'ютерна фонетика, комп'ютерна морфологія, комп'ютерна морфемологія, комп'ютерний словотвір,

[2] Natural-language processing – оброблення природної мови.

комп’ютерний синтаксис, комп’ютерна семантика. У комп’ютерній лінгвістиці на сьогодні немає чіткого розмежування між морфемним та словотвірним аналізами, тому що словотвірний аналіз у комп’ютерних системах базується на морфемному аналізі. Ця закономірність відображена й у назвах електронних довідкових систем, створених в українській комп’ютерній лінгвістиці: морфемно-словотвірний фонд української мови, автоматизована система морфемно-словотвірного аналізу. Тому завдання автоматичного морфемного і словотвірного аналізів об’єднуються в одному лінгвістичному напрямі, який можна назвати комп’ютерною морфемологією.

Комп’ютерна морфемологія [3], спрямована на вивчення морфемної підсистеми мови та структури тексту на морфемному рівні організації, ставить завдання створити комп’ютерну модель, яка адекватно імітує аналітичну діяльність лінгвіста і вможливлює здійснення автоматичного морфемного та словотвірного аналізу і синтезу. Теоретичні і практичні здобутки сучасної комп’ютерної лінгвістики в українському та російському мовознавстві визначають такі завдання в галузі комп’ютерної морфемології:

1) створення дослідницьких науково-довідкових морфемних (морфемно-словотвірних) баз даних: [Алексієнко 2001], [Алексієнко 2004а], [Зубань 2006], [Клименко 2014б], [Карпіловська 2006] та ін.;

2) автоматичне / автоматизоване укладання традиційних (паперових) та електронних морфемних і словотвірних словників: [АРСУН 2013], [Зубань 2015], [Зубань 2016а], [Зубань 2017], [Карпіловська 2019], [Кретов 1999], [Пазельская 2009а], [Поликарпов 1998], [САМУК 1998], [Широков 2011], [Zuban 2017] та ін.;

3) створення текстоорієнтованих автоматичних морфемних сегментаторів та аналізаторів тексту в корпусах мов: [Алексеенко 2004], [Гришина 2009], [Дарчук 2013], [Дарчук 2008], [Зубань 2017а], [Кукушкина 2005], [Кукушкина 2006а], [СтилеАнализатор-2 2014], [Татевосов 2009], [Токтонов 2006], [Ляшевская 2016], [Zuban 2019] та ін.;

4) створення морфемноорієнтованих семантичних мереж в інформаційно-пошукових системах: [Пацкин 2002], [Пацкин 2004];

5) створення комп’ютерних інструментів морфемно-словотвірного синтезу слів [Карпіловська 1986], [Карпіловська 1990].

Сформульовані завдання мають прикладне та теоретико-лінгвістичне (дослідницьке) спрямування, причому ці два вектори взаємодоповнюють один одного: теоретичні знання про морфемну будову слова є підґрунтям для досягнення практичної мети, а практичні результати поглиблюють

[3] Із метою розмежування омонімії термінів, термін "морфемологія" вживається на позначення розділу лінгвістики, а термін "морфеміка" на позначення морфемної підсистеми мови. Морфемологія [від гр. μορφή = morphḗ ’форма’ і гр. λόγος ’слово, вчення’] – гносеологічне поняття, наука про морфемну будову мови. Уперше цей термін був запропонований О. Моїсеєвим [Моисеев 1987] і використовується філологами-україністами КНУ імені Тараса Шевченка [Козленко 2014: 30].

теоретичні знання. Визначені завдання є взаємозумовленими й пов'язаними з укладанням лінгвістичних баз даних, які в кінцевому результаті можуть бути представлені користувачеві у формі електронних словників, тому на перетині завдань комп'ютерної лексикографії та комп'ютерної морфемології сформувалась галузь, яку можна назвати комп'ютерна морфемна лексикографія.

Практика використання комп'ютера в укладанні паперових морфемних та словотвірних словників використовується в лексикографії з 70-х років XX ст.: частотний словник З. Оліверіуса «Морфемы русского языка» [Оливериус 1976], кореневий словник Д. Уорта, О. Козака, Д. Джонсона «Русский словообразовательный словарь» [Уорт 1983]; морфемний словник чеської мови Е. Славічкової [Slavíčková 1975]. Процедура укладання та короткий огляд цих словників представлено Н. Клименко в статті «Составление словарей морфем с помощью ЭВМ» [Клименко 2014].

Українська морфемна лексикографія, започаткована відомими традиційними словниками «Морфемний аналіз: Словник-довідник» [Яценко 1980], «Морфемний словник» [Полюга 1983], «Українсько-російський словотворчий словник» [Сікорська 1995] із розвитком теорії морфемології та словотвору поповнилася лексикографічними працями, які ставлять нові лексикографічні завдання: «Словник афіксальних морфем української мови» [САМУК 1998]; «Кореневий гніздовий словник української мови» [Карпіловська 2002]; «Шкільний словотвірний словник сучасної української мови» [ШСССУМ 2005]; «Етимологічний словник запозичених суфіксів і суфіксоїдів в українській мові» [Селігей 2014]; «Українські словогрона. Словогрона духу» [Різників 2015]. Серед цих словників взірцем словникарства нової доби – комп'ютерної української лексикографії – стали лексикографічні праці ([САМУК 1998], [Карпіловська 2002], [ШСССУМ 2005]) науковців відділу структурно-математичної лінгвістики Інституту мовознавства ім. О.О.Потебні НАН України, а нині відділу лексикології, лексикографії та структурно-математичної лінгвістики Інституту української мови НАН України. Ці словники існують у традиційному паперовому форматі, а також у форматі комп'ютерних копій паперових словників, вони укладалися за допомогою комп'ютера на матеріалі електронної бази даних морфемно-словотвірного фонду української мови (МСФ), створення якого (1988 – 1991 рр.) започаткувало новий етап у розвитку української морфемної лексикографії – комп'ютерну морфемну лексикографію.

Важливість конструювання комп'ютерних баз даних у морфемології та словотворі була визначена в колективній статті «Морфемно-словотвірний фонд української мови як дослідницька та інформаційно-довідкова система» (автори: Н. Клименко, Є. Карпіловська, Л. Комарова, Т. Недозим, Т. Іванова). «У сучасній україністиці гостро відчувається потреба в різноаспектній лексикографічній параметризації морфемного та словотвірного рівнів мови. Для її реалізації необхідні словники кожного з цих мовних рівнів, які б подавали їх одиниці разом із характеристиками парадигматичних відношень

(перш за все через спектр їхнього аломорфного варіювання), і такі правила виведення з простих одиниць комплексів різної складності, що ґрунтувалися б на кількісних, частотних позиційних та комбінаторних характеристиках морфем. Лише за умови всебічного обстеження морфемної та словотвірної будови одиниць мови можливе достатньо обґрунтоване розв'язання кардинальних проблем української морфеміки [...]. Йдеться про створення машинної інформаційної системи зі спеціальними процедурами її ведення, виконання певних типів логіко-класифікаційних задач або надання користувачеві певної довідки на його запит. Обробка однієї такої фактичної бази за різними процедурами дозволить одержувати різні машинні продукти: словники, інвентарі, таблиці сполучуваності мовних одиниць (морфем, наприклад), довідкову інформацію» [Клименко 2014б: 545].

Інформаційно-довідкова дослідницька система МСФ, яка постійно поповнюється новою лексикою, була також використана у фундаментальних дослідженнях, за якими написані відомі підручники з морфемології, статті та монографії, зокрема: «Основи морфеміки сучасної української мови» [Клименко 1998], «Словотвірна морфеміка сучасної української мови» [Клименко 1998а], «Суфіксальна підсистема сучасної української мови: будова та реалізація» [Карпіловська 1999], «Динамічні процеси в сучасному українському лексиконі» [Клименко 2008], «Сучасна українська словотвірна номінація: ресурси та тенденції розвитку» [Кислюк 2017] та ін.. Морфемно-словотвірний фонд української мови – це база знань, спрямована на виконання функції своєрідного довідника для лінгвіста-дослідника, і, безсумнівно, є надзвичайно важливою для організації повномасштабного лінгвістичного дослідження.

Визнання сучасної лінгвістики лінгвістикою тексту [Плунгян 2009] уже стало аксіомою в мовознавстві. Тому завдання морфемного та словотвірного аналізу переносяться на вивчення функціонування морфемних структур, словотвірних моделей, різних типів морфем у конкретних текстах, жанрах, стилях з урахуванням статистичних даних. У такій зміні акцентів по-новому формулюється й лексикографічне завдання – створення текстоорієнтованих електронних морфемних словників за текстовими даними корпусів мов. Особливої уваги сьогодні заслуговують ті електронні лінгвістичні продукти, які спрямовані на аналіз тексту і є пошуковими системами, здатними в автоматичному або автоматизованому режимі вилучати з тексту інформацію про мовні одиниці. Корпусна лексикографія стала новим етапом у розвитку словникарства: «Спочатку комп'ютери скромно асистували процесу укладання словників, потім забезпечували перекачування паперової інформації на комп'ютерні носії, і нарешті, стали основою укладання електронних словників, що називається, з нуля. У сучасних роботах із теоретичної лексикографії зазначається, що вже тридцять років лінгвістичні корпуси служать основою укладання словників. Вони фактично разом із

системою пошуку стали невід'ємною частиною того, що називають «lexicographic tool» [11][4] » [Голубкова 2014: 76 – 77].

Автоматичне укладання текстоорієнтованих морфемних словників можливе лише в тих корпусах текстів, які оснащені лінгвістичними процесорами автоматичної морфемної сегментації та автоматичного морфемного аналізу. У Корпусі української мови [КУМ 2019], який створено колективом комп'ютерних лінгвістів Інституту філології Київського національного університету імені Тараса Шевченка під керівництвом доктора філологічних наук Н. Дарчук [Дарчук 2013], [Дарчук 2010],[Дарчук 2010а], тексти параметризуються на чотирьох рівнях текстової структури: морфологічному, морфемному, синтаксичному та семантичному. Автоматичний аналіз текстів проводиться з використанням великих лінгвістичних баз даних, над створенням яких наш колектив працює багато років. Результатом роботи в цьому напрямі в галузі морфемології та словотвору було створення впродовж 2001 – 2020 рр. автоматизованої системи морфемно-словотвірного аналізу (АСМСА) української мови, яка використовується у функціях:

1) інформаційно-довідкової лексикографічної системи з морфемології та словотвору української мови;

2) бази даних у системі автоматичного морфемного сегментатора початкових форм / текстових слововживань української мови.

3) бази даних у системі автоматичного конструювання частотних морфемних словників за лексичними реєстрами початкових форм (лем) текстових вибірок КУМ.

Мета монографії – обґрунтувати інфологічні і даталогічні принципи комп'ютерного лексикографічного моделювання та етапи конструювання лінгвістичної бази даних АСМСА й лексикографічної системи текстоорієнтованих електронних частотних морфемних словників української мови. Відповідно до мети в монографії ставляться завдання:

1) обґрунтувати теоретико-методологічні засади комп'ютерного моделювання об'єктів та процесів морфемної системи української мови;

2) обґрунтувати лінгвістичні основи концептуальної моделі АСМСА;

3) проаналізувати схему даних даталогічної моделі та етапи конструювання АСМСА;

3) проаналізувати дослідницькі функції АСМСА та використання цієї системи в текстоорієнтованому автоматичному морфемному аналізі;

4) обґрунтувати концептуальну лексикографічну модель електронного морфемного словника;

5) проаналізувати схему даних даталогічної моделі текстоорієнтованої електронної лексикографічної системи частотних морфемних словників (ЧМС);

[4] Tarp S. Beyond Lexicography: New Visions and Challenges in the Information Age // H. Bergenholtz, S. Nielsen & S. Tarp (eds.) Lexicography at a Crossroads: Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow. Bern : Peter Lang, 2009. P. 17–32

6) проілюструвати пошукові та класифікаційні можливості інтерактивних електронних ЧМС;

7) за статистичними даними електронних ЧМС провести стилеметричне дослідження ідіостилів українських поетів.

Монографія складається із чотирьох розділів, висновків та додатків, у яких подано ілюстративний матеріал результатів роботи АСМСА та скріншоти інтерфейсу ЧМС. У створенні АСМСА брали участь викладачі кафедри української мови та прикладної лінгвістики Київського національного університету імені Тараса Шевченка О. Зубань, Л. Алексієнко, Н. Дарчук, а також інженер-програміст лабораторії комп'ютерної лінгвістики університету В. Сорокін, тому в другому та третьому розділах монографії описуються результати експериментальної колективної роботи. Дослідження, представлені в першому та третьому розділах, є самостійними.

У першому розділі проаналізовано теоретико-методологічні засади комп'ютерного моделювання елементів морфемної системи мови, ґрунтовані на принципах формалізації лінгвістичного аналізу знакових одиниць; визначено концептуальні та даталогічні особливості таких комп'ютерних моделей: моделі морфемної структури (ММС) слова, моделі автоматичної морфемної сегментації слова, моделі АСМСА, лексикографічної моделі електронного морфемного словника.

Комп'ютерна модель морфемної структури слова є гносеологічним аналогом онтологічної морфемної будови слова й виконує функцію базового конструкта у наступних етапах моделювання. Саме ця модель забезпечує ізоморфність моделі автоматичної морфемної сегментації, яка ставить завдання адекватно зімітувати діяльність лінгвіста в процесі морфемного аналізу слова. У процесі створення автоматичного морфемного аналізу слів української мови було побудовано два типи даталогічних ММС слова. Ці моделі побудовані за різними методологічними принципами:

1) модель морфемної структури посткореневої зони дієслівних основ побудована за принципом дистрибутивно-опозиційного протиставлення графем на морфемному шві між коренем та суфіксальною послідовністю основи слова;

2) модель морфемної структури слова створена за функціонально-кількісно-графемними формулами морфемної сегментації.

Перша модель реалізована лише в алгоритмічній моделі автоматичної морфемної сегментації, а друга модель реалізована програмно у побудові АСМСА та в лінгвістичному процесорі автоматичної морфемної сегментації початкових форм (лем) у Корпусі української мови.

Особлива увага в першому розділі надається розробленню лексикографічної концептуальної моделі електронного морфемного словника. У цій моделі враховується структурна організація морфемної системи мови й ставиться завдання описати морф – базову одиницю морфеміки – в усіх типах внутрішньосистемних відношень: синтагматичних, парадигматичних, ієрархічних. Багатоаспектність структурних відношень

визначає принцип інтегральності лексикографічного опису, а електронна інтерактивна форма словника змінює аспект принципу інтегральності в усталеному лексикографічному розумінні: вимога інтегральності ставиться не до словникової статті, а до макроструктури лексикографічної системи. Лексикографічна електронна система складається із баз даних (БД), у яких описувана реєстрова одиниця наскрізно пов'язана із усіма полями баз даних. Словникова стаття конструюється користувачем самостійно в інтерактивному режимі інтерфейсу, який має багато входів у систему і об'єднує три зони-словники, поєднані інтерактивною навігацією через гіперпокликання:

1) словник морфів (морфем), диференційований за функціональними типами морфем, із представленням повного реєстру слів, у яких реалізований кожен морф;

2) словник морфемних структур слів, типізованих в окремі групи за символьними моделями морфемних структур.

3) тлумачно-морфонологічний словник морфем, у якому морфема представлена як інваріантно-варіантний конструкт.

Комп'ютерна реалізація лексикографічної моделі на великих лексичних і текстових масивах уможливлює доповнення визначених типів словників статистичною інформацією, тому всі три типи словників є частотними.

Програмні інструменти АСМСА забезпечують автоматичне конструювання інтегральної електронної лексикографічної системи ЧМС. У проекції на лексичний масив мови це лексикографічне завдання ще не було виконано, але було розроблено проєкт лексикографічної системи «Морфограф» [Зубань 2017]. У дослідженні українськомовного тексту в Корпусі української мови інтегральна лексикографічна модель набула текстової орієнтації і була реалізована у двох типах електронних частотних словників: словнику морфів (морфем) та словнику морфемних структур слів, які представлені в мережі Інтернет [ЧСКУМ 2017].

Другий розділ монографії присвячено опису даталогічної моделі АСМСА: структури даних, процедури автоматизованого укладання та етапів розбудови БД, класифікаційно-пошукових можливостей системи.

АСМСА складається із двох БД: морфемної бази даних (МБД) та словотвірної бази даних (СБД), кожна з яких структурується на БД-таблиці. Бази даних укладалися в автоматизованому режимі: лінгвіст-укладач проводив морфемний та словотвірний аналізи за допомогою комп'ютерних інструментів, а результати цього аналізу імпортувалися в бази даних. У межах АСМСА можна здійснювати такі операції:

1) автоматичну класифікацію лексики в спільнокореневі, спільноафіксальні та спільноструктурні вибірки;

2) автоматичне формування реєстру афіксальних та кореневих морфем;

3) автоматичне / автоматизоване укладання словотвірних гнізд.

Типи виконуваних завдань, обсяг реєстру морфемних структур слів (≈ 200 тис.) та структурування даних на 9 БД-таблиць демонструють

поліфункціональність цієї системи й широкий спектр проаналізованих морфемних та словотвірних об'єктів, явищ, процесів, що визначає АСМСА базою знань із морфеміки та словотвору української мови.

Методика автоматизованого укладання баз даних АСМСА стала предметом вивчення в університетських курсах із комп'ютерної лексикографії та автоматичного морфемного аналізу й описана в університетських підручниках [Дарчук 2008], [Перебийніс 2009]. АСМСА[5] як інформаційно-довідкова система з морфеміки та словотвору української мови активно використовується філологами КНУ ім. Т. Шевченка в різноманітних наукових дослідженнях, у читанні університетських курсів із морфемології та словотвору української мови.

У третьому розділі представлено апробацію АСМСА у функції автоматичного морфемного аналізатора в Корпусі української мови. У процесі створення автоматичної системи морфемного аналізу в КУМ із метою оптимізації пошуку на великих текстових масивах ми відмовилися від методу морфемної анотації текстових слововживань. Морфемна розмітка не проводиться, тексти Корпусу виступають тільки матеріалом для укладання частотних морфемних словників. На сьогодні укладено двадцять ЧМС за різними текстовими вибірками КУМ. Із цими словниками можна ознайомитися на лінгвістичному порталі mova.info [ЧСКУМ 2018]. Такі словники можуть бути укладені для будь-якої текстової вибірки Корпусу української мови на замовлення користувача.

Лексикографічна БД ЧМС укладається автоматично за даними текстового модуля КУМ, морфологічного модуля КУМ та морфемного автоматичного аналізу, роботу якого забезпечує АСМСА. Схема даних БД ЧМС є універсальною для всіх словників, а імпорт даних для кожної текстової вибірки здійснюється окремо: у результаті кожен ЧМС працює за автономною БД. СКБД цієї бази даних, як і БД АСМСА, розроблена на основі MS Access, а також в укладанні баз даних використано авторське програмне забезпечення, створене В. Сорокіним за допомогою об'єктно-орієнтованих мов програмування C++ та C #. За моделлю організації даних обидві БД є реляційними. Інтерфейс лексикографічної системи морфемних словників створений у вигляді веб-додатка ASP.Net. Він побудований з урахуванням інтерактивного характеру електронних словників: користувач за обраними опціями в інтерактивному режимі автоматично будує потрібні словникові статті.

Кожен ЧМС складається із трьох інтерактивних функціональних зон:

1) реєстр одиниць за обраним функціональним типом морфеми або моделлю морфемної структури слова;

[5] АСМСА ще не представлена як електронний лексикографічний продукт для широкого кола користувачів, але колектив лабораторії комп'ютерної лінгвістики КНУ ім. Т. Шевченка може надати доступ до цієї системи, з якою можна працювати в он-лайн режимі (електронна адреса для звернення: oxana.mell.zuban@gmail.com).

2) реєстр слів текстової вибірки, у якому реалізована, вибрана в першій зоні, морфемна одиниця;

3) конкорданс до вибраного в 2-ій зоні слова. Перша і друга зони подають кількісну та статистичну інформацію про одиниці реєстру.

Інтерактивна навігація між двома словниками (ЧС морфем, ЧС морфструктур) та трьома зонами у межах кожного словника здійснюється через гіперпокликання.

Лексикографічна система ЧМС в Корпусі української мови – це зручний інформаційний лінгвістичний інструмент, який покликаний допомогти користувачеві в автоматичному режимі проводити різноманітні дослідження в галузі морфемології, лексикології, семантики, синтаксису, стилеметрії на базі великого обсягу ілюстративного текстового матеріалу; здійснювати різноманітні класифікаційні операції з лексичним матеріалом за текстовими даними різних стилів та дискурсів; отримувати нові знання про семантичну та формальну структуру українського слова.

У четвертому розділі монографії представлено результати використання статистичних даних частотного словника (ЧС) морфемних структур слів у стилеметричному дослідженні поетичних ідіостилів Тараса Шевченка, Лесі Українки, Ліни Костенко, Василя Стуса. Статистичне дослідження стилістично маркованого тексту керується тенденцією об'єднання двох аспектів:

1) вивчення морфемної системи ідіостилю з урахуванням усіх морфемних структур слів, які використані автором у текстах, що визначає індивідуальний добір цих одиниць за статистикою їх вживання, а отже, робить кожну морфемну структуру слова стилістично маркованою одиницею;

2) вивчення тих морфемних структур слів, які формують функціонально-стилістичні та емоційно-експресивні стилерозрізнювальні ознаки ідіостилю.

Базовими в стилеметричному дослідженні виступають два параметри:

1) кількісно-структурний: кількість морфем у моделі морфемної структури слова та особливості організації функціональної морфемної будови;

2) статистичний: абсолютна частота (f) і відносна частота (p – %) ММС у текстовій вибірці та лексичному реєстрі ідіостилю. Відносна частота ММС у лексичному реєстрі інтерпретується як питома вага або лексична продуктивність, а в текстовій вибірці – індекс покриття тексту.

За кількісно-структурним параметром проведено аналіз морфемної довжини слів, морфемної глибини слів, валентності функціональних типів морфем у морфемній структурі слова.

Статистичні параметри ММС слів аналізується за схемою розробленої моделі морфемної статистичної структури ідіостилю, що визначає розподіл ММС слів на три статистичні групи (високочастотні, середньочастотні,

низькочастотні) через зіставлення відносної частоти слів однієї ММС у двох рангових списках ММС, укладених за:

1) лексичним реєстром початкових форм (лем), що формує поняття – морфемна статистична структура лексикону;

2) текстом (слововживаннями), що формує поняття – морфемна статистична структура тексту.

На базі визначеного операційного поняття – "моделі морфемної статистичної структури ідіостилю" – проаналізовано статистичну "поведінку" ММС в ідіостилях чотирьох українських поетів і проведено порівняльний аналіз статистичних даних, який визначає стилеметричні ознаки кожного ідіостилю.

Модель морфемної статичної структури ідіостилю може використовуватися як еталонна статистична модель у вивченні різних стилів та ідіостилів однієї мови, а також у типологічних дослідженнях флективних мов, за умови використання однакових символьних моделей морфемної структури слова. Ця статистична модель демонструє вищий рівень узагальнення кількісної моделі тексту, ніж лексична статистична модель, тому що кількість ММС у тексті, порівняно з кількістю слів, зменшена в сотні разів.

***

Висловлюю щиру подяку моїм УЧИТЕЛЯМ – кандидату філологічних наук Людмилі Антонівні Алексієнко та доктору філологічних наук Наталії Петрівні Дарчук, їхня колежанська й дружня підтримка, а також конструктивна критика сформували мене як лінгвіста-дослідника й сприяли реалізації задуму цієї монографії. Особливу вдячність адресую інженеру-програмісту Вікторові Михайловичу Сорокіну, який упродовж багаторічної співпраці сумлінно виконує обов’язки програміста в наших проектах і втілює у "віртуальне життя" всі наші ідеї.

Також висловлюю щиру вдячність рецензентам – доктору філологічних наук, професору, завідувачеві кафедри прикладної лінгвістики Національного університету «Львівська політехніка» Олені Петрівні Левченко та доктору філологічних наук, старшому науковому співробітнику Інституту української мови НАН України Кислюк Ларисі Павлівні за глибокий аналіз монографії та висловлені поради.

Окремо хочу подякувати студентам Інституту філології, які брали активну участь у цьому проекті з першого етапу його створення, а також допомагають у редагуванні та розбудові АСМСА на нинішньому етапі.

У М О В Н І П О З Н А Ч Е Н Н Я

f – абсолютна частота
F – флексія
I – інтерфікс
p – відносна частота
P – префікс
R – корінь
S – суфікс
X – постфікс
АГАТ – автоматичний граматичний аналіз тексту
АОТ – автоматичне оброблення текстів
АСМСА – автоматизована система морфемно-словотвірного аналізу
БД – база даних
БЗ – база знань
ВС – Василь Стус
КУМ – Корпус української мови
ЛК – Ліна Костенко
ЛУ – Леся Українка
МБД – морфемна база даних
МБЗ – морфемна база знань
ММС – модель морфемної структури
МС – морфемний сегментатор
МСФ – морфемно-словотвірний фонд
НКРМ – Національний корпус російської мови
ПМ – природна мова
СБД – словотвірна база даних
СБЗ – словотвірна база знань
СКБД – системи керування базою даних
СУБД – системи управління базою даних
ТШ – Тарас Шевченко
ФМ – формальна мова
$x_{i.}$ – варіанта варіативного ряду
ЧМС – частотний морфемний словники
ЧС – частотний словник

Р О З Д І Л 1

МОДЕЛЮВАННЯ ОБ'ЄКТІВ ТА ПРОЦЕСІВ МОРФЕМНОЇ СИСТЕМИ МОВИ У КОМП'ЮТЕРНІЙ МОРФЕМОЛОГІЇ

1.1. Теоретико-методологічні засади моделювання об'єктів морфемної системи мови

Проблема статусу морфеми та процедури морфемного аналізу ставилась у центрі досліджень відомих лінгвістів XX ст.: Б. де Куртене [Б. де Куртене 1963], Л. Блумфілда [Блумфилд 1968], З. Харріса [Харрис 1965], В. Скалічки [Скаличка 1967], Ч. Хоккета [Hockett 1947], Г. Глісона [Глисон 1961], А. Мартіне [Мартине 1963], Л. Єльмслєва [Ельмслев 1960], Ж. Вандрієса [Вандриес 1964], Дж. Грінберга [Гринберг 1963], М. Трубецького [Трубецкой 1967] та ін.. На сучасному етапі розвитку лінгвістики модифікації в дефініції морфеми, у практиці її виділення, у функціональній класифікації морфем і описі їх суттєвих характеристик зумовлені тенденцією до загальної переорієнтації морфології як науки в межах антропоцентричної лінгвістичної парадигми. Сьогодні актуально постає проблема встановлення залежності морфологічних структур від фонетичних, синтаксичних, семантичних, дискурсивних і прагматичних факторів. Увага лінгвістів зосереджується на динамічних аспектах мовної діяльності, спрямована на поняття морфологічного процесу й морфологічного правила, що визначило використаня методу моделювання в описі морфемної системи мови та словотворення.

На думку відомого американського морфолога Ст. Андерсона [Anderson 1982], у морфології визначаються дві основні проблеми:

1) проблема фонологічної будови морфеми і її аломорфів;

2) проблема об'єднання морфів у певні морфологічні структури.

Акцент із техніки морфологічного аналізу, орієнтований на сегментацію словоформ та ідентифікацію виділених відрізків, переноситься на пояснення процесів організації морфологічних структур на синхронному рівні мови. У такій зміні акцентів морфема виступає як відомий, раніше заданий мовний об'єкт, як одиниця вже встановлена і зафіксована в спеціальних словниках, а об'єктами вивчення морфемології стає морфемна структура слова та фонологічна / графемна структура морфеми, що описуються за допомогою різноманітних моделей.

Вивчення морфемної структури слова – одна з центральних проблем лінгвістичних досліджень у сучасній україністиці: питання морфемної будови слова української мови, словотвірних процесів, фонологічної та графемної будови морфів, принципів об'єднання морфів у морфологічні структури, комбінаторики морфів у морфемній структурі слова розглядаються в працях багатьох українських лінгвістів. Бібліографія цих праць надзвичайно велика, тому перерахуємо найбільш значущі, на нашу

думку: І. Ковалик [Ковалик 1958], [Ковалик 1961], [Ковалик 1971]; Н. Клименко [Клименко 1973], [Клименко 1975], [Клименко 1996], [Клименко 1998]; Є. Карпіловська [Карпіловська 1999]; І. Козленко [Козленко 2014]; Л. Кислюк [Кислюк 2017]; В. Горпинич [Горпинич 1999]; В. Ґрещук [Ґрещук 1995]; І. Савченко [Савченко 1990]; М. Пещак [Пещак 1966]; колективна монографія «Морфемна структура слова» [МСС 1979] Т. Бондаренко [Бондаренко 1974]; О. Зубань [Зубань 1998]. Зміна аспекту вивчення морфемної системи мови визначила чільне місце в сучасному українському мовознавстві лексикографічних праць нового покоління: «Словник афіксальних морфем української мови» [САМУК 1998]; «Кореневий гніздовий словник української мови» [Карпіловська 2002]; «Шкільний словотвірний словник сучасної української мови» [ШССУМ 2005], які започаткували розвиток комп'ютерної української лексикографії.

Н. Клименко в низці статей 90-их років XX ст. чітко окреслила завдання сучасної морфемології як частини категорійної граматики української мови, що визначають новий предмет дослідження – морфемну структуру слова: «У розділі «Морфеміка» повинні описуватись як елементарні одиниці (морфеми), так і комплексні (морфемні структури слів). [...] Частиною морфеміки є опис таких комплексних її одиниць, як морфемні структури слів. Вони дають уявлення про слово як послідовність морфем, побудовану за певними правилами сполучуваності окремих їхніх типів і класів. Сукупність допустимих морфем, властивих кожній частині мови, показує ланцюжки морфем, за якими моделюються слова у мові, окреслюють можливість афіксального та кореневого зростання слова, спільні та відмінні риси простих і складних слів, співіснування в мові так званих кореневих слів та афіксальних утворень. Морфемні структури слів, прочитувані як ієрархічні утворення з центральним компонентом – коренем і підпорядкованими йому в препозитивній і постпозитивній частині службовими, афіксальними, морфемами, дають змогу показати закономірності конструювання цих структур, використання їх у системі мови і текстах» [Клименко 1996: 199].

Вивчення структурних властивостей об'єкта стало однією з центральних задач багатьох галузей сучасної науки, що перейшли від простого опису безпосередньо спостережуваних фактів до пізнання глибинних властивостей об'єкта й принципів його організації. Проте терміни "структура" і "система" в наукових дослідженнях часто не мають чіткого розмежування або ж виступають як синонімічні. Залежно від того, як визначаються ці поняття, висувається вимога до вибору тих чи інших методологічних принципів, на яких базується конкретне дослідження, і зокрема методу моделювання.

Деякі вчені, диференціюючи ці поняття, протиставляють їх як онтологічну і гносеологічну категорії. А. Мартіне зауважує, що «[...] у мові такої речі як "структура" не існує [...], те, що так називається, не що інше, як схема опису, яка придумана лінгвістом для того, щоб полегшити йому класифікацію матеріалу» [Мартине 1963: 454 – 455].

У дослідженні, яке описується в цій монографії, поняття структури розуміється як онтологічне, що зіставляється з поняттям системи як співвідношення цілого і його частин. Система – цілісний спосіб організації об'єкта. Структура – реляційний каркас системи, мережа відношень між елементами системи. «Під системою розуміється єдине ціле, яке домінує над своїми частинами, що складається з елементів і відношень, які їх пов'язують. Сукупність відношень між елементами системи створює її структуру» [Степанов 1975а: 103].

Структура мови – це об'єктивно існуюча категорія, яка відображає онтологічний статус мови. Проте ця об'єктивна характеристика не може бути виявлена через безпосереднє спостереження. Наявність структури мовної системи можна виявити емпірично за допомогою моделювання реального мовного явища. У такому "виявленому", опосередкованому вигляді, структура як зафіксована мережа відношень може розглядатися (повторно) як теоретична побудова, як структурна модель, як спосіб відображення об'єктивних мовних властивостей.

Структурний підхід визначає спрямованість лінгвістичного аналізу на відношення між об'єктами в межах якоїсь цілісності. У центрі уваги знаходиться не субстанціональна природа елементів, їх фізичних властивостей, а загальний каркас їх відношень. Характеристика елемента встановлюється через його місце у визначеній схемі відношень (структурі), через його зв'язок з іншими елементами й цілим, тобто, через поняття "значущості" елемента в системі, де «кожний залежить від інших і може бути тим, чим він є, тільки завдяки відношенням з іншими елементами» [Ельмслев 1960: 59].

Субстанція мови як вторинної матеріальної системи не є організуючою силою системи. Система мови утворюється не за рахунок фізичних відмінностей системних елементів, а за рахунок семантично значущих фізичних опозицій. Чим вища диференціювальна здатність матеріальної субстанції, тим складнішу семантичну інформацію здатна виражати семіотична система мови. Субстанція системи і елементи системи – це різні поняття. Субстанція не залежить від системи й структури. Елемент системи – структурно зумовлена частина системи, що є формою існування субстанції. Однак це не означає, що структура мовної системи абсолютно не залежна від субстанції, як стверджували глосематики. Структура – це лише один із атрибутів семіотичної системи мови, яка характеризується субстанціональною природою елементів. Структурна характеристика елементів через поняття "значущості" отримується винятково завдяки побудованим у визначеному порядку матеріальним одиницям, але у свою чергу, впливає на ці одиниці, змінюючи їх первинні властивості і функції. Тому структура системи в онтологічному розумінні не може вважатися абсолютно не залежною від субстанції, проте в гносеологічному розумінні незалежність структури від субстанції уможливлює моделювання цієї структури: представлення субстанції елементами інших символьних систем.

Уведення в лінгвістику методологічного правила про можливість окремого вивчення атрибутів мовної системи, що визначає структурну організацію мовних елементів окремими об’єктами вивчення за допомогою структурної методики аналізу, висуває необхідність вирішення низки проблем морфемології на синхронному рівні, зокрема вимагають теоретичного обґрунтування такі проблеми:

а) пояснення процесів організації морфем у слові як одиниці лексичної системи мови та мовлення (тексту);

б) пояснення взаємовідношень інваріанта (морфеми) і варіанта (морфа) як взаємодії парадигматики і синтагматики, що визначає статус морфонологічних явищ і морфонологічного правила;

в) пояснення взаємовідношень між морфемними структурами спільнокореневих та спільноафіксальних слів у лексичній системі мови та мовлення (тексті).

Визначення структури мови як мережі відношень між її елементами дає підстави вважати, що ця мережа створюється окремими системами й підсистемами, а структурні зв’язки елементів, відповідно, відображаються й у зв’язках міжрівневих. Морфема вступає в структурні відношення трьох типів: парадигматичні, синтагматичні, ієрархічні, які можуть бути описані за допомогою методу моделювання.

Лінгвістичне моделювання морфемної структури слова, що відображає як синагматичні, так й ієрархічні відношення морфеми, вимагає представлення чітких дефініцій у трактуванні самого поняття морфеми. Термін "морфема", у трактуванні Бодуена де Куртене, введено в лінгвістику задовго до того, як методологічною основою лінгвістичного дослідження стало фундаментальне для сучасного мовознавства поняття моделі мовної структури.

У більшості лінгвістичних шкіл утвердилось синкретичне за своєю суттю визначення морфеми, що поєднує принципово гетерогенні явища – форму і зміст. Синкретизм у понятті "морфема" – це характеристика морфеми як онтологічної одиниці мови. У науковій метамові, зокрема й у моделюванні, передбачаються суттєво варіативні процедури вичленування, що логічно приводять до побудови різних гносеологічних об’єктів, які можуть претендувати тільки на методологічний статус, статус моделі для відповідного онтологічного поняття морфеми.

Конкретність лінгвістичного знака, яку зазначав Фердинанд де Соссюр, пов’язується насамперед із його формою, що характеризується стабільністю й цілісністю та дозволяє вичленувати цю форму з мовленнєвого потоку. Визначення морфеми як елементарної значущої одиниці мови через цілісність форми спостерігається в лінгвістичних досліджень на початкових етапах розвитку теорії морфеми (Бодуен де Куртене [Б. де Куртене 1963], Празька лінгвістична школа [Вахек 1964], Л. Блумфілд [Блумфилд 1968]), адже поняття форми закладено в самому терміні "морфема" (від грецького morphe – форма).

Визначення морфеми як одиниці опису граматичної будови мови керується ознакою асиметричного дуалізму мовного знака, що не дозволяє встановити відношення між означальним і означуваним морфеми, як відношення один до одного. Тому проблема статусу морфеми як гносеологічної одиниці вирішувалась:

1) шляхом розкладу морфеми на дві одиниці, одна з яких характеризує план змісту, а друга план вираження: морфема = сема + морфема (В. Скалічка [Скаличка 1967а], [Скаличка 1967]); морфема = морфема + морф (Ч. Хоккет [Hockett 1947], З. Харріс [Харрис 1962], Г. Глісон [Глисон 1961]);

2) шляхом переносу акценту з форми на зміст у визначенні морфеми (монеми) – А. Мартіне [Мартине 1963].

В історії розвитку поняття морфеми початково ця одиниця визначалась як елементарний лінгвістичний знак через цілісність форми, а кінцево як одиниця граматичної системи, що виділялась з орієнтацією на функцію. Лінгвісти Празької лінгвістичної школи та американського дескриптивізму, усвідомивши неможливість побудови граматичної системи в термінах лінгвістичного знака, прийшли до його роздвоєння й створення на його основі двох одиниць, а А. Мартіне визначив монему через цілісність функції.

Семіотичний статус морфеми як двобічної одиниці визначався переважною більшістю мовознавців. По-різному трактувалось поняття обсягу морфеми залежно від її відношень із наступною за рангом мовною одиницею – словом: морфема – частина слова, (Бодуен де Куртене [Б. де Куртене 1963], Празький лінгвістичний гурток [Вахек 1964]); морфема безпосередній компонент речення – (Л. Блумфілд [Блумфилд 1968], З. Харріс [Харрис 1965] та інші представники американського дескриптивізму). Полеміка з цього приводу продовжувалася й у лінгвістиці кінця XX ст.. Американський семасіолог Дж. Фодор вважає, що зміст речення – функція морфем, які складають це речення, а семантична відмінність двох речень – це наслідок того, що вони складаються з різних морфем або характеризуються різною локалізацією одних і тих самих морфем [Fodor 1980: 75 – 76].

Розуміння поняття "морфема" як елемента морфологічної системи мови в рамках тієї чи іншої лінгвістичної концепції набуває різної інтерпретації, що зумовлюється прагненням зробити однотипним опис різноманітних засобів вираження граматичних відношень, які дістали назву "grammatical process". Тому в граматичних описах мов з'являються такі морфеми, як пуста морфа (Ч. Хоккет [Hockett 1947]), перервана морфема (З. Харріс [Харрис 1965], А. Мартіне [Мартине 1963]), складна морфема і субморфема (Є. Курилович [Курилович 1965]), нульова морфема (Бодуен де Куртене [Б. де Куртене 1963]), синтаксична аналітична морфема (Л. Теньєр [Теньер 1988]). Поряд із морфемою, що лінійно вичленовується в слові або реченні, статусу морфеми набувають суперсегментні засоби вираження граматичних значень: наголос, інтонація, порядок "семантем", редуплікація, інкорпорація (Ж. Вандрієс [Вандриес 1964], Г. Глісон [Глисон 1961], Ч. Хоккет [Hockett 1947], С. Богданов [Богданов 1980], Ю. Маслов [Маслов 1961]).

У монографії приймається домінуюче в сучасній лінгвістиці теоретичне положення, що «морфемою може вважатися фонологічна послідовність, яка повторюється з певним змістом і локалізована у визначеному місці морфологічних структур, чітко вичленовується, існуючи у вигляді безперервної матеріальної сегментної форми й асоціюється з дискретним, хоч і складним, ”квантом” значення» [Кубрякова 1991: 171]. Необхідно зауважити, що таке визначення, як і представлені попередньо дефініції поняття морфеми, репрезентує синтагматичну, а не парадигматичну одиницю, тобто морф, а не морфему. Морфема – психолінгвістичний інваріант, що характеризується відсутністю конкретно маніфестованої субстанції й актуального референційного значення, одиниця парадигматична, що реалізується в конкретних варіантах – морфах. Морф – онтологічна лінійна субстанціальна одиниця мовлення, яка є конструктом слова й характеризується визначеним значенням. У сучасній морфемології терміни "морфема" і "морф" часто ототожнюються, і в значенні терміна "морф" вживається термін "морфема". Стосовно структурної організації слова на морфемному рівні також типовим є використання терміна "морфемна структура слова", а не "морфна структура слова". Враховуючи традиційність використання терміна "морфемна структура слова", використовуємо його в нашому дослідженні.

Специфіка онтологічної сутності структурної організації мовної системи вимагає обов’язкового використання методики моделювання в лінгвістичному аналізі, предметом дослідження якого виступає морфемна структура слова. Морфемна структура слова як результат сегментації слова на морфеми встановлюється на основі дистрибутивних (синтагматичних), опозиційних (парадигматичних) та ієрархічних відношень морфем у системі мови.

Встановлення опозиційних характеристик морфів базується на понятті значущості лінгвістичних елементів, що притаманне для обох планів мови, як характеристика, що служить для визначення парадигматичного впорядкування елементів. Розуміння мови як системи знаків зводиться до визначення цієї системи сукупністю відомих значущостей, що встановлюються через опозиції. Це призводить до заміни поняття мовної системи поняттям мовної структури. Тому єдиним об’єктом лінгвістики у Фердинанда де Соссюра є мова, яка розглядається сама в собі як ”автономна сутність”. Таке розуміння мовної системи дозволяє застосувати структурні та математичні методи дослідження плану змісту та плану вираження мовного знака як внутрішньолінгвістичних категорій. Хоча мовний знак завжди має екстралінгвістичні семи, описом яких займається зовнішня лінгвістика, це не порушує науковості та експланаторності методів структурної лінгвістики.

Різна інтерпретація мовного знака представляє дві різні семіотичні теорії – менталістичну (де значення наповнюється соціально-культурним змістом, що переносить його у сферу соціальних ситуацій та культурних реалій етносу) й механістичну (визначення знака через форму без врахування

його значення). У першому й другому випадках визначення мовного знака претендує на гносеологічний статус, і тому не заперечує онтологічної суті знака, його двоплановості. Моделювання морфемної структури слова передбачає формалізований аналіз, що ґрунтується на механістичній теорії мовного знака.

Принципи формалізованого опису морфології мови, відповідно до норм дескриптивної лінгвістики, полягають у виділенні мінімальних значущих елементів мовлення (морфемних сегментів), визначенні їх функціонування в різних позиціях, групуванні морфемних сегментів у морфеми, групуванні морфем у граматичні розряди. Проблема значення в трактуванні лінгвістичних одиниць у межах дескриптивної теорії має двояке вирішення. Прихильники формалізованого опису мови вважають, що точне формулювання результатів аналізу, як і весь хід дослідження, повинні враховувати факти значення. Ч. Фріз зауважував, що «[...] йдеться не про опозицію між повною непотрібністю значення і його часткового чи повного застосування, а про те, щоб встановити, наскільки й у якій формі потрібно враховувати значення для адекватного аналізу» [Фриз 1962: 60].

Необхідно визнати, що, за винятком небагатьох робіт теоретичного характеру, значення в тій чи іншій інтерпретації використовується майже всіма дескриптивістами. Але хоча семантичний критерій нерідко вводиться дескриптивістами до лінгвістичного аналізу, використання значення у формальній техніці аналізу залишається невизначеним, як і поняття значення, набуваючи потрактування то предметної віднесеності (referens), то диференційного значення, то відповідей інформанта, то реакції мовця, то кількості інформації, то взаємозамінності. Дескриптивісти обмежують обсяг семантики в лінгвістичних дослідженнях значеннями, вираженими в структурі мови, допускаючи, що відмінності в значенні не суттєві і не повинні братися до уваги до тих пір, допоки цій відмінності не відповідатиме відмінність у формі. Таке розуміння ролі значення в лінгвістичному аналізі доречне, тому що воно вимагає більш точних методів вивчення лінгвістичних значень, що враховують насамперед формальні елементи, завдяки яким значення реалізується й виявляється в мові.

Залежність значення мовного знака від його форми найактуальніше проявляється в морфемі. «Особливе внутрішньоструктурне призначення, а відповідно, і специфічне значення мають словотвірні і словозмінні морфеми, які реалізують своє значення в комбінації (і протиставленні) з іншими знаками, і в силу цього їх називають інколи напівзнаками» [ЛЭС 1990, 154]. Слово і морфема – нерівноцінні знаки у формально-граматичному і семіотичному аспектах. Слово – граматично оформлена одиниця мови, яка характеризується синтаксичною самостійністю й здійснює цілісну референцію до дійсності, виконує номінативну функцію. Морфема – конструктивний елемент слова і, як знак, має значення в системі мови, але не здійснює цілісної референції до дійсності. Це протиставлення визначає різні семіотичні характеристики слова і морфеми: слова називають, виконують

номінативну функцію, морфеми мають асоціативний характер значення, не виконують номінативної функції. У межах знакової теорії перехід від морфеми до слова, як ієрархічний перехід від елементів нижчої за рангом системи до елементів вищої системи, супроводжується набуттям нової якості елементами нижчого рівня – зміною характеру плану змісту: перехід від асоціації понять до номінації реалій дійсності. Ця зміна здійснюється на основі комбінаторики, яка виступає процесом актуалізації значення морфеми, що характеризується категорією інформативності одиниць мови – мірою реалізації змісту морфеми в кожному конкретному вживанні. З іншого боку, комбінаторність морфів у межах слова забезпечує появу нової дискурсивної інформації, крім тієї, що закладена у значенні морфеми. У силу асоціативного характеру значення морфеми, номінативна функція виступає у морфемі лише потенційно можливою реалізацією в структурі слова та речення. Тому в морфемі, як гносеологічному понятті аналізу, посилюється фактор категорії значущості, а не значення.

Будь-який афікс, взятий поза морфемною структурою слова, не характеризується поняттєвим значенням, але, разом із тим, кожний афікс – елемент морфологічної системи мови. Через відношення, у які він вступає в цій системі з іншими афіксами на парадигматичній і синтагматичній площинах, афікс визначається значеннєвим елементом. «Визначальним у семантиці морфеми (морфа) є не значення, а значущість. [....] Значущість морфеми виступає як вияв морфемної парадигматики мови й вираження морфологічної структурно-конструктивної релевантності» [Герд 1983: 51].

Розуміння морфеми значущою одиницею – необхідна умова дистрибутивно-опозиційного моделювання в дослідженні морфемної структури слова. Категорія значущості дійсна й вагома тільки за умови, що предметом дослідження є структура об’єкта – один із атрибутів його системної характеристики. Морфема, як елемент структури морфологічної системи, визначається категорією значущості, цінності, а в мовленні, куди вона входить тільки через посередництво структури слова, морфема набуває того чи іншого значення, хоча онтологічно, у семантиці морфеми значення і значущість є нероздільними аспектами.

Розвиток концепції морфеми в різних лінгвістичних школах свідчить, що морфема як елементарна знакова одиниця не може утримувати свою цілісність у граматичній системі мови. Оскільки принципом побудови мовної системи є мінімальність відмінностей між безпосередньо зіставлюваними одиницями, то системні відношення в плані змісту не збігаються із системними відношеннями в плані вираження. Протиставлення функціональних елементів знака зіставляється з будь-якими структурними відмінностями формальних елементів, і навпаки, опозиції форм відповідають будь-яким відмінностям плану змісту. За принципами глосематики, неможливо одночасно й паралельно описувати систему вираження і систему змісту, одну систему в термінах іншої системи, одні диференційні ознаки в термінах інших диференційних ознак. Усі відношення – парадигматичні,

ієрархічні, синтагматичні, що у своїй сукупності створюють мовну систему, пов'язують між собою не двобічні одиниці мови, а одиниці форми або змісту мовних знаків.

Завдання лінгвістичної моделі морфемної структури слова – визначити форму мовного знака (морфеми), зафіксувати її і встановити функцію морфеми в структурі слова через зв'язок із значенням слова. Проблема виявлення цього зв'язку в протиставленні ідеального і матеріального в мовному знаку. Взаємодія форми і змісту пов'язана з діалектичним законом переходу кількісних змін у якісні. Раптовий характер переходу цих змін не дозволяє зафіксувати складні процеси взаємодії форми і змісту, що відбуваються безперервно. Форма характеризується кількісними показниками, тоді як зміст – якісними. Без посередництва форми людина не здатна сприймати зміст. За визначенням Ю. Степанова, між морфемою і словом лежить "нічия земля", де проходить міжрівневий перехід сукупності морфемних асоціативних значень у цілісну семантику слова. Необхідно знайти внутрішньомовні критерії, які б дозволили тим чи іншим шляхом формалізувати семантичні зв'язки, що лежать в основі тематичного об'єднання морфів у слово. «Відношення між змістовим боком знака, який прихований від безпосереднього спостереження, і тим боком знака, який даний у безпосередньому спостереженні, називається репрезентацією» [Степанов 1975: 239]. Встановлення репрезентації сучасна наукометрія дозволяє перевірити і підтвердити методом моделювання.

Формалізація об'єкта моделювання вимагає представлення співвідношення між планом вираження і планом змісту морфеми через формальні структури, тобто через опозицію значення та вираження морфеми, що й становить його значущість. Вимога формалізації процедури моделювання морфемної структури слова передбачає визначення морфеми гносеологічним поняттям, яке претендує винятково на методологічний статус, статус моделі для відповідного онтологічного поняття морфеми, що, з позицій структуралізму, визначається через поняття значущості, диференційного значення у межах системи.

Формалізований морфемний аналіз у застосуванні методу моделювання має виражено структурний характер, адже дослідження структурних відношень, як правило, пов'язане з пошуком їх формального вираження. Залежно від мети та аспектів опису, цей формалізм може бути віднайдений як на рівні досліджуваних одиниць, так і на інших – вищих та нижчих рівнях мовної системи. Морфемні структури розглядаються як складні об'єкти – окремі підсистеми зі своєю власною структурою елементів – морфів, які, у свою чергу, репрезентовані через кількісно-графемні характеристики. Структурний підхід передбачає визначення зовнішніх і внутрішніх структурних властивостей морфем. Зовнішні структурні властивості морфеми – місце морфа (дистрибуція) в морфемній структурі слова (від перестановки морфа в структурі об'єкта залежить його функціональна

характеристика). Внутрішні структурні властивості морфеми – відношення між графемами / фонемами, які маніфестують конкретний морф.

«Існує два розуміння терміна "формальний". Перше розуміння [...] ототожнює формальний метод із методом дослідження мови в аспекті форми (структури вираження) без врахування семантики (структури змісту). Інше розуміння [на якому ґрунтовано моделювання морфемної структури слова у монографії][1] пов'язано з представленням опису строгим, точним, аксіоматичним способом, який не залишає місця двозначностям. У цьому розумінні можна говорити про формальне представлення елементів структури вираження і елементів структури змісту [...]. Для того, щоб формально представити співвідношення між формою і значенням, [...] необхідно широко користуватись "інтуїтивними" знаннями про мову [...]. У цьому розумінні можна говорити про можливість і необхідність створення формальних описів граматик мов неформальним шляхом» [Откупщикова 1963: 45].

Формалізм структурної морфемології, яка в сучасному мовознавстві дістала назву морфотактика, полягає в роздільному моделюванні плану змісту та плану вираження морфемних структур слів, що ґрунтовано на концепції роздільного членування тексту Л. Єльмслєва: «...будь-який текст на першому етапі поділяється завжди на дві, і тільки на дві частини, мінімальне число яких гарантує їх максимальну протяжність, а саме – на лінію вираження і лінію змісту, що пов'язані між собою солідарністю через посередництво знакової функції. Потім лінія вираження і лінія змісту, кожна у свою чергу, поділяється далі з врахуванням їх взаємодії в знаках. У такий спосіб перше членування лінгвістичної системи приведе нас до встановлення двох її парадигм: вираження і змісту» [Ельмслев 1960: 317].

1.2. Комп'ютерні моделі об'єктів та процесів морфемної системи мови

Терміни "модель" і "метод моделювання" запозичено лінгвістикою із математики американськими дескриптивістами З. Харрісом та Ч. Хокеттом. Теоретичним підґрунтям створення різних типів морфемних моделей у сучасній комп'ютерній морфемології виступають постулати структурної лінгвістики, у якій побутує таке визначення моделі: «Модель у структурній лінгвістиці – це символічний апарат, або метамова певної теорії, яка виступає як семіотичний аналог структури, що закладена об'єктивно в природі мови – [...] всяка форма запису є частина змісту запису» [Степанов 1975: 130]. Структура мови, представлена через модель, виступає як "автономна сутність", гносеологічний конструкт. На відміну від математики, де термін модель має значення "інтерпретація теорії", в емпіричних науках цей термін має інше трактування. У математичних науках «[...] теорія – це оригінал для

[1] Додаткова інформація до цитати вставлена автором моногафії.

моделі, тобто об’єкт, відображенням якого служить модель, що використовується як знаряддя дослідження цього об’єкта. В емпіричних науках теорія і модель повинні бути підведені під одне поняття як гіпотетико-дедуктивні системи з еквівалентною пізнавальною функцією» [Шаумян 1965: 28]. У математиці теорія – дедуктивна система, яка відмежована від предметної дійсності, а місце гіпотез у ній займають аксіоми. Теорія в структурній лінгвістиці – гіпотетико-дедуктивна система, в якій висновки – наслідок поєднання даних про спостережувані факти із сукупністю фундаментальних гіпотез. «Модель – це теорія, яка має наочний зміст у вигляді образів, що служать аналогами об’єктів, які приховані від прямого спостереження» [Шаумян 1965: 28].

З розвитком комп’ютерної лінгвістики, яка ґрунтована на теоретичних засадах структуралізму, виникає новий тип моделей – комп’ютерні лінгвістичні моделі. «Важливе лінгвістичне завдання у сфері проблематики комп’ютерної лінгвістики – це побудова таких лінгвістичних процесорів, які здатні були б забезпечити людино-машинне спілкування природною мовою. Для успішного вирішення цієї стратегічної мети необхідно створити теоретичні передумови, а саме – розробити адекватну модель взаємодії мовних одиниць і реалізувати її в комп’ютерній граматиці і комп’ютерних словниках. Основним поняттям у сучасній комп’ютерній лінгвістиці є поняття моделі, а метод моделювання є конструктивною необхідністю, оскільки об’єкт науки в спілкуванні людина-машина-людина недоступний для безпосереднього спостереження» [Дарчук 2013: 27].

Комп’ютерне моделювання як лінгвістичний метод ставить завдання розв’язувати задачі лінгвістичного аналізу або синтезу на основі використання комп’ютерної моделі структури мови, що виступає обов’язковим складником створюваної інформаційної системи. Зазвичай, за архітектурою ANSI/X3/SPARK, «...в інформаційній системі виділяють три рівні опису даних: концептуальний, внутрішній та зовнішній. Зовнішні моделі відтворюють погляди кінцевих користувачів [...] на інформаційну систему. У внутрішньому рівні визначаються структури зберігання й маніпулювання даними та адекватне алгоритмічне й програмно-операційне середовище. [...] Визначення концептуальної моделі передбачає задання набору категорій (об’єктів) моделі, набору операцій над визначеними об’єктами, множини обмежень цілісності, які відображають семантику предметної галузі» [Широков 1998: 81].

Відомі представники комп’ютерної лінгвістики, такі як В. Перебийніс [Перебийніс 1969], Р. Піотровський [Пиотровский 1999], Ю. Апресян [*Апресян* 1990], С. Шаумян [Шаумян 1965], Н. Дарчук [Дарчук 2013], Ю. Караулов [Караулов 1981], А. Баранов [Баранов 2001], В. Широков [Широков 1998] розробили низку вимог, які ставляться до концептуальних інформаційних моделей:

- модель повинна бути спрощеним аналогом, а не копією оригіналу, тобто простішою за оригінал;

- моделювати можна тільки ті явища, істотні властивості яких вичерпуються їхніми структурними (функціональними) характеристиками;
- модель повинна бути формальною, точною й однозначною;
- модель оперує не поняттями про об'єкти природної мови, а конструктами, тобто поняттями про ідеальні об'єкти, одержаними на підставі загальних гіпотез на основі спостережень та інтуїції дослідника, тому модель завжди є ідеалізацією об'єкта;
- модель повинна застосовуватись не до якогось конкретного об'єкта, а до певного класу об'єктів природної мови;
- модель пов'язується з експериментальними даними за допомогою тієї чи іншої інтерпретації: визначення правил підстановки реальних об'єктів мови замість об'єктів моделі;
- модель повинна мати: пояснювальну силу – здатність передбачувати, виявляти й пояснювати властивості природної мови та евристичні (пошукові) властивості – генерувати нові знання про оригінал;
- всі об'єкти моделі, а також операції над ними повинні передбачати інтерпретацію у вигляді алгоритмів скінченної складності;
- модель завжди спрямована на зв'язок із лінгвістичним процесором, тобто програмно-апаратним комплексом, який є інтегрованим середовищем, призначеним для вирішення лінгвістичних задач за допомогою комп'ютера.

Традиційним у лінгвістиці є визначення моделі в Лінгвістичному енциклопедичному словнику: «Модель (франц. modèle, від лат. modulus – міра) в лінгвістиці: 1. Штучно створений лінгвістом реальний або уявний пристрій, що відтворює, імітує своєю поведінкою (зазвичай у спрощеному вигляді) поведінку якогось іншого ("справжнього") пристрою (оригіналу) в лінгвістичних цілях. 2. Зразок, що слугує стандартом (еталоном) для масового відтворення; те саме, що й "тип", "схема", "парадигма", "структура" тощо (наприклад, "модель дієвідміни або відміни", "словотвірна модель", "модель речення" тощо)» [ЛЭС 1990: 304]. Наведена дефініція репрезентує неоднозначне розуміння терміна "модель": перше значення – модель як схема процесу; друге – модель як зразок статичного об'єкта.

Хрестоматійною в структурній лінгвістиці є стаття: Чжао-Юань-Жень «Моделі в лінгвістиці і моделі загалом» [ЧЮЖ 1965], у якій наведено до тридцяти визначень терміна "модель". Варіативність поняття лінгвістичної моделі пояснюється відсутністю єдиної типології моделей. Узагальнено можна визначити три основні аспекти, що визначають тип лінгвістичної моделі: 1) об'єкт моделювання; 2) мета і спосіб побудови теорії лінгвістичного дослідження; 3) використання / невикористання комп'ютера в лінгвістичному дослідженні.

Об'єкти моделювання визначаються структурною організацією мовної системи, для того, щоб їх виявити необхідно проводити дослідження на кожному рівні мовної системи. Формалізований опис одиниць кожного рівня в комп'ютерному моделюванні ускладнюється виявленням правил

сполучуваності цих одиниць з урахуванням подальшого програмного забезпечення цих правил. Об’єктами моделювання в комп’ютерній морфемології можуть виступати:

1) структурні одиниці, що відображають структурні властивості (синтагматичні, парадигматичні, ієрархічні) морфеми: морфемні структури слів; фонологічні / графемні структури морфем; морфеми як парадигматичні класи аломорфів; морфеми як інваріантні морфонологічні конструкти;

2) мовленнєва діяльність, спрямована на морфемний синтез і аналіз словоформ;

3) текст на морфемному рівні організації;

4) морфеміка як цілісна підсистема мови.

Перший тип об’єктів визначає поняття "структурна модель". Такі моделі називаються статичними: «статична модель (= структурна, класифікаторна, таксономічна модель) – модель, яка відображає будову, устрій оригінала-мовного об’єкта чи об’єктів у статиці, або стані спокою, класифікує його за певними ознаками» [Карпіловська 2006: 33].

Моделювання мовленнєвої діяльності в традиційній і в комп’ютерній лінгвістиці передбачає створення процесуальної, динамічної моделі, спрямованої на визначення морфем у структурі слова / словоформи, чи, навпаки, створення слів / словоформ за заданою системою морфемних одиниць, а також загалом на породження / синтез нових слів-об’єктів за заданими вихідними елементами (морфемами, основами, словотвірними формантами тощо) та правилами словотвірного синтезу. «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 упорядкованої сукупності, то динамічні моделі унаявнюють процеси виведення одних мовних об’єктів з інших, їхні взаємоперетворення» [Карпіловська 2006: 33].

Моделювання морфемної системи мови та тексту передбачає створення лексикографічної моделі, орієнтованої на системний опис великого реєстру морфемних одиниць, які належать до першого типу об’єктів. Лексикографічні моделі традиційних паперових словників мають статичний характер, лексикографічні моделі електронних інтерактивних словників є динамічними (процесуальними).

За метою і способом побудови теорії лінгвістичного дослідження моделі поділяються на три типи: індуктивні, дедуктивні та індуктивно-дедуктивні моделі. Індуктивні моделі організовують дослідження за схемою: від конкретного мовного матеріалу до формування певної гіпотези про закономірності організації та функціонування лінгвістичних явищ. Дедуктивні моделі будуються на основі висунутої гіпотези, яка перевіряється на фактичному мовному матеріалі. Індуктивно-дедуктивні моделі, будуються на основі синтезу даних про спостережувані мовні явища й сукупності гіпотез про "поведінку" та структуру мовних об’єктів. У лінгвістичних дослідження,

які не використовують інструментів комп'ютерної лінгвістики, усі статичні моделі є індуктивними, а динамічні моделі синтезу та трансформаційні моделі завжди є дедуктивними. Комп'ютерні статичні і динамічні моделі завжди є індуктивно-дедуктивними. У комп'ютерній моделі індуктивний і дедуктивний підходи не протиставляються, а взаємодоповнюються: індуктивні узагальнення, зроблені на концептуальному рівні моделі, ізоморфно формалізуються на етапі софтвера[6] й багаторазово перевіряються на новому мовному матеріалі. У результаті перевірки продукуються нові знання про модельовані на концептуальному рівні об'єкти й виводять нові теоретичні положення з інших, які вже були відомі.

Застосування комп'ютерного моделювання лінгвістичних явищ пов'язано із базовими поняттями софтвера: "лінгвістичний алгоритм", "лінгвістична база даних" та "лінгвістичний процесор". Лінгвістичний процесор працює на основі лінгвістичних баз даних, що систематизують формалізовану інформацію про мовні об'єкти, представлені у вигляді статичних лінгвістичних моделей, та програмного забезпечення, розробленого за схемами динамічних моделей лінгвістичних правил – лінгвістичними алгоритмами.

У лінгвістичних процесорах бази даних називають ще резидентними (прихованими) словниками, або автоматичними словниками. Таким чином, робота сучасних лінгвістичних процесорів прямо пов'язана із комп'ютерною лексикографією: автоматичний словник (база даних) лежить в основі роботи лінгвістичного процесора, а електронний словник є зовнішньою комп'ютерною моделлю – результатом роботи лінгвістичного процесора. Відмінність між лексикографічними моделями двох типів комп'ютерних словників у тому, що автоматичний словник систематизує формалізований опис лінгвістичних об'єктів у вигляді моделей, які можуть бути не зрозумілі користувачам, але "зрозумілі" лінгвістичному процесору, натомість, вихідний електронний словник систематизує лінгвістичну інформацію, представлену в зрозумілому для користувача вигляді.

Розвиток комп'ютерної лінгвістики уможливив проведення лінгвістичного дослідження на великих обсягах мовного матеріалу як текстового, так і лексикографічного. Результати автоматичного лінгвістичного аналізу у таких дослідженнях систематизуються у великі бази даних, які можна вважати проявом лексикографічного ефекту методології сучасної лінгвістики.

Поняття "лексикографічний ефект" було введено відомим українським науковцем В. Широковим [Широков 1998] у межах інформаційної теорії лексикографічних систем (Л-систем). Лексикографічний ефект розглядається як тенденційне явище в інформаційних системах. Будь-яка інформаційна система характеризується наявністю інформаційних одиниць та відношень

[6] Софтвер – (від англ. software: soft – м'який; ware – продукт) комп'ютерна програма, яка реалізовує алгоритм дій оператора ЕВМ, на противагу хардверу – інструментальній технічній частині комп'ютера.

між цими одиницями подібно до лексикографічної системи. Обмін інформацією й перетворення комбінацій інформаційних одиниць (феноменів) з однієї форми в іншу зумовлює породження нових лексикографічних систем, причому сприйняття та аналіз людиною будь-якої інформаційної (лексикографічної) системи має суб'єктивну інтерпретацію.

Природну мову як інформаційну систему особливого типу формують мовні одиниці, що характеризуються дискретністю. Дискретність визначає властивості мовних одиниць різних рівнів мовної системи, і ці одиниці є проявами лексикографічних ефектів кожного мовного рівня: система фонем – лексикографічний ефект фонетики; система морфем – лексикографічний ефект морфеміки. У цьому виявляється феноменологічний аспект лексикографічного ефекту мовної системи. «Універсальність явища лексикографічного ефекту спричиняє неодноразово відзначену нами тенденцію до лексикографування будь-якого мовного феномену – саме цей факт пояснює побутування в лексикографічній практиці прикладів створення словників, у яких лексикографуються навіть і такі одиниці мови, які не мають безпосереднього вербального вираження» [Широков 2005: 41].

Комп'ютерна лінгвістика, вивчаючи мову як інформаційну систему, а мовні одиниці і явища як феномени інформаційної системи методом комп'ютерного моделювання, продукує нове розуміння лексикографічного ефекту: лексикографічний ефект методології лінгвістичного дослідження. Є. Купріянов, акцентуючи увагу на методологічному аспекті концепції лексикографічного ефекту, зауважив: «Можна стверджувати, що, досліджуючи будь-які предметні галузі, фахівці вивчають лексикографічні ефекти, які виникають у цих галузях. Таким чином, лексикографічний ефект має не лише феноменологічний складник, а й методологічний, оскільки володіє певним «потенціалом операціональності», для того щоб визначати відповідні комплекси елементарних інформаційних одиниць у процесі моделювання тих чи інших систем, при цьому враховуючи, конкретизуючи та репрезентуючи властивості цих одиниць. У цьому прояві концепція лексикографічного ефекту є методом абстрагування даних» [Купріянов 2018: 64].

Комп'ютерне моделювання морфемної системи мови ґрунтується на феноменології лексикографічного ефекту цієї системи, що мотивує, визначає концептуальну лексикографічну модель як базовий конструкт методу, і проявляється, як результат застосування моделі, у створенні лексикографічної внутрішньої моделі, що в сучасній комп'ютерній лінгвістиці називається лінгвістичною базою даних. Враховуючи, що комп'ютерна модель повинна застосовуватись не до якогось конкретного об'єкта, а до певного класу об'єктів природної мови, тобто якоїсь системи, можна стверджувати, що в лінгвістичних процесорах завжди використовується якась внутрішня лексикографічна модель. У такому розумінні поняття лексикографічна модель та лексикографічна система використовуються ширше, ніж у лексикографії, яка ставить завдання укласти

для користувача паперовий або електронний словник, який виступає зовнішньою моделлю. Таким чином, комп'ютерна лексикографічна модель, спрямована на укладання електронного словника для людини-користувача, має дворівневий принцип будови: 1) комп'ютерна внутрішня лексикографічна модель бази даних, що є обов'язковим атрибутом лексикографічного процесора; 2) комп'ютерна зовнішня лексикографічна модель електронного словника, представлена для користувача, – результат роботи лексикографічного процесора.

Створення лінгвістичної бази даних також ґрунтовано на двох рівнях опису лінгвістичної інформації й передбачає поєднання двох типів моделей, або двох етапів моделювання: інфологічний, на якому створюється концептуальна модель; та датологічний (софтвер), на якому створюється інформаційна (алгоритмічно-програмна) модель. За визначенням Є. Карпіловської: «Завдання інфологічного етапу проектування бази даних полягає у відборі об'єктів опису, типів інформації про їхню будову та функціонування. 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, даталогічного, 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» [Карпіловська 2006: 35].

Комп'ютерне моделювання тексту на морфемному рівні організації або морфеміки як цілісної підсистеми мови має лексикографічне спрямування й передбачає укладання морфемної бази даних та представлення її для користувача у вигляді паперового або електронного морфемного словника. У досягненні цієї мети дотримується послідовність моделювання різних об'єктів і процесів (правил) морфемної системи мови.

Створення комп'ютерного лексикографічного опису морфемної системи мови на базі встановленого у словниках реєстру слів із визначеною морфемною будовою проходить у три етапи:

(1) моделювання морфемної структури слова;

(2) моделювання морфемної бази даних;

(3 а) моделювання структури електронного словника.

У створенні лексикографічного опису морфемної системи тексту завдання ускладнюється, тому що об'єктом аналізу стають слововживання тексту, які ще не поділені на морфеми. Текстоорієнтований комп'ютерний лексикографічний морфемний опис передбачає таку послідовність етапів моделювання:

(1) моделювання морфемної структури слова;

(2) моделювання морфемної бази даних;

(3б) моделювання процесу морфемної сегментації слова (леми або слововживання);

(4) моделювання лексикографічної бази даних морфемного словника за лексичним реєстром текстової вибірки;

(5) моделювання електронного морфемного частотного словника, у якому морфемні одиниці мають зв'язок із текстовим конкордансом.

(1) Модель морфемної структури слова. В обох типах лексикографічного комп'ютерного моделювання (словниковоорієнтованого та текстоорієнтованого) на першому етапі створюється модель морфемної структури слова, яка описує цю структуру в символах, доступних для автоматичних операцій.

Моделювання морфемної структури слова в українській морфемології вперше було використано Н. Клименко та Є. Карпіловською: «Встановлення інвентаря морфемних структур слова слід проводити на статистично впорядкованому мовному матеріалі. Коректна організація його визначає надійність одержаних висновків. Одним із шляхів розв'язання цього питання може стати перезапис одиниці словника української мови, в якому подаються поділень на морфеми слова, у вигляді формули морфемної структури слова. При перезаписі кожна морфема слова повинна бути переведена в одиницю певного класу. Таких класів передбачається п'ять: корені, префікси, суфікси, з'єднувальні голосні в складному слові, флексії. Позначимо їх, відповідно, символами: K, P, S, I, F» [Клименко 2014а]. Модель морфемної структури слова стала основним методологічним конструктом у вивченні законів морфемної та словотвірної будови українського слова у фундаментальних дослідженнях [7] відділу структурно-математичної лінгвістики Інституту мовознавства НАН України та у створенні комп'ютерного морфемно-словотвірного фонду української мови.

Модель у формі символьного перекодування функціональних типів морфем є типовою моделлю в галузі морфотактики, хоча в різних дослідженнях символи можуть змінюватися. У комп'ютерному моделюванні при побудові АСМСА ([Алексієнко 2001], [Алексеенко 2004], [Zuban 2015]) було використано такі моделі: P – префікс, R – корінь, S – суфікс, F – флексія, I – інтерфікс, X – постфікс, наприклад *уч/и/тель/к/а* – RSSSF.

За допомогою методу моделювання морфемної структури слова створюється формальний лінгвістичний опис цієї структури, який представляє структуру форми і структуру змісту морфемної структури солова у вигляді формальної функціональної формули будови морфемної структури. В основу такого опису покладено фундаментальну ідею, розроблену відомими лінгвістами В. Солнцевим [Солнцев 1971], Ю. Степановим [Степанов 1975]: формула будови, що регулярно

[7] Наукові досягнення цього колективу системно описані у статті Є. Карпіловської «Здобутки академічної структурної та математичної лінгвістики у моделюванні українського слова» [Карпіловська 2019].

повторювана і відтворювана в мові, визначається як окрема одиниця мовної структури поряд із фонемою та морфемою. Формули будови – інваріант класу морфемних структур конкретних слів: одна модель морфемної структури слова може представляти великий клас конкретних морфемних структур слів, що онтологічно існують у субстанціальних одиницях (словах), які виступають варіантами формул будови.

Модель морфемної структури слова є індуктивною моделлю, що формалізує дистрибутивно-опозиційні відношення функціональних класів морфем у морфемній структурі слова та загалом у морфемній системі мови. Систематизований опис лексики мови в символах моделей морфемної структури слова створює концептуальну модель морфемної підсистеми мови або тексту.

(2) Моделювання морфемної бази даних. У галузі української морфемної та словотвірної лексикографії з 1988 р. успішно розвивається практика і теорія укладання баз даних із використанням методу комп'ютерного моделювання: 1988 – 2020 рр. – морфемно-словотвірний фонд української мови [Клименко 2014б]; 2001 – 2020 рр. – автоматизована система морфемно-словотвірного аналізу (АСМСА) [Алексієнко 2001]. Ці дві морфемно-словотвірні бази побудовані на основі однієї концептуальної моделі: морфемна структура слова описується за допомогою символів статичної моделі, про яку йшлося вище, проте на даталогічному етапі створення бази даних використовуються різні методики у створенні комп'ютерної моделі, що зумовило різну структуру баз даних і різні завдання: морфемно-словотвірний фонд української мови [8] – словниковозорієнтована БД, а АСМСА – орієнтована на створення електронного словника за лексичним реєстром МБД АСМСА [9] та на проведення автоматичного морфемного аналізу слів за текстовими вибірками в Корпусі української мови й автоматичне укладання електронного морфемного словника за текстовими вибірками.

(3б) Моделювання процесу морфемної сегментації слова. У процесі роботи над створенням лінгвістичного процесора, здатного здійснювати автоматичний морфемний аналіз словоформ / лем українськомовного тексту та автоматично укладати частотні морфемні словники, було розроблено два типи динамічних моделей морфемної сегментації слів української мови [10], що використовують різні даталогічні моделі морфемної структури слова: 1) модель автоматичного морфемного

[8] Принципи даталогічної моделі морфемної структури слова, а також структура й функції бази даних МСФ української мови описані в публікаціях співробітників відділу структурно-математичної лінгвістики Інституту мовознавства ім. О. О. Потебні АН України [Клименко 2014б], [Карпіловська 2006].

[9] Інфологічний та даталогічний етапи створення МБД АСМСА будуть описані в наступних параграфах монографії.

[10] Опис цих моделей, а також етапів комп'ютерного моделювання (3а), (4), (5) представлено в наступних параграфах першого розділу та в другому розділі монографії.

сегментування суфіксальної зони дієслівних словоформ тексту на основі квазіфлексій (фізична стадія даталогічного етапу не реалізована) [Зубань 1998]; 2) модель автоматичного морфемного сегментування лем текстових слововживань на основі функціонально-кількісно-графемних моделей слів [Zuban 2015], [Zuban 2017].

1.3. Моделювання процедури автоматичного морфемного аналізу посткореневої зони українського дієслова

1.3.1. Алгоритмічна модель морфемного сегментатора посткореневої зони дієслівних словоформ української мови

На першому етапі розвитку комп'ютерної морфемної лексикографії комп'ютер використовувався для укладання морфемних традиційних (паперових) словників. За допомогою комп'ютера були укладені морфемні словники російської мови: частотний словник З. Оліверіуса «Морфемы русского языка» [Оливериус 1976], кореневий словник Д. Уорта, О. Козака, Д. Джонсона «Русский словообразовательный словарь» [Уорт 1983]; морфемний словник чеської мови Е.Славічкової [Slavíčková 1975]; а також відомі українські словники нового покоління: «Словник афіксальних морфем української мови» [САМУК 1998], «Кореневий гніздовий словник української мови» [Карпіловська 2002] та «Шкільний словотвірний словник сучасної української мови» [ШСССУМ 2005]. При укладанні цих словників уперше було застосовано автоматизований морфемний аналіз.

При укладанні морфемних словників Е. Славічкової, З. Оліверіуса та зазначених словників української мови комп'ютер було використано для різноманітних класифікаційних та ідентифікаційних операцій, а також кількісних обчислень на основі баз даних, які систематизують морфемні структури слів, що були розділені вручну на морфеми, яким приписані різні характеристики-індекси даталогічної моделі. Укладання цих словників забезпечують лексикографічні процесори, оснащені програмами пошуку й підрахунку заданих у БД дискретних одиниць – морфем.

«Русский словообразовательный словарь» Д. Уорта, О. Козака, Д. Джонсона [Уорт 1983] ілюструє досвід автоматичного членування слова на морфеми й групування слів за спільними коренями. Базовою ідеєю автоматичного морфемного сегментування в цьому словнику є дистрибутивний метод: на морфемних швах можливі тільки деякі визначені послідовності графем, завдяки яким можуть бути виявлені морфеми.

Морфемні аналізатори відомих словників, укладених за допомогою комп'ютера, розраховані на аналіз слова, що розглядається поза текстом. У межах системи автоматичного граматичного аналізу тексту (АГАТ), розробленого на початку 90-років групою лінгвістів Інституту мовознавства ім. О.О.Потебні НАНУ (Т. Грязнухіна, Н. Дарчук, В. Критська, Л. Орлова, Т. Пуздирєва) й програмістів (Г. Колєнов, В. Сорокін, Т. Недозим), було

поставлено завдання створити автоматичну систему морфемного сегментування словоформ тексту. Ця система була розроблена для дієслівних словоформ на алгоритмічному рівні (без реалізації програмного забезпечення) в дисертаційному дослідженні О. Зубань [Зубань 1998]. Базисними принципами автоматичного морфемного сегментування словоформ стали теоретичні аспекти побудови багатоступеневого алгоритму морфологічного аналізу Г. Бєлоногова [Белоногов 1983], методологія дистрибутивного аналізу З. Харріса [Харрис 1962], [Харрис 1965], що була використана в досвіді автоматичного членування слова на морфеми Д. Уортом [Уорт 1983], М. Андреєвим [Андреев 1963], П. Шевельовою [Шевелёва 1973], Б. Сухотіним [Сухотин 1984], та в методиці автоматичного сегментування слова, представленої «Структурною граматикою сучасної української літературної мови (проспект)» [СГСУЛМ].

Морфемний сегментатор (МС) – один із блоків морфологічного модуля системи АГАТ, що забезпечує автоматизацію морфемного аналізу словоформ. На вході МС – текст, у якому на першому етапі роботи морфологічного модуля словоформам приписані коди граматичного класу і підкласу (додаток 2). На виході – морфемне сегментування словоформ із збереженою інформацією про граматичне значення.

В основі концептуальної інформаційної моделі автоматичного морфемного сегментування текстових слововживань лежить частиномовна й функціональна характеристика морфемної структури словоформ української мови: посткоренева зона диференційована за частинами мови, а префіксальна зона спільна для всіх частин мови Ця інфологічна модель може бути представлена у вигляді такої блок-схеми:

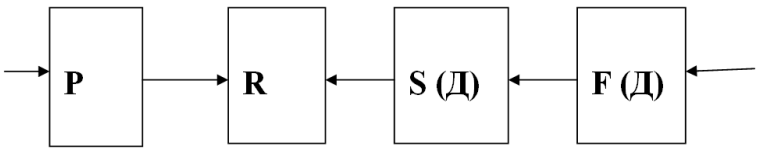


Рис.1.1. Блок-схема інфологічної моделі морфемної сегментації текстових слововживань

Чотири блоки моделі МС відповідають функціональному навантаженню чотирьох морфемних зон:

1) флективна зона (блок F);
2) суфіксальна зона (блок S)
3) коренева зона (блок R)
4) префіксальна зона (блок P).

Блок-схема (рис. 1.1) демонструє, що роботу МС посткореневої зони українського дієслова забезпечують два блоки: блок F (Д), блок S (Д). Коренева зона словоформ виділяється як частина, що залишилася після сегментації флексії, суфіксальної та префіксальної послідовностей.

МС аналізує морфемну структуру словоформи в інверсійному порядку, тому на першому етапі роботи морфемного сегментатора працює блок F.

Вхідною інформацією для зони флексії є текст із приписаними словоформам граматичними кодами. Вихідна інформація для блоку F (Д) – дієслова з визначеною межею флексії й збереженою інформацією граматичного коду про частиномовну характеристику та граматичні дієслівні значення, упорядковані в алфавітному порядку. Ця інформація є вхідною для другого етапу роботи МС – блоку S(Д). На виході блоку S – дієслівні словоформи, у яких сегментовано посткореневу морфемну структуру, із збереженою інформацією про граматичне значення.

Створення інфологічної та датологічної моделі МС вимагає формалізованого опису морфемної структури посткореневої зони українського дієслова. Формалізація опису проводиться на основі даних: 1) про дистрибутивно-опозиційні характеристики графем на морфемному шві між коренем та суфіксальною послідовністю основи; 2) про комбінаторику словозмінних суфіксів та флексій у дієслівних словоформах. Формалізований опис передбачає проведення дослідження в декілька етапів та розв'язання таких задач:

1) встановлення закономірностей сполучуваності дієслівних основ із словозмінними суфіксами та флексіями;

2) побудова алгоритму автоматичного виділення флексій у дієслівних словоформах української мови;

3) побудова алгоритму автоматичного виділення словозмінних суфіксів у дієслівних словоформах української мови;

4) встановлення дистрибутивної діагностики графем на морфемних швах у суфіксальних послідовностях дієслівних основ;

5) побудова алгоритмічного графа (дерева) суфіксальної зони дієслівної основи інфінітива;

6) представлення графа основи інфінітива в термінах квазіфлексій та програмних процедур;

7) побудова алгоритмічного дерева суфіксальної зони дієслівної основи теперішнього часу;

8) представлення графа основи теперішнього часу в термінах квазіфлексій та програмних процедур.

Роботу блока F(Д) МС моделює алгоритм виділення флексій у дієслівних словоформах української мови. Роботу блока S(Д) – алгоритм виділення суфіксів у дієслівних словоформах української мови.

Створенню алгоритмів автоматичного виділення флексій і словозмінних суфіксів дієслівних словоформ передувало встановлення інвентаря суфіксів і флексій, який добирався за даними «Украинской грамматики» [УГ 1986: 161 – 172]. Для побудови алгоритмів необхідно було змоделювати графемну структуру морфемних словозмінних блоків, а також одиничних словозмінних суфіксів та флексій у вигляді правил-процедур (формул), у яких знаком (*) позначається кожна графема морфеми, а знаком (-) – межа між двома морфемами.

Флексії особово-часової парадигми теперішнього / майбутнього (доконаного виду) часу першої та другої дієвідміни [11] змодельовані в табл. 1.1. та табл. 1.2:

Таблиця 1.1. Флективні моделі першої дієвідміни

№	особово-числове значення	приклад словоформи	флексія	кількісно-графемна модель
1	2 особа однини	читає-ш	-ш	-*
2	1особа множини	читає-мо	-мо	-**
3	2 особа множини	читає-те	-те	-**
4	3 особа множини	читаю-ть	-ть	-**

Таблиця 1.2. Флективні моделі другої дієвідміни

№	особово-числове значення	приклад словоформи	флексія	кількісно-графемна модель
1	1особа однини	сидж-у	-у	-*
2	2 особа однини	сиди-ш	-ш	-*
3	3 особа однини	сиди-ть	-ть	-**
4	1 особа множини	сиди-мо	-мо	-**
5	2 особа множини	сиди-те	-те	-**
6	3 особа множини	сидя-ть	-ть	-**

Словозмінні морфемні комплекси в дієсловах майбутнього часу недоконаного виду можуть описані формулами, поданими в табл. 1.3.

Таблиця 1.3. Формотвірні моделі дієслівної парадигми майб. часу недок. виду

№	особово-числове значення	приклад словоформи	суфікс інфінітива	словозмін. суфікс	флексія	кількісно-графемна модель
1	1 ос. од.	чита-тиму	-ти-	-му-		-**-**
2	2 ос. од.	чита-тимеш	-ти-	-ме-	-ш	-**-**-*
3	3 ос. од.	чита-тиме	-ти-	-ме-		-**-**
4	1 ос. мн.	чита-тимемо чита-тимем	-ти- -ти-	-ме- -ме-	-мо- -м-	-**-**-** -**-**-*
5	2 ос. мн.	чита-тимете	-ти-	-ме-	-те-	-**-**-**
6	3 ос. мн	чита-тимуть	-ти-	-му-	-ть-	-**-**-**

Словозмінні суфікси та флексії родово-числової парадигми минулого часу представлено в алгоритмі у вигляді моделей, поданих у табл. 1.4. Флексії парадигми наказового способу моделюються в табл. 1.5.

[11] Словоформи 1ос.од. та 3 ос.од. першої дієвідміни теперішнього/майбутнього часу характеризуються відсутністю особової флексії, яка накладається на препозитивний дієслівний суфікс *(чит-аю, чит-ає)*, тому ці словоформи не потрапляють до блоку F морфемного сегментатора.

Таблиця 1.4. Формотвірні моделі дієслівної парадигми минулого часу

№	родово-числове значення	приклад словоформи	словозмін. суфікс	флексія	кількісно-графемна модель
1	чол.р.од.	чита-в	-в	0[12]	-*
2	жін.р.од.	чита-ла	-л-	-а	-*-*
3	сер.р.од.	чита-ло	-л-	-о	-*-*
4	множ.	чита-ли	-л-	-и	-*-*

Таблиця 1.5. Флективні моделі наказового способу

№	особово-числове значення	приклад словоформи	флексія	кількісно-графемна модель
1	2 особа однини	пиш-и	-и	-*
2	1 особа множини	пиш-імо сядь-мо	-імо -мо	-*** -**
3	2 особа множини	пиш-іть сядь-те	-іть -те	-*** -**

Формалізований опис флексій у символах кількісно-графемних моделей дозволяє змоделювати морфемну сегментацію флексій у дієслівних словоформах у вигляді алгоритмічної динамічної моделі (рис. 1.2).

Опис алгоритму:

1) початок.

2) введення опису процедур:

PROC 51 /=**/ (комп'ютер відділяє постфікс -ся й ставить перед ним позначку /=/);

PROC 52 /-***/ (комп'ютер відділяє триграфемні флексії наказового способу -імо, -іть);

PROC 53/-**/ (комп'ютер відділяє двографемні флексії теперішнього і майбутнього часу -мо, -те, -ть ,та флексію інфінітива);

PROC 54 /-*/ (комп'ютер відділяє одноградемні флексії: теперішнього часу -у, -ш, -м; майбутнього часу -ш; минулого часу -а, -о, -и; другої особи однини наказового способу -и ;

PROC 55 /-0/ (комп'ютер приписує позначку нульової флексії в словоформах чоловічого роду минулого часу та другій особі однини наказового способу);

PROC 56 /блок S/ (перехід у режим роботи блоку S МС).

3) чи закінчує слово постфікс -ся/сь? Якщо так – п. 4, якщо ні – п.5.

4) PROC 51. п.5.

[12] Нульова флексія не моделюється в моделях-процедурах, тому що у графемному представленні словоформ тексту вона відсутня. Як необхідний елемент граматичної будови словоформи, нульова флексія приписується до відповідних словоформ у режимі роботи блоку F МС.

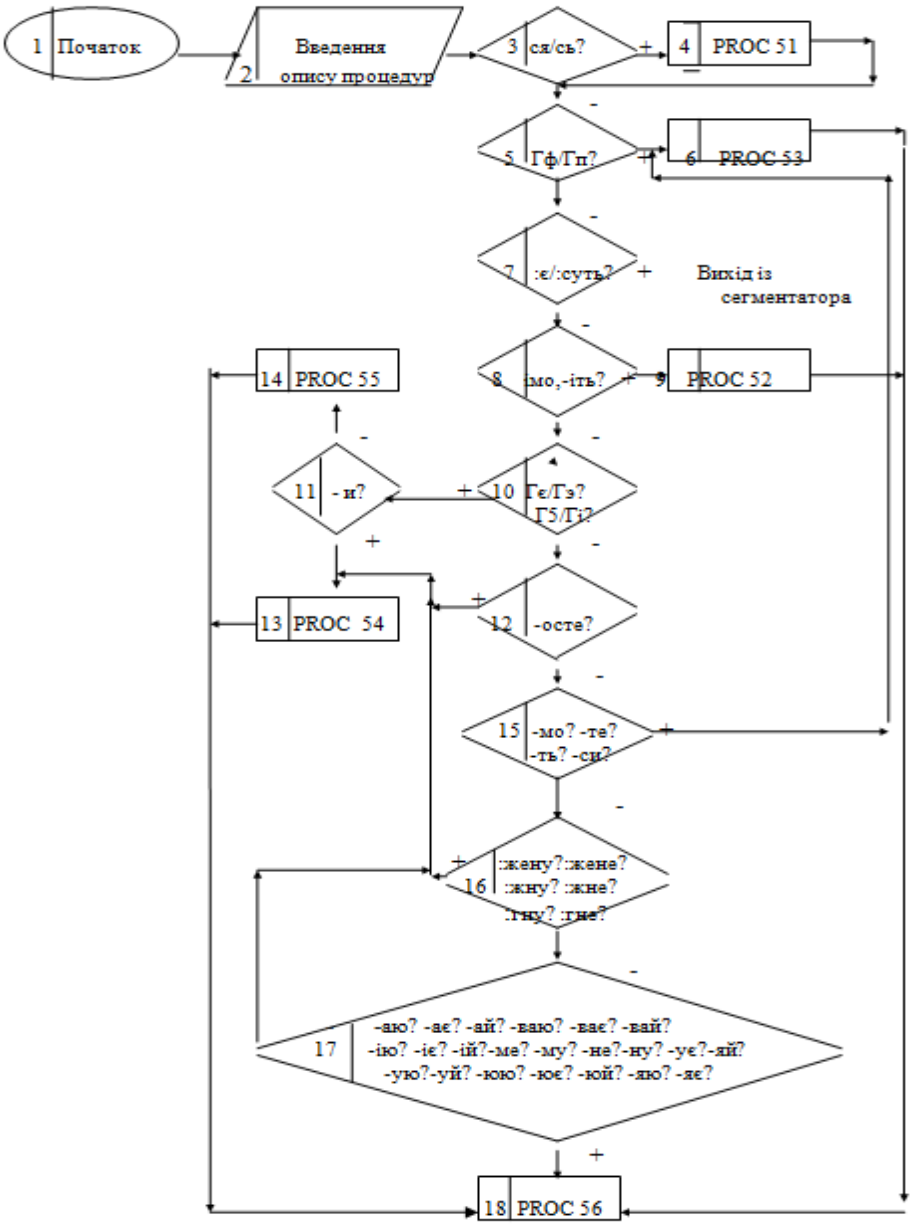


Рис. 1.2. Алгоритм виділення флексій у дієслівних словоформах

5) чи відповідає словоформа граматичним кодам Гф/Гп? Якщо так – п.6, якщо ні – п.7.

6) PROC 53.п.18.

7):є; :суть; – становлять лексему? Якщо так – вихід із сегментатора, якщо ні – п.8.

8) чи закінчує слово -імо, -іть? Якщо так – п. 9, якщо ні – п.10.

9) PROC 52. п 18.

10) чи відповідає словоформа граматичним кодам Г5/Гі, Гє/Гэ ? Якщо так – п.11, якщо ні – п.12.

11) чи закінчує слово графема -и? Якщо так – п.13, якщо ні – п.14.

12) чи закінчує слово комплекс графем -осте? Якщо так – п.13, якщо ні – п.16.

13) PROC 54. п. 18.

14) PROC 55. п.18.

15) чи закінчується слово графемами -мо? -те?-ть? -си? Якщо так – п.6, якщо ні – п.17.

16) :жену, :жене, :жну, :жне, :гну, :гне, входить до складу словоформи? Якщо так – п. 13, якщо ні – п. 17.

17) чи закінчується словоформа на: ну? не? ую? ує? ваю? ває? ію? іє? аю? ає? яю? яє? юю? ює? му? ме? Якщо так – п.18, якщо ні – п.13.

18) PROC 56.

Алгоритм виділення флексій у дієсловах української мови – це процесуальна аналітична модель, яка відображає схему аналізу та структуру дієслівної флективної парадигми, а також синтагматичні відношення флективних морфів із сусідніми суфіксами. Будь-яка дієслівна словоформа може бути проаналізована за цим алгоритмом, що моделює сегментацію дієслівних флексій на основі п’яти правил-процедур.

За структурою блок-схеми МС (рис. 1.1) та процедурою (56) описаного алгоритму після роботи в режимі блоку F сегментатор переходить у режим блоку S(Д), роботу якого моделює алгоритм виділення суфіксів у дієслівних словоформах української мови.

Дієслівна часова парадигма характеризується флективними морфемами, а також словозмінними суфіксами (минулий час та майбутній час недоконаного виду синтетичного способу), які разом зі словотвірними суфіксами дієслівних основ утворюють суфіксальні послідовності дієслівних словоформ. Цей факт було враховано при побудові алгоритму автоматичного виділення суфіксів, що умовно можна поділити на дві частини:

1) алгоритм автоматичного виділення словозмінних суфіксів;

2) алгоритмічне представлення морфемної структури суфіксальних послідовностей дієслівних основ.

Алгоритм виділення суфіксів у дієсловах української мови складніший, порівняно з попереднім алгоритмом. Флективна зона словоформи завжди характеризується наявністю однієї флексії, яка локалізована в кінці словоформи (постфікс вичленовується окремою процедурою алгоритму). Тому для встановлення межі між флексією й словозмінним суфіксом або суфіксом дієслівної основи достатньо змоделювати парадигму флективних морфів за їх кількісно-графемними характеристиками, не звертаючись до значень флексій: сегментація флексій проходить через процедуру ідентифікації флективних морфів словоформ із флексіями-еталонами алгоритму.

Суфіксальна зона українського дієслова має складну будову. Словозмінні суфікси виконують ту саму функцію, що і флексії – функцію передачі граматичних значень, і характеризуються визначеною позицією в морфемній структурі словоформи. Відповідно, перша частина алгоритму створюється подібно до алгоритму виділення флексій. Складним завданням алгоритмічного моделювання морфемної структури дієслівних словоформ є моделювання морфемної структури суфіксальних послідовностей дієслівних основ.

Суфіксальні послідовності дієслівних основ в українській мові характеризуються складною комбінаторикою морфів: основу дієслова, у переважній більшості випадків, закінчує дієслівний суфікс, якому можуть передувати словотвірні суфікси інших частин мови. Кількісно-суфіксальні послідовності можуть мати від однієї до чотирьох морфем у різних комбінаціях. Проте комбінаторика суфіксальних морфів системно обмежена, і тому може бути алгоритмічно описана за умови, що морфи однієї морфеми будуть представлені як різні елементи морфологічної системи мови. Таке спрощення структурних відношень між елементами мовної системи на цьому етапі побудови моделі вимагається формалізованим характером опису, інакше структура алгоритму МС набагато ускладнюється.

Для здійснення автоматичного сегментування суфіксальних послідовностей дієслівних основ, враховуючи взаємозалежність між формою та значенням морфів, необхідно побудувати даталогічну модель механізму морфемного аналізу, що репрезентує морфему як формальну одиницю (на рівні графем) і ґрунтується на:

1) формальних закономірностях комбінаторики морфів у структурах суфіксальних послідовностей;

2) дистрибутивно-опозиційних характеристиках графем на морфемному шві між коренем і суфіксальною послідовністю.

Концептуальну інфологічну модель сегментації посткореневої зони дієслівних основ утворюють два алгоритмічні дерева-графи, що репрезентують морфемну структуру суфіксальних послідовностей двох дієслівних основ:

1) основи інфінітива (граф А);

2) основи теперішнього часу (граф Б).

Роздільне моделювання основи інфінітива і основи теперішнього часу пояснюється емпіричними спостереженнями над суфіксальною зоною дієслівних основ: основу теперішнього часу можна представити як морфонологічний варіант основи інфінітива. Тому у формальному дослідженні, де морфи однієї морфеми вважаються різними елементами синтагматичної осі мовної системи, суфіксальні послідовності двох типів дієслівних основ характеризуються різною номенклатурою суфіксальних морфем.

Для представлення суфіксальних послідовностей дієслівних основ у алгоритмічному дереві була введена робоча одиниця – квазіфлексія, яка служить для автоматичного розрізнення омонімії між суфіксами й графемними комплексами або графемами при морфемному членуванні. Квазіфлексія репрезентує графемну діагностику на морфемному шві між коренем і суфіксальною послідовністю і є базовим поняттям динамічної моделі морфемного сегментування. Завдяки введенню поняття квазіфлексії

досягнуто мінімізації змістової інформації, необхідної при аналізі знакових одиниць – морфем, без урахування значення кореня[13].

Кожна гілка алгоритмічного графа відповідає окремій квазіфлексії, що моделює морфемну структуру суфіксальної послідовності основи та визначену дистрибуцію графем на морфемному шві між коренем і посткореневим суфіксом. Графемна дистрибуція однієї квазіфлексії через опозиційні відношення з дистрибуціями графем сусідніх квазіфлексій, служить маркером у визначені морфемної структури основи. Побудова алгоритмічного графа представляє техніку дистрибутивно-опозиційного методу лінгвістичного аналізу (рис.1.3).

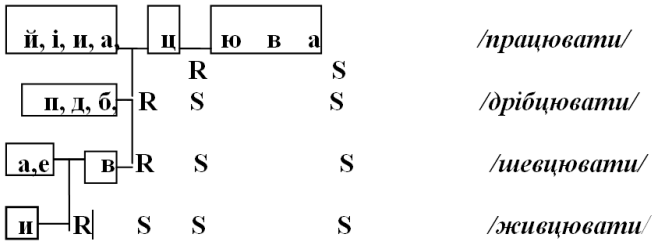


Рис.1.3. Фрагмент графічної моделі сегментації дієслівної основи інфінітива

Фрагмент графічного алгоритму складається з чотирьох гілок і містить інформацію про морфемне сегментування дієслівних основ інфінітива із суфіксом *-ова-,* що сполучається в препозиції з графемою /ц/. Ставиться завдання – представити повну графемну діагностику на морфемних швах між коренем і посткореневим суфіксом. Ця діагностика повинна протиставляти морфемні структури сусідніх інфінітивів, що формують алфавітний інверсійний список.

Перша гілка графа представляє суфіксальну структуру основ інфінітива, які на морфемному шві між коренем і суфіксом мають графему /ц/, що належить до кореня. Але для діагностування суфікса *-ова-* однієї кінцевої графеми /ц/ недостатньо, тому що існує суфікс *-ц-*, який репрезентує друга гілка графічного алгоритму. Із цієї причини для діагностики суфікса *-ова-* на першій гілці задається не одна графема /ц/, а мінімально-достатнє сполучення з двох кореневих графем (рис.1.4):

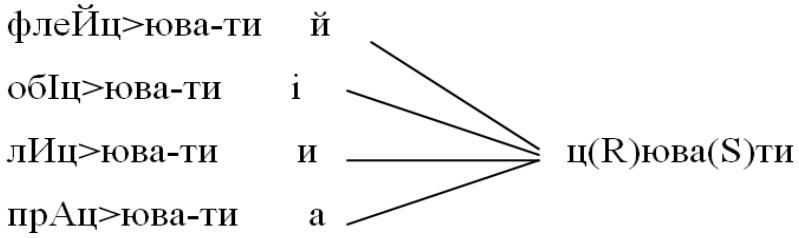


Рис.1.4. Морфемна структура першої гілки алгоритмічного графа

[13] Створення МС базується на даних морфемного словника І. Яценка, тому до алгоритмічного графа потрапляють дієслівні основи вже просегментовані на морфи за методом зіставлення, який враховує значення кореня. Тому значення кореня все-таки імпліцитно присутнє в МС.

Друга гілка алгоритму представляє сегментацію суфіксальної зони інфінітива за квазіфлексіями, у яких графемі /ц/, що становить суфікс, передують графеми /п/,/д/,/б/ (рис.1.5).

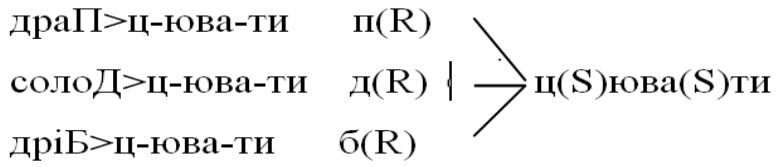


Рис.1.5. Морфемна структура другої гілки алгоритмічного графа

Третя гілка показує, що наявність перед графемою /ц/ графеми /в/ не є достатньою для проведення морфемної сегментації, тому що квазіфлексія /-вцювa/ характеризується двома типами морфемної структури суфіксальної послідовності: R2S, R3S. Для того, щоб зняти омонімію в цій дистрибуції, необхідно задати сполучення з трьох графем перед суфіксом *-ювa-* (рис. 1.6):

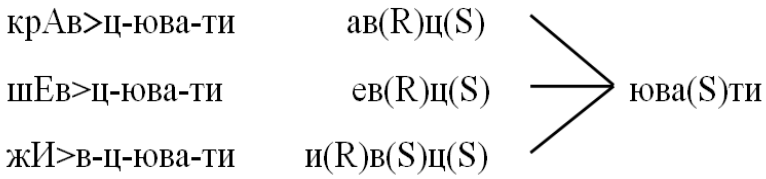


Рис.1.6. Морфемна структура третьої гілки алгоритмічного графа

Третя гілка представляє квазіфлексії /-евцювa/ та /-авцювa/, у яких графеми /е/ та /а/ визначають, що морфемний шов між коренем і суфіксом проходить між графемами /в/ і /ц/, де *-ц-* – суфікс: *шев>ц-ювa-ти, крав>ц-ювa-ти*.

Четверта гілка (див. Рис.1.3) моделює сегментацію морфемної структури суфіксальної послідовності інфінітива, що здійснюється на основі квазіфлексії /-ивцювa/, у якій диференційну функцію виконує графема /и/. Наявність у графемному складі цієї квазіфлексії графеми /и/ зумовлена також тим, що морфемна структура дієслова *жи>в-ц-ювa-ти* має суфікс *-в-*, для виділення якого необхідно мати графемне представлення лівобічного оточення графеми /в/.

Алгоритмічний граф суфіксальної зони основи теперішнього часу (рис.1.7) – дистрибутивно-опозиційна модель морфонологічного варіанта морфемної структури основи інфінітива, у якій враховано всі фонологічні альтернації, що відбуваються у дієслівній основі при утворенні словоформ теперішнього часу та наказового способу.

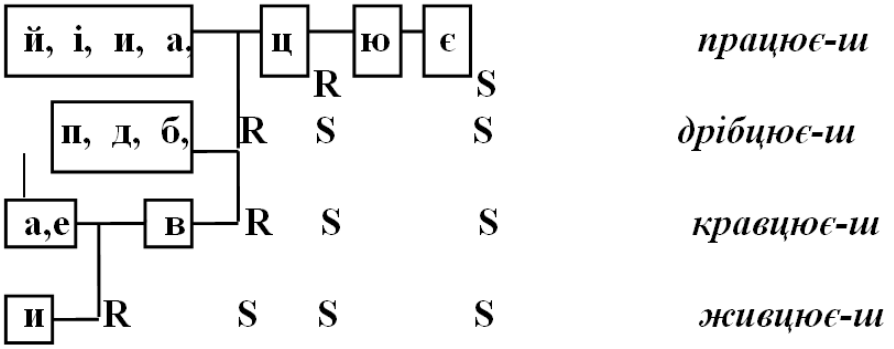


Рис. 1.7. Фрагмент алгоритмічного графа морфемної структури основи теперішнього часу

Загальний принцип побудови графічної дистрибутивної моделі основи теперішнього часу такий самий, як і у варіанті основи інфінітива: відмінність графемного вираження сусідніх квазіфлексій хоча б за однією графемою визначає тип морфемної структури суфіксальної зони основи:

ацює – RS *працю>є-(ш);*

бцює – R2S *дріб>ц-ює-(ш)*;

евцює –R2S *шев>ц-ює-(те)*;

ивцює – R3S *жи>в-ц-ює-(те)*.

Як свідчить інфологічна модель морфемної сегментації, квазіфлексії – це синтагматичні відрізки слова. Правобічна межа суфіксальної послідовності основи слова завжди збігається із межею квазіфлексії. Лівобічна межа суфіксальної послідовності дієслівної основи може збігатися з межею квазіфлексії (що зустрічається дуже рідко) або нарощуватися графемами кореневого морфа. Наприклад, морфемна сегментація основ дієслів *працювати, шевцювати, живцювати* моделюється квазіфлексіями: /евцюва/, /авцюва/, /ивцюва/. Графеми /е/,/а/,/и/, що належать до кореня словоформи, виступають диференційною ознакою квазіфлексій і сигналізують про різну морфемну сегментацію суфіксальної зони дієслівних основ, які на морфемному шві між суфіксом *-ювa-* й наступною морфемою (сегментація проходить у інверсійному порядку) мають графему /ц/.

У межах динамічної моделі морфемної сегментації квазіодиниці такого типу називаються квазіфлексіями, але це не значить, що вони мають таке функціональне навантаження, як і реальні флексії. Квазіфлексії не виконують словозмінної функції, а приклад фрагмента дерева-графа показує, що конструктивно квазіфлексії можуть представляти всю суфіксальну послідовність і навіть частину кореня. Оскільки у флективних мовах граматична інформація зосереджена переважно в кінці словоформ, моделювання морфемної сегментації передбачає, що лінгвістичний процесор буде ”зчитувати” словоформу з кінця, в інверсійному порядку, тому мінімальне сполучення графем, необхідне для сегментації суфіксальної зони основи дієслова – формально виділена одиниця – асоціюється із частиною, що закінчує слово – флексією. Саме тому ця квазіодиниця називається квазіфлексією, а не квазісуфіксом, квазіосновою чи квазікоренем.

Для створення лінгвістичного процесора автоматичного морфемного сегментування посткореневої зони українського дієслова необхідно представити концептуапльну лінгвістичну модель морфемної структури суфіксальних послідовностей дієслівних основ – два алгоритмічні графи – у більш формалізованому вигляді, придатному для використання у розробленні програмного забезпечення, тобто описати морфемні структури суфіксальних послідовностей дієслівних основ у термінах даталогічних моделей – формул програмних процедур. Кожна квазіфлексія як одиниця лінгвістичної моделі має субстанціональне вираження – ланцюжок графем та структурну організацію, що відображає морфемну структуру суфіксальної послідовності дієслівної основи. Структура кожної квазіфлексії може бути представлена через програмну процедуру. 4037 квазіфлексій, визначені за інверсійним списком ≈ 27 тисяч дієслів української мови, були змодельовані на даталогічному етапі 71 програмною процедурою. У даталогічній моделі знаком (*) позначається графема в ланцюжку графем, знаком (>) – морфемний шов між коренем і суфіксом, знаком (-) – морфемний шов між двома суфіксами, наприклад: >*-*-* бали (*недб>а-л-и-ти*). Програмні процедури можуть моделювати морфемні структури як однієї структурної основи, так і двох структурних основ, наприклад: процедура >** моделює морфемну структуру посткореневої зони основи інфінітива /*да>ва давати*/ та основи теперішнього часу /*руб>ає рубаєш/* через зіставлення з різними квазіфлексіями. Програмні процедури – метамова даталогічної моделі МС, яка моделює всі типи морфемних структур суфіксальної зони дієслівних основ через комбінаторику кількісно-графемних типів суфіксальних морфем: 71 програмна процедура (Додаток 1.1) представляє 71 структурний тип сегментації суфіксальних послідовностей дієслівних основ у графемному представленні.

Програмні процедури моделюють процес морфемної сегментації, що буде реалізований за допомогою комбінації простих операцій над графемним складом дієслівної основи: операція зсуву і операція приписування символу. Операція зсуву вказує на те, скільки графем із кінця дієслівної основи потрібно зсунути, щоб поставити мітку кореня чи суфікса. Операція приписування символу ставить мітку кореня чи суфікса на місці визначеного морфемного шва.

Запропонована форма запису процедур морфемної сегментації дієслівної основи в експліцитному вигляді виявляє структурні характеристики морфемного складу суфіксальних послідовностей відносно кількості морфем, їх графемно-кількісного представлення, валентності морфем одного типу в межах слова. Іншими словами, виведені процедури можуть розглядатися як метамова опису морфемної структури дієслівних основ української мови, а також як об’єкти лінгвістичного дослідження. За умови побудови аналогічних морфемних сегментаторів для інших флективних мов, представлені в такий спосіб процедури членування

морфемної структури можуть розглядатися як еталонні в типологічних дослідженнях.

На основі двох списків квазіфлексій (список А квазіфлексій основи інфінітива та список Б квазіфлексій основи теперішнього часу), а також графемних моделей словозмінної парадигми дієвідмінювання, було створено алгоритм автоматичного виділення суфіксів у дієсловах української мови, що забезпечує роботу блока S морфемного сегментатора системи АГАТ.

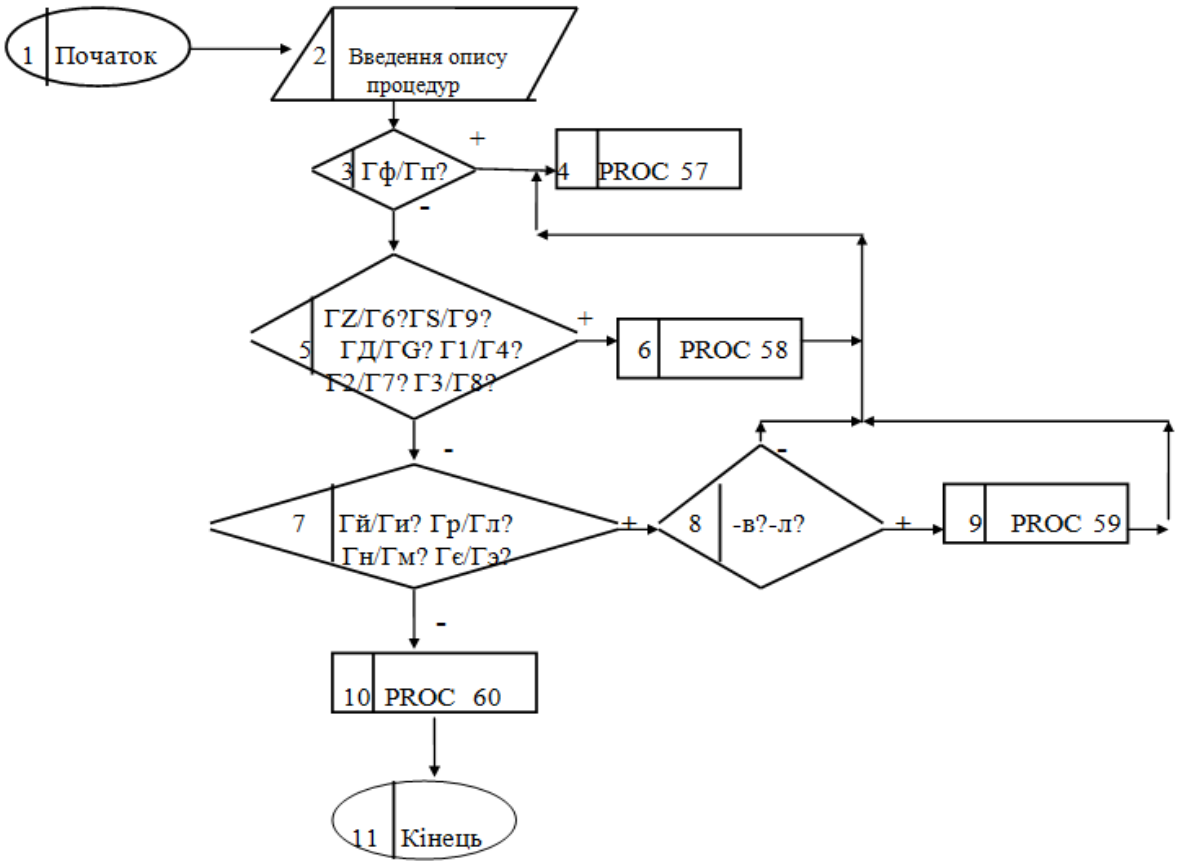


Рис.1.8. Алгоритм автоматичного виділення суфіксів у дієслівних словоформах української мови

Опис алгоритму:

1.Початок.

2. Введення опису процедур:

1)PROC 57/список А/ (сегментація за списком квазіфлексій основи інфінітива);

2) PROC 58 /-**-**/ (сегментуються словозмінні суфікси дієслівної парадигми майбутнього часу доконаного виду: -ти-му; -ти-ме);

3) PROC 59 /-*/ (відділяється словозмінний суфікс минулого часу -в,-л);

4) PROC 60 /список Б/ (сегментація за списком квазіфлексій основи теперішнього часу).

3. Чи відповідає словоформа граматичним кодам Гф/Гп? Якщо так – п.4, якщо ні – п.5.

4. PROC 57.

5. Чи відповідає словоформа граматичним кодам Гz/Г6, Гs/Г9, ГD/ГG, Г1/Г4, Г2/Г7,Г3/Г8 ? Якщо так – п.6, якщо ні – п.7.

6. PROC 58, п.4.

7. Чи відповідає словоформа граматичним кодам Гй/Ги, Гр/Гл, Гн/Гм, Гє/Гэ? Якщо так – п.8, якщо ні – п.10.

8. Чи передують визначеній флексії графеми -в-, -л-? Якщо так – п.9, якщо ні – п.4.

9. PROC 59, п.4.

10. PROC 60.

11. Кінець.

Алгоритм морфемного сегментування суфіксальної зони дієслівних словоформ – повністю формалізований процес морфемного аналізу, який представляє логічну стадію даталогічної динамічної моделі автоматичного морфемного аналізу: у кожному дієслівному слововживанні українського тексту посткоренева зона може бути просегментована на морфеми. Це означає, що метод комп'ютерного моделювання дає точну процедуру виявлення подібностей і відмінностей лінгвістичних визначень, встановлення їх еквівалентності, трансформування цих визначень у більш зручну та елементарну форму, що відповідає вимогам автоматизації лінгвістичного аналізу. Створена динамічна модель може бути реалізована в програмно-операційному середовищі даталогічного етапу.

1.3.2. Принципи формалізації морфемного аналізу посткореневої зони дієслівних словоформ української мови

Принципи створення даталогічної моделі морфемного сегментатора демонструють формалізацію морфемного аналізу, що породжує нові гносеологічні одиниці – квазіфлексії, які є елементами формальної мови (ФМ) моделі. Будь-яка формальна процедура сама по собі виявляє тільки формальні елементи – елементи штучної метамови, але не може визначити спосіб переходу від елементів ФМ до онтологічних значеннєвих одиниць мови-об'єкта – природної мови (ПМ). «[...] замінюючи одну мову іншою мовою, ми можемо працювати із цією мовою й встановлювати різні взаємозв'язки між її елементами. Але це не дає права автоматично переносити закони "нової мови" на факти "старої" мови: таке перенесення можна визнати достовірним, завчасно довівши ізоморфізм обох мов» [Сова 1970: 156]. Якщо в процесі лінгвістичного дослідження виникає потреба перевести природну мову в штучну, то необхідно зафіксувати цей перехід, сформувавши правила переходу від ПМ до ФМ і навпаки – від ФМ до ПМ.

Саме в неправомірній заміні формально-значеннєвих одиниць природної мови формальними одиницями криється причина того, що лінгвістам не вдається перенести апарат системи абстракцій, створений

логіками, на мовні об’єкти й побудувати на основі логіко-математичних абстракцій формальну теорію мови. Це стосується також і формальних описів мови, розроблених Ч. Фрізом [Fries 1952], Н. Хомським [Хомский 1961], З. Харрісом [Харрис 1962], у яких аналізуються конструкти формальних (штучних) мов.

Беручи до уваги той факт, що тип концептуальної інфологічної моделі мовного опису повинен обиратися залежно від сфери його подальшого застосування (у нашому випадку побудова лінгвістичного процесора автоматичного морфемного сегментування), у процесі формалізованого опису посткореневих морфемних послідовностей українського дієслова формалізація здійснювалась у два етапи, які представляють два типи моделей:

1) дерева-графи – статичні концептуальні моделі морфемної структури суфіксальних послідовностей дієслівних основ;

2) алгоритми морфемної сегментації – динамічні даталогічні моделі, що представляють не даність мовного феномена (морфемної структури), а процес аналізу цього феномена.

В основу формалізованого опису було покладено дві гіпотези:

ГІПОТЕЗА 1. Суфіксальні послідовності дієслівних основ функціонують у мові як сталі синтагматичні одиниці, що характеризуються певною морфемною структурою.

ГІПОТЕЗА 2. Морфемна структура суфіксальних послідовностей дієслівних основ може бути визначена через опозицію модельних конструктів-квазіфлексій, диференційну функцію в яких виконує дистрибуція графем на морфемному шві між коренем і суфіксальною послідовністю.

Суть гіпотетичного опису морфемної структури суфіксальних послідовностей дієслівних основ – формалізоване представлення відношень між значенням і формою морфем та формалізоване представлення поєднання морфем у суфіксальні послідовності.

На основі ГІПОТЕЗИ 1 було звернуто увагу на те, що кількість суфіксів, префіксів і флексій у мові обмежена, а коренева система відкрита й за кількістю одиниць набагато більша. Суфікси і префікси, сполучаючись, утворюють певні комбінації, кількість яких теж можна точно визначити. Тому в процесі автоматизації морфемного аналізу немає необхідності програмувати повну морфемну комбінаторику кожної словоформи. Потрібно змоделювати лише морфемні структури префіксальних і суфіксальних послідовностей, а корінь буде вичленовуватися як частина, що залишилась після виділення докореневої і посткореневої частин.

Для створення автоматичної системи морфемного сегментування інформація про комбінаторику афіксальних морфів є недостатньою, тому що на морфемних швах між коренем і афіксами можлива омонімія між морфемами та графемними сполуками або окремими графемами, яка в

традиційному лінгвістичному аналізі знімається першочерговим виділенням кореня за умови наявності інформації про лексичне значення слова.

У моделі морфемної сегментації суфіксальних послідовностей дієслівних основ цю проблему вирішує квазіфлексія як результат дистрибутивно-опозиційного аналізу графем на мофемному шві між коренем і посткореневим афіксом – ГІПОТЕЗА 2.

Квазіфлексія виступає центральною формальною одиницею інфологічної та даталогічної моделей. Можливість репрезентації форми вираження квазіфлексії на графемному рівні мови, де первинна звукова субстанція мовної системи представлена через іншу субстанцію – графічну, зумовлюється фонематичним характером принципів української орфографії. Кожна графема в переважній більшості випадків (33графеми = 38 фонемам) репрезентує фонему як варіант класу звуків, що розглядається в його диференційній функції. Фонема – елемент мовної системи, звук – мовна субстанція. Фонему як абстрактну сутність не можна вимовити, тобто представити за допомогою тієї субстанції, від якої вона абстрагована. Графема – елемент іншої фізичної субстанції, тому індивідуальні особливості звукових варіантів, зумовлені позицією, та особливості індивідуальної вимови нейтралізуються в графемі як моделі фонеми. Графема – модель фонеми, і тому виконує диференційну функцію, ідентичну до функції фонеми, яку вона репрезентує. Тому морфемна структура, визначена на фонемному і графемному рівні, репрезентує ідентичні відношення між морфемами, відмінною буде лише в деяких випадках структурна організація плану вираження морфів.

Графема як субстанція квазіфлексій набуває дещо нових функцій: визначена через дистрибутивно-опозиційні структурні відношення у дереві-графі, графема, що закінчує квазіфлексію (в інверсійному прочитанні), ідентифікує морфемну структуру основи аналізованої словоформи з програмною процедурою квазіфлексії, і таким способом диференціює морфемні структури сусідніх квазіфлексій дерева-графа. У цій функції графема виступає як елемент даталогічної моделі МС.

Наприклад:

форма вираження квазіфлексії	структурна організація квазіфлексії (програмна процедура)	графема, що виконує ідентифікуючу та диференціальну функції	морфемна структура основи аналізованої словоформи
вивіша	>*	в	вивіш>а-(ти)
живіша	>*-**-*	ж	жи>в-іш-а-(ти)
ливіша	>**-**-*	л	сміл>ив-іш-а-(ти)
сивіша	>**-*	с	сив>іш-а-(ти)

Наведений приклад показує, що дистрибутивно-опозиційний аналіз графем на морфемному шві між коренем і суфіксом визначає правобічне і лівобічне графемне представлення квазіфлексії:

в|ивіша;

ж|ивіша;

л|ивіша;

с|ивіша;

Правобічна частина квазіфлексії – графемне представлення частини дієслівної основи, що характеризується ідентичністю субстанціонального (графемного) вираження сусідніх квазіфлексій моделі-дерева (ивіша). Лівобічна частина – визначені в препозиції до правобічної частини через опозицію в дереві-графі різні графеми (в, ж, л, с).

Правобічне графемне представлення квазіфлексії виділяється на основі недискретного принципу дослідження мовних явищ, у якому основна увага зосереджується не на виділенні одиниць мовних рівнів, а на фактах частково тотожного представлення форми вираження об'єктів мовної системи. Недискретному принципу мовного опису відповідає недискретність в організації мови-об'єкта. На морфемному рівні мови виділяються кінцеві автономні одиниці морфемного аналізу – морфи, які ідентифікуються в парадигматичні одиниці – морфеми. Недискретне членування морфемної структури дозволяє виділити квазіморфеми. «Квазіморфема може бути визначена як будь-який сегмент лексичної одиниці, який у плані вираження є повним або частковим омофоном [у нашому дослідженні омографом] [14] тотожного сегмента іншої лексичної одиниці» [Степанов 1975: 52].

Недискретний принцип аналізу дозволяє виділити в дієсловах */вивішати/, /живішати/, /сміливішати/, /сивішати/* квазіморфему -ивішати, а в основах відповідних дієслів квазіморфему -ивіша. Виділена таким способом одиниця репрезентує тільки графемну маніфестацію суфіксальної послідовності дієслівної основи, не вказуючи на її морфемну структуру.

Квазіморфема як одиниця моделі морфемного сегментування – квазіфлексія – характеризується наявністю графемного вираження і морфемної структури, яка описана програмною процедурою. Диференціація різних типів морфемних структур здійснюється через диференціацію графемного контексту лівобічної частини квазіфлексії.

Отже, квазіфлексія – одиниця формальної мови, одиниця "штучна", яка не виділяється на жодному рівні мовної системи й вводиться в модель на основі наукового поняття ідеалізації.

«Ідеалізація – це метод абстракції, суть якого в тому, що об'єкти науки утворюються в результаті відчуження від принципової неможливості створити їх експериментальним шляхом» [Шаумян 1965: 28]. Такі об'єкти визначаються як ідеалізовані. У математиці – це крапка, пряма, лінія... У формалізованому дослідженні морфемної системи мови – це квазіфлексія.

[14] Вставлення автора.

Префікс "квазі-" в терміні "квазіфлексія" вжито за аналогією до вживання цього префікса в математичних термінах "квазіпорядок", "квазігрупа" тощо. У подібних термінах префікс "квазі-" означає послаблення вимог до поняття.

Лінгвістична одиниця – флексія визначається як морфема, що закінчує словоформу й виконує словозмінну функцію. У понятті "квазіфлексія" функціональне навантаження змінюється. Квазіфлексія – позиційно кінцева частина дієслівної основи, яка разом із програмною процедурою, що репрезентує структуру квазіфлексії,виконує такі функції:

1) описує групу словоформ, які мають однакову морфемну структуру і графемну маніфестацію суфіксальних послідовностей дієслівних основ. Наприклад, квазіфлексія /зичи - */ описує словоформи, у яких суфіксальна послідовність дієслівної основи представлена одним суфіксальним морфом – -и-, а лівобічне графемно-дистрибутивне оточення – графемним контекстом /зич/: *зич>и-ти, зазич>и-ти, позич>и-ти, запозич>и-ти, язич>и-ти;*

2) диференціює типи сегментації (морфемні структури) суфіксальної зони дієслівних основ:

/зичи -*/ *зич>и-ти;*

/узичи -**-* /*муз>ич-и-ти;*

/:личи -*/*лич>и-ти;*

/велич -**-* /*звел>ич-и-ти;*

/теличи -*/*спантелич>и-ти.*

Диференційна функція квазіфлексії визначає її в межах моделі як значущу одиницю.

Постулювання поняття квазіфлексії проходить відповідно до трьох етапів ідеалізації, визначених Д. Горським: «1) змінюючи деякі умови, у яких знаходиться предмет, що вивчається, ми робимо їх дію монотонно спадною; 2) при цьому виявляється, що деякі властивості цього предмета також монотонно змінюються; 3) передбачаючи, що дія умов на предмет, який вивчається, зведена до нуля, ми здійснюємо перехід до деякого ідеалізованого об’єкта» [Горский 1961: 180].

Етап 1: якщо розглядати суфіксальну послідовність як автономну сутність (гіпотеза 1), то порушується семантичний зв’язок між коренем і суфіксальною послідовністю й, відповідно, знімається актуалізація асоціативних значень суфіксальних морфем як двопланових одиниць – етап 2. Оскільки встановити межу між коренем і суфіксом за таких умов не можливо через звернення до семантики слова, уводиться поняття "квазіфлексії" як робочої одиниці дистрибутивно-опозиційної моделі – етап 3 (гіпотеза 2). Квазіфлексія, виконуючи описову і диференційну функції, дозволяє формалізувати процес сегментації морфемної структури суфіксальних послідовностей дієслівних основ.

Ідеалізація дає можливість розмірковувати про ідеалізовані об’єкти як про існуючі в дійсності, хоча в дійсності існують лише їх прообрази. Але якщо модель, одиницями якої є абстрактні поняття, функціонуючи, буде сегментувати словоформи на ті самі сегментні відрізки, що й лінгвіст при

морфемному аналізі, то можна вважати, що у відношенні сегментації посткореневого фонду дієслова модель є ізоморфною, відповідно, гіпотетичний опис механізму морфемного аналізу – правильний.

Якщо формалізована процедура побудована як формалізований опис результатів інтуїтивної процедури, то вона буде аналізувати всі можливі словоформи тексту й представляти дані про морфемну структуру мови, адекватні даним першого інфологічного етапу опису мови. Виходячи з граматики, яка проводить опис мовних елементів ”від змісту до форми”, лінгвістичний опис, спрямований на формалізацію мовного явища приходить до граматики, що репрезентує опис мовних елементів ”від форми до змісту”. Ці граматики взаємопов’язані й становлять один граматичний опис мови: якщо граматика 1 дає необхідні дані про мовні елементи для їх подальшої формалізації, то формальна граматика доповнює першу граматику новими даними про структурні відношення між морфемами суфіксальних послідовностей дієслівних основ, які можна отримати тільки методом моделювання структури мовної системи.

Формальна лінгвістична теорія, тобто модель, дає процедуру відкриття формальної граматики, що виступає посередником між мовою (граматикою 1) і формалізованим механізмом аналізу цієї мови. Формальна граматика «встановлює правила кореспонденції між символічним апаратом теорії і одиницями тексту» [Засорина 1971: 266].

Лінгвістична інтерпретація алгоритмічного опису посткореневої зони дієслова є аналітичною граматикою української мови, у якій в експліцитній формі подано формальні процедури виділення морфів, а також аналіз і класифікацію одиниць морфемного рівня, що задовольняє вирішення двох головних задач: практичної – автоматичне виділення морфем у тексті, теоретичної – лінгвістичне дослідження морфемної структури слова.

Комп’ютерна морфеміка посткореневої зони українського дієслова представляє нові лінгвістичні дані про структурну організацію дієслівних основ і словоформ, які можуть бути отримані тільки методом моделювання, що зумовлено онтологічною сутністю структури мовної системи: структурні відношення "приховані" від безпосереднього спостереження.

Створення дистрибутивно-опозиційної моделі механізму морфемного сегментування стало можливим завдяки введенню робочої одиниці аналізу – квазіфлексії, що дозволило побудувати динамічну даталогічну алгоритмічну модель МС посткореневого фонду дієслова. Ця модель може визначати будь-який посткореневий морф дієслівної основи на базі 4037 квазіфлексій описаних 71 програмною процедурою, будь-який словозмінній суфіксальний морф на основі двох програмних процедур, будь-яку дієслівну флексію на основі 5-ти формальних визначень-процедур.

У процесі створення моделі були виявлені обмеження розробленої формальної процедури, адже мова має багато таких феноменів, вивчення яких не досягло ще, а, можливо, і ніколи не досягне рівня формальної моделі. Формалізації морфемного аналізу в моделі МС не підлягають омографи,

морфемну структуру яких не можна протиставити за графемною диференційною ознакою. Неможливо зняти графемну омонімію на морфемних швах між коренем і суфіксом за допомогою дистрибутивно-опозиційного аналізу в основах таких дієслів-омографів:

1) */вид>`а-ти/ – /в`ида>ти/;*
2) */доп`і>к-а-ти/ – /допік>`а-ти/;*
3) */г`у>к-а-ти/ – /гук>`а-ти/;*
4) */спі>ш-`и-ти/ – /сп`іш>и-ти/.*

У цих випадках розрізнення омонімії та морфемних структур слів можливе тільки за наявності дистрибутивної ознаки – наголошена / ненаголошена фонема, яка не враховується в графемному представленні слів описуваної моделі.

У дієсловах, які є повними лексичними омонімами:

1) /нарив>а-ти/ – /нари>ва-ти/;
2) /перед>ува-ти/ – /переду>ва-ти/;
3)/висну>ти/ – /вис>ну-ти/; розпізнавання омонімії за формальними дистрибутивними ознаками в межах слова взагалі не можливе.

Теоретична і практична значущість застосованої методики формалізації лінгвістичного опису морфемної структури посткореневого фонду українського дієслова полягає не лише в алгоритмічній об'єктивності формальних операцій визначення морфемних сегментів у дієслівних словоформах, а також у тому, що створена даталогічна модель виступає способом побудови структурної морфемології суфіксальної зони українського дієслова, яка репрезентує нові дані про структурні відношення суфіксальних морфів у морфемній системі української мови.

1.4. Комп'ютерне моделювання морфемної структури початкових форм слів української мови в АСМСА

1.4.1. Концептуальна модель АСМСА: методологічні засади морфемного та словотвірного аналізів

У сучасній лінгвістиці із 70-их років минулого століття усталилася тенденція утвердження морфемного аналізу як самостійної сфери дослідження, на противагу словотвірному аналізу. У центрі уваги – проблеми морфемної структурної організації слова: валентність, дистрибуція, обмеження сполучуваності морфів, структурні типи слів, моделі морфемної будови слів та інших структурних морфемних одиниць, статистичні характеристики різного типу морфів, укладання різноманітних морфемних словників. Вивчення закономірностей організації морфем у структурі словоформи здійснюється на базі вже відомої номенклатури морфем, зафіксованих у морфемних словниках, які почасти засвідчують визначення різних морфемних структур у тих самих словах: *легенд-арн-ий* (Л. Полюга, «Морфемний словник» [Полюга 1983]), *легенд-ар-н-ий* (І. Яценко,

«Морфемний аналіз. Словник-довідник» [Яценко 1980]). Така невідповідність пояснюється змішуванням принципів словотвірного та морфемного аналізів у визначенні морфемної структури слова.

Морфемний і словотвірний аналізи оперують поняттям морфеми як знакової одиниці, тому встановлення морфемної і словотвірної будови слова вимагає виокремлення сегментів, які мають значення, і це значення в більшості випадків встановлюється в словотвірних відношеннях: словотвірна морфема мотивованого слова має словотвірне значення, яким це слово відрізняється від мотивувального. Словотвірний підхід використовується у визначенні морфемної будови слова, що призводить до змішування різних типів аналізу і їхніх одиниць, і як наслідок – до визначення різної морфемної структури слова.

Аналізуючи розвиток теорії морфологічного членування російської словоформи, С. Богданов [Богданов 1997] зазначає, що протиставлення морфемного і словотвірного типів аналізу в радянському мовознавстві в період 1950 – 1960-х рр. і пізніше здійснювалось у двох напрямах:

1) визначення морфемного аналізу як однієї зі стадій (початкової) словотвірного аналізу [Шанский 1959];

2) визнання цих двох видів аналізу незалежними одна від іншої сферами дослідження, які вирішують різні завдання шляхом застосування різних методик сегментування словоформ [Кубрякова 1974].

Сам С. Богданов уже в кінці 90-х рр. визначає морфемний аналіз як «...кінцевий етап лінгвістичної сегментації слова (тобто кінцевий результат морфемного аналізу – визначену морфемну структуру, утворювану протиставленням функціонально охарактеризованих складників – морфем, – можна отримати тільки після здійснення формо- і словотвірного членування). Відповідно [...] 1-й і 2-й етапи морфемного членування є реалізацією формотвірного і словотвірного типів аналізу внутрішньої структури словоформи» [Богданов 1997: 34].

Розмежування двох типів морфологічного аналізу в більшості лінгвістичних досліджень має дещо декларативний характер. Активний розвиток теорії і практики словотвірного аналізу, порівняно з морфемним, зумовлює визначення морфемної структури в термінах словотвірного аналізу, чи навпаки – словотвірної структури в термінах морфемного аналізу. Так, у мовознавчих студіях побутує думка про опис словотвірної структури слова як багатоступеневої деривaційної організації мотивованого (Є. Карпіловська, Н. Клименко [Клименко 1998а]; В. Лопатін [Лопатин 1977], А. Тихонов [Тихонов 1990]), а не двочленної конструкції в термінах словотвірної основи та словотвірного форманта. Відповідно, морфемна сегментація словоформ здійснюється методами словотвірного аналізу, ґрунтуючись на понятті словотвірної пари. Застосування методики, що використовує поняття морфемної структури слова як результату послідовного здійснення актів словотворення (визначена морфема повинна відображати один словотвірний такт) [Головин 1977] зумовило в сучасній

теорії мовознавства визначення морфеми як центральної одиниці словотвірного рівня [Тихонов 1971].

Статус морфеми як словотвірної одиниці й узаконення комплексних морфем (біморфем (конфіксів) [Шуба 1975], тричленних морфем [Бирюкова 1975], поліморфем [Котова 1978]) суперечить традиційному визначенню морфеми Бодуена де Куртене, який визначав морфему неподільним мінімальним мовним знаком. Такий підхід зумовлює не тільки певні термінологічні невідповідності, а й спричинює автоматичне перенесення, накладання формотвірної і словотвірної структур на морфемну структуру словоформи чи навпаки, і таким чином зрівнює одиниці різних типів аналізу та зумовлює різне морфемне членування у словниках та граматиках.

Внутрішня структура словоформи на морфологічному рівні організації може визначатися трьома типами аналізу: 1) формотвірним; 2) словотвірним; 3) морфемним. Формотвірний аналіз використовує поняття граматичної основи слова та форматива й здійснюється на базі визначення структурно-системних відношень у межах парадигми словоформ одного слова. Словотвірний аналіз визначає словотвірну основу і словотвірний формант, зіставляючи слова одного словотвірного гнізда за принципами словотвірної мотивації. Морфемний аналіз ставить завдання визначити структуру словоформи на рівні організації мінімальних знакових одиниць морфем, ґрунтуючись на структурних відношеннях морфем у системі морфеміки в цілому. Три типи морфологічного аналізу визначаються різними цілями, методикою й різними структурними одиницями, і хоча ці різновиди морфологічного аналізу використовують поняття "морфема", проте одиниці різних морфологічних структур – словотвірна база та словотвірний формант, граматична основа словоформи та форматив – за своїми функціями та межами переважно не дорівнюють морфемі.

Морфемна структура словоформи за своїм кількісно-морфемним складом не обов’язково відображає морфемні межі, що актуалізуються в словотвірних та формотвірних структурах. Морфемна структура словоформи, що визначає морфеми як базові онтологічні одиниці морфемного рівня мови, репрезентує максимально можливу кількість морфемних меж словоформи (*дерев-ин-к-а, аналіз-ова-н-ий*), а кожна функціональна структура цієї словоформи реалізує лише їх частину, але не фіксує меж відсутніх у морфемній структурі словоформи: словотвірна структура – *деревин-к(а), аналізова-н(ий)*; формотвірна – *аналізован-ий, деревинк-а*. До того ж, на базі однієї морфемної структури словоформи можуть реалізовуватися різні формотвірні структури за умови синтаксичної неактуалізації формотвірної морфеми: так морфемна структура словоформи *колом* може реалізовуватися у двох формотвірних структурах – *кол-ом* (іменник орудного відмінка однини, де -ом – флексія), а у випадку адвербіалізації – *стати кол-ом* – -ом не виконує функції флексії.

Словотвірні та формотвірні форманти можуть мати власну морфемну структуру, що включає декілька морфів. Зокрема це стосується

різноманітних конфіксів та словотвірних чи формотвірних формантів, що об'єднують інтерфікс із суфіксом або ж суфікс із флексією. Морфемна структура таких складних компонентів слугує засобом експлікації структурно-семантичних системних відношень словоформи. Так, в утворенні дієслівної форми минулого часу *фарбувала* бере участь форматив –ла, який складається з двох морфем -л- та -а: суфікс -л- відображає структурно-семантичні відношення словоформи в часовій парадигмі і є носієм граматичного значення минулого часу, а флексія –а – структурно-семантичні відношення словоформи в межах парадигми минулого часу й виражає граматичне значення жіночого роду та однини.

Семантична цілісність конфіксальних компонентів руйнується як на рівні парадигматичних відношень у системі морфеміки, так і на рівні синтагматичної реалізації морфем, оскільки компоненти конфіксів виконують різне функціональне навантаження: префікси виконують розрізнювальну, модифікаційну функцію, а суфікси – узагальнювальну, класифікаційну або суміщують класифікаційну функцію з модифікаційною [Карпіловська 1999].

Морфеми парадигматично протиставляються в межах своєї підсистеми як на позиційно-формальному рівні, так і функціонально-значеннєвому. У такому парадигматичному протиставленні корелятивна взаємовизначеність префікса і суфікса, що позиційно пов'язані з одним коренем, не заперечує структурно-морфемної самостійності цих афіксів. Навпаки, їх здатність поєднуватися з іншими коренями чи автономно, чи обопільно характеризує обов'язкові структурні відношення: додаткової дистрибуції афіксальних морфем та контрастної дистрибуції суміжних кореневих морфем [Зубань 2000]. Якщо суфікси (наприклад -ок-, що виражає значення 'предметності') знаходяться у відношенні додаткової дистрибуції, тобто сполучаються у препозиції з різними коренями біл- (*біл-ок*), жовт- (*жовт-ок*), -глин- (*су-глин-ок*), то вони визначаються як синтагматичні знакові одиниці, що репрезентують морфему -ок-, незважаючи на префікс су-, який бере участь у префіксально-суфіксальному способі словотвору (*суглинок*).

Про парадигматичну самостійність префікса су- та суфікса -ок- свідчить їх здатність вступати у синонімічні та омонімічні відношення в морфемній підсистемі української мови, зокрема префікс су- може виражати такі словотвірні значення: 1) 'вказує на вияв ознаки, названої мотивувальною основою, більшою мірою' (*суглинок, супісок, сукровиця*); 2) 'вказує на сукупність, цілісність ознак, названих мотивувальною основою' (*сув'язь, сузір'я, супліддя*); 3) 'вказує на неповний вияв ознаки, названої мотивувальною основою' (*сугорб, сумерк, сутінки*); 4) 'вказує на суміжність, зв'язок із тим, що названо мотивувальною основою' (*супутник, сусід*). Суфікс -ок- може виражати такі словотвірні значення: 1) 'вказує на предметність (речовину) за ознакою, названою мотивувальною основою' (*суглинок, білок, жовток)*; 2) 'вказує на демінутивне значення предмета, названого мотивувальною основою' (*дубок, горбок*); 3) 'вказує на

опредметнену дію, названу мотивувальною основою’ (*стрибок, ривок*); 4) ’вказує на предмет як результат дії, названої мотивувальною основою’ (*моток, обрубок*); 5) ’вказує на особу за виконуваною дією, названою мотивувальною основою’ (*ходок, стрілок, недоросток*).

Морфемна структура словоформи не завжди пояснюється деривaційними процесами творення цієї словоформи. Наприклад, у словах *доглядальниця, залицяльниця* на морфемному рівні визначається суфіксальна послідовність –а-льн-иц-. Оскільки в мові не зафіксовано таких слів, як *доглядальний* та *залицяльний*, словотвірний аналіз визначає формант –льниц-, на базі словотвірних пар *доглядати – доглядальниця*, *залицятися – залицяльниця*. Визначення морфів –льн- та –иц- здійснюється на основі системно-структурних відношень цих знакових сегментів. Оскільки в мові існують такі слова, як *регулювальниця, перевіряльниця*, утворені від прикметників *регулювальний, перевіряльний*, то слова *доглядальний і залицяльний* розглядаються як потенційно можливі, а морфемне членування слів *доглядальниця, залицяльниця* здійснюється за принципами дистрибуції аналогічної сегментації та орієнтації на інвентар афіксів, уже встановлений у мові.

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 в 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 в синхронічному словотворі ці слова є непохідними.

Морфеми -льн-, -иц- та -ок- визначаються не за словотвірними відношеннями, а на основі дистрибутивних та парадигматичних відношень цих морфем у системі морфеміки в цілому, що дозволяє використати метод аналогій у морфемній сегментації слів. Метод аналогій у морфемному аналізі ґрунтується на загальносистемній мотивованості морфемних структур. Морфемна структура є константною, і вона відображає ті морфологічні процеси, які спрацьовують у системі мови в цілому. Тому морфемна структура, конкретно взятої словоформи, може відображати не тільки актуальні морфологічні процеси, а й діахронічні, і ті, що потенційно можливі. Отже, морфемна структура словоформи може розглядатися як стала структура, що не конструюється, а переймається (успадковується) новоутворенням від свого аналогічного зразка.

Морфологічна структура словоформ може бути описана як результат різних морфологічних процесів, що відбуваються в мові. Мета морфологічного аналізу визначити природу цих процесів: зіставити кожен сегмент словоформи з тим морфологічним процесом, який породжує й констатує цей сегмент. Для визначення формотвірної та словотвірної структури словоформи необхідно проаналізувати морфологічні процеси, що

пояснюють динаміку породження конкретної словоформи. Визначені в результаті двох типів аналізу сегменти функціонально значущі для морфологічних процесів словозміни і словотвору, які й зумовили утворення конкретної словоформи, що аналізується. На формотвірному рівні організації морфологічної структури словоформи визначають двоелементну будову: формотвірний елемент – словозмінний або формотвірний афікс, та вторинну константну одиницю – основу, яка є предметом словотвірного і морфемного аналізу.

У морфемному аналізі основи слова визначають два підходи:

1) коренезорієнтований, а по суті, словотвірнозорієнтований [Винокур 1946], [Шанский 1959], за яким подільність основи слова на морфеми враховує дистрибуцію кореневої морфеми в спільнокореневих словах, відповідно, якщо немає двох слів з одним і тим же вільним або зв'язаним коренем, то основа вважається неподільною;

2) суфіксозорієнтований [Смирницкий 1948], [Земская 1973], [Яценко 1980], за яким подільність основи слова на морфеми враховує дистрибуцію і кореня в спільнокореневих словах, і афікса в спільноструктурних словах, відповідно, якщо немає двох слів з одним і тим же вільним або зв'язаним коренем, але є слова, у яких повторюється афікс, то корінь визначається за надлишковим принципом, а основа вважається подільною.

Другий підхід І. Яценко визначає релевантним для морфемного аналізу, адже він « … визначає рівноправність усіх морфем – і кореневих, і службових – як структурних частин слова і як носіїв своїх певних значень. Отже, науковий підхід до розуміння морфемного складу слова зобов'язує брати до уваги співвідношення слів не лише за кореневими морфемами, а й за службовими» [Яценко 1980: 11].

Морфема – знакова одиниця морфемного рівня мовної системи і її видільність визначається повторюваністю у структурах різних слів і відтворюваністю у свідомості мовців. Ці дві ознаки – повторюваність і відтворюваність – відображають два типи структурних відношень морфем у морфемній системі мови: синтагматичні відношення – здатність морфа сполучатися з іншими морфами у структурі слова; парадигматичні – здатність морфеми асоціюватися з певним квантом значення, цілісною формою морфеми у свідомості носія мови й на цій підставі відтворюватися як цілісна одиниця (морфема) та протиставлятися іншим морфемам. У синтагматиці морф – це структурний компонент слова, що вступає із цим словом в ієрархічні відношення: морфи, поєднуючись один з одним за правилами дистрибуції, утворюють конструкт нової якості – одиницю вищу за рангом – слово. Морф не існує поза словом, а тільки в ньому, і виконує щодо слова конститутивну функцію. Тому основним принципом морфемного членування слова на морфи (морфеми) є принцип додаткової дистрибуції: якщо структурний сегмент слова асоціюється з одним і тим же значенням і повторюється в різному оточенні інших структурних сегментів, то такий сегмент є морфом (морфемою).

Морфемна сегментація слова за принципом додаткової дистрибуції спрямована на визначення морфемної структури слова безвідносно до типів морфем – кореневих чи афіксальних, її завдання вичленувати мінімальні за формою значущі (системно протиставлені) елементи структури слова. Наступним завданням морфемного аналізу є встановлення наявності значення визначеного сегмента. Причому значення морфеми встановлюється не тільки у конкретно взятому слові, а в системі морфеміки загалом, тому що в одних словах морфи однієї морфеми можуть виражати значення, а в інших словах нівелюють ці значення, але це не означає, що морфема не має значення. Морфема як інваріант, тобто як парадигматична одиниця, що реалізовується в мовленні в конкретних варіантах – морфах, може варіювати в плані змісту і плані вираження. Одні морфи зазнають фонемних змін – варіативність форми, інші семантичних змін – варіативність значення, або суміщують варіативність двох планів мовного знака.

Морфемний анатомізм слова зумовлює членування його на морфи в плані семантики і форми, проте семантична структура слова не завжди релевантна формальній структурі, що пояснюється асиметричним дуалізмом морфеми як знакової одиниці мови: є морфеми, які виражають декілька значень, притаманних слову, наприклад флексії, які в українській мові є кумулятивними, або, навпаки, є морфеми, які не актуалізують значення в семантичній структурі конкретного слова. Такі елементи для семантичної структури слова є надлишковими компонентами, проте слово як цілісна знакова одиниця без них не існує: якщо ці морфеми забрати зі слова, то слово-знак руйнується. Отже, такі компоненти слова є значущими у системі морфеміки загалом і в структурі конкретного слова, де вони виконуються основну функцію морфеми – конститутивну. Крім того, у системі мови ці морфи асоціюються із певним значенням, оскільки інші морфи цієї самої морфеми актуалізують значення морфеми в інших словах. Таким чином значення морфеми встановлюється як системномовне, парадигматичне, а не синтагматичне, причому еталоном у встановленні морфемного значення виступає морф у сильній морфемній позиції – фінальній позиції морфемної структури основи слова. Реалізуючись у цій позиції у похідних словах морф виконує словотвірну функцію – виступає словотвірним формантом мотивованого слова, тому значення морфеми в більшості теорій визначається як словотвірне. Однак значення морфеми і словотвірне значення базуються на семантиці різних процесів: словотворення і конструювання слова. Значення морфеми має більш узагальнений і абстрагований характер, порівняно із словотвірним значенням, оскільки значення морфеми узагальнює всі морфи системи мови в одну інваріантну одиницю – морфему, безвідносно до відношень словотвірної похідності конкретних слів, чи слів одного словотвірного типу.

«Морфемний різновид морфологічного аналізу є найбільш послідовною реалізацією принципу сегментації мовної форми – сегменти морфемної структури словоформи можуть бути охарактеризовані як її кінцеві

складники. Для здійснення такого членування необхідно задіяти для зіставлення максимально широке, практично необмежене коло словоформ. Це пов’язано з тим, що на рівні кінцевих складників морфологічна структура слова є наслідком не тільки процесів слово- і формотворення, а також і результатом аналогічного впливу вже існуючих морфологічних структур» [Богданов 1997: 58].

Метод аналогій у морфемному аналізі ґрунтується на загальносистемній мотивованості морфемних структур, що на рівні загальносистемних закономірностей відображають формо- і словотвірні процеси на синхронному зрізі. Морфемна структура є константною, і вона конденсує, відображає ті морфологічні процеси, які спрацьовують у системі мови в цілому. Морфемна структура словоформи може розглядатися у двох аспектах:

1) як константна структура, що не конструюється, а "переноситься" цілісно новоутворенням від свого аналогічного зразка;

2) як «застигла в морфемах словотвірна генеалогія слова в його живих для мовної свідомості ланках» [Моисеев 1968: 7].

Морфемний аналіз, який керується принципом повноти лінгвістичного аналізу, повинен розмежовувати ці два аспекти й проходити у два етапи:

1) сегментація словоформи на основі методу додаткової дистрибуції та методу аналогій, що дозволяє визначити межі морфем у структурі словоформи;

2) визначення функціональної структури на морфемному рівні організації словоформи через актуалізацію деривaційних та формотвірних процесів, що дозволяє проаналізувати функціональні типи морфем, уже встановленої на першому етапі аналізу, морфемної структури слова.

Принцип розмежування двох типів аналізу морфемного та словотвірного як процедур із різними одиницями (морфема – твірна основа та словотвірний формант), різною метою (визначення морфемної будови слова – визначення способу творення слів), різними методами (методи додаткової дистрибуції та аналогій – метод словотвірної похідності) покладено в основу комп’ютерного моделювання в АСМСА.

На інфологічному етапі моделювання було створено концептуальну модель системи АСМСА, що розмежовує два типи морфологічного аналізу й структурується на два модулі та чотири блоки:

1) морфемний модуль – морфемна база даних, де аналізується морфемна структура словоформ;

2) словотвірний модуль – словотвірна база даних, де аналізується словотвірна структура словоформ.

Кожен модуль складається з двох блоків: 1) блок-словник; 2) блок-аналізатор. Структуру АСМСА можна представити у вигляді такої блок-схеми:

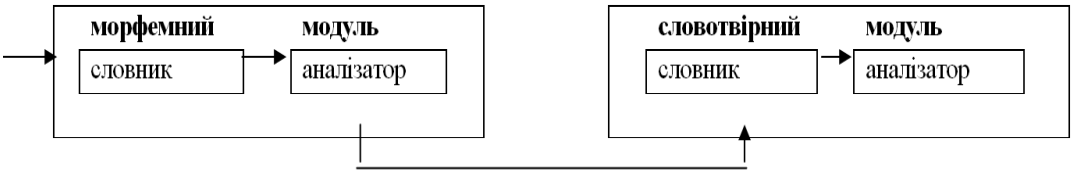


Рис.1.9. Блок-схема концептуальної моделі АСМСА

Передбачається така логіка виконання завдань: (1) укладання словника морфемного модуля – реєстру слів із визначеною морфемною будовою → на базі цього словника створюється морфемний аналізатор (2), що виконує функцію лінгвістичного класифікатора в процесі укладання спільнокореневих вибірок → словника словотвірного модуля (3) → на базі цього словника створюється система автоматизованого словотвірного аналізу (4). Така послідовність завдань зумовлена робочою гіпотезою: побудова словотвірного гнізда як статті електронного словотвірного словника здійснюється на базі вибірки всіх спільнокореневих слів мови. Тому на першому етапі необхідно було укласти таку базу даних, яка б забезпечувала автоматичне групування однокореневих слів на масиві найповнішого реєстру лексики української мови.

1.4.2. Даталогічна модель морфемної структури слова в АСМСА

Робота над створенням АСМСА розпочалася з формування лексичного реєстру обсягом ≈ 170 тис. початкових форм слів української мови, представлених у графемному записі з граматичною індексацією частин мови. Графемний запис слів був автоматично конвертований у спрощену фонематичну транскрипцію. Алгоритм перетрансляції графемного запису у спрощений фонематичний ураховує тільки позиції йотованих я, ю, є, ї: здійснює перезапис одного символа у два символи: я → ja, ю → jy, є → je, ї → ji. Ця процедура є обов'язковою, тому що вона дозволяє правильно визначити межі морфів у тих випадках, коли одна буква представляє дві фонеми, які належать до різних морфем: *клеїти → клej-і-ти; блукаючий → блук-aj-уч-ий.* Усі інші особливості фонемного запису не беруться до уваги, зокрема м'які фонеми (зберігається графічне позначення м'якого знака: *сіль*) та наголос, тому що АСМСА будувалась за концепцією текстоорієнтованої системи з перспективою автоматичного морфемного аналізу слововживань українськомовного тексту в графемному представленні.

У процедурі перекодування символів принцип ізоморфності двох моделей (фонемної та графемної) мовної субстанції не порушується, тому що «... писемна форма мови моделює її усну форму, тому взагалі справедлива послідовність: <модель дійсності – мислення> → <модель мислення – усна мова> → <модель усної мови – писемна мова>. Оскільки наведені моделі фізично реалізуються в єдиній системі (пов'язаній з індивідуумами, соціальними спільнотами, системами культури тощо), цілком природні та

необхідні їх взаємодія та взаємовплив. Таким чином, писемний варіант мови так само виступає у функції і моделі мислення, і моделі дійсності» [Широков 2011].

Першочерговим завданням було створення на основі сформованого лексичного реєстру морфемної бази даних, у якій у кожному слові буде визначено морфемну структуру у формі такого запису, який би дозволив комп'ютеру автоматично просегментувати слово на морфеми, визначити функціональний тип просегментованих морфем і здійснювати різноманітні класифікації морфем та слів за заданою морфемною ознакою (спільний корінь, спільний афікс, спільна морфемна структура та ін.). Морфемна модель слова повинна кодувати інформацію про межі й тип кожної морфеми в слові, причому запис цієї моделі повинен бути формалізований на такому рівні, що уможливлює програмування процедури автоматичної морфемної сегментації слів у нових лексичних реєстрах, укладених за текстовими вибірками.

Морфемна база даних укладалася автоматизовано: лінгвіст проводив морфемний аналіз слів сформованого лексичного реєстру, використовуючи процедурний і символьний апарат спеціально створеної комп'ютерної програми, за допомогою якої кодував визначену морфемну структуру в символах розробленої даталогічної моделі (автоматизована система укладання МБД описана у 2-му розділі).

У моделюванні морфемної структури слова було використано формалізовані принципи опису структурних відношень морфів на двох площинах організації слова як мовного знака: формалізація плану вираження та плану змісту морфемної структури слова, визначеної лінгвістом. У формалізації плану змісту морфемної структури слова було використано традиційну статичну символьну модель функціональних типів морфем, де латинськими літерами позначається функціональний тип морфеми: P – префікс, R – корінь, S – суфікс, F – флексія, I – інтерфікс, X – постфікс, наприклад, *анти-об-лід-н-ювa-ч-ø* PPRSSSF. Проте приписування такої моделі до слова не забезпечує кодування процесу автоматичної сегментації слова на морфеми: у програмному забезпеченні необхідно задати умову не тільки про функціональний тип кожної визначеної морфеми в слові, а й чітко визначити межі цієї морфеми у слові.

В основу створення даталогічної моделі морфемної структури слова було покладено розуміння слова в комп'ютерній лінгвістиці: слово є послідовністю графемних символів від пробілу до пробілу. В автоматичному лінгвістичному аналізі комп'ютер працює з одноплановою одиницею – ланцюжком графем, кожна з яких займає свою порядкову позицію: а1н2т3и4о5б6л7і8д9н10ю11в12а13ч14 (*антиобліднювач*). Відповідно, встановлювані лінгвістом межі морфем проходять між порядковими позиціями графем: анти4об6лід9н10юва13ч14*ø (анти-об-лід-н-юва-ч-ø)*. Ця закономірність була покладена в основу моделювання плану вираження морфемної структури слова. Функціональну організацію слова на

морфемному рівні представляє функціональна модель – PPRSSSF, а субстанціальне вираження цієї структури представлено через кількісно-графемну модель (анти-4; об-6; лід-9; н-10; ова-13; ч-14; ø-14). Позиція нульового афікса (суфікса або флексії) кодується цифрою, що визначає межу попередньої морфеми. Зіставлення і поєднання двох типів моделей (функціональної і кількісно-графемної) відбувається автоматично, і в результаті формується функціонально-кількісно-графемна даталогічна модель морфемної структури слова: *антиобліднювач* P4P6R9S10S13S14F14. У цій моделі кожна морфема представлена двосимвольним кодом, де перший символ – графема – визначає функціональний тип морфеми, а другий символ – цифра – місце морфемного шва за порядковим номером графеми в слові: анти – P4; об –P6; лід – R9; н – S10; ова – S13; ч – S14; *ø* – F14. Графемно-цифрові межі морфем за вимогами програмного забезпечення були перекодовані в знаки однієї символьної системи: подані через латинську літеру за порядковим номером у латинському алфавіті (за винятком деяких літер):

P4P6R9S10S13S14F = PE(4)PG(6)RJ(9)SK(10)SN(13)SO(14)FO = PEPGRJSKSNSOFO.

У такий спосіб кожному слову автоматизовано приписується модель, яка виступає робочою одиницею морфемної бази даних АСМСА (див. Розд. 2). Резидентний словник МБД систематизує інформацію за чотирма зонами:

антиобліднювач,Й,PEPGRJSKSNSOFO/лід1/
лід,Й,RDFD/лід1/
лідер,Й,RDSFFF/лід2/
лідерство,Л,RDSFSIFJ/лід2/
лідерський,А,RDSFSIFK/лід2/
підлідний,А,PDRGSHFJ/лід1/
підлідник,Й,PDRGSHSJFK/лід1/

1) графемний запис слова: антиобліднювач; 2) граматичний код частини мови: Й; 3) програмна процедура морфемної сегментації: PEPGRJSKSNSOFO; 4) ідентифікатор кореневої морфеми: /лід1/ (функція четвертої зони буде описана в 2-му розділі).

Даталогічна функціонально-кількісно-графемна модель морфемної структури слова, зіставлена з графемним записом кожного слова у МБД, дає необхідну лінгвістичну інформацію про організацію морфемної структури слова для проведення автоматичного морфемного сегментування й визначається як програмна процедура морфемної сегментації. Комп'ютерна програма морфемної сегментації слова проводить операції над ланцюжком графем за двосимвольними кодами даталогічної моделі: перший символ визначає функціональний тип морфеми, який встановлюється через процедуру приписування символу; другий символ визначає межу морфеми, яка встановлюється через процедуру підрахунку графем із початку слова, і відділення від слова графемного сегмента за встановленою кількісно-графемною межею. Нульовий афікс, який закодований однаковою з попередньою морфемою кількісно-графемною позицією, визначається через

процедуру приписування символа – ø, наприклад, гра RCSCFD: гр - RC; ø-SC; а-FD.

1.5. Моделювання лексикографічного опису морфемної системи української мови

1.5.1 Концептуальна лексикографічна модель електронного морфемного словника

Використання АСМСА в автоматичному морфемному аналізі початкових форм (лем) українськомовного тексту зумовило формулювання нового лексикографічного завдання: створити інтерактивні електронні морфемні словники за текстовими вибірками.

Структура електронних морфемних словників визначається завданнями цих словників та теоретичними засадами концептуальної лексикографічної моделі, основною вимогою якої є створення повного й цілісного опису морфемної системи української мови. Тому одиниці і принципи лексикографічного опису повинні визначатися системно-структурними закономірностями рівневої організації морфеміки та завданнями морфемології.

Усталеним у сучасному мовознавстві, як зазначалось у першому параграфі, є теоретичне визначення морфеми як інваріантно-варіантної знакової одиниці. Розмежування морфеми як синтагматичного та парадигматичного явищ, продукує два теоретичні поняття: "морф" і "морфема". Для формування концептуальної лексикографічної моделі цей методологічний підхід є надзвичайно важливим, тому на інфологічному етапі необхідно дотримуватися термінологічного розрізнення цих понять. Морфема – інваріант, парадигматична одиниця мови, що реалізується в конкретних варіантах – морфах. Морф – мінімальна лінійна знакова одиниця мови, яка є елементом слова й не має синтаксичної самостійності. У системі мови морф вступає в синтагматичні, парадигматичні та ієрархічні відношення: зі словоформами як одиницями вищого рівня і фонемами як одиницями нижчого рівня мови. Ці відношення формують структуру морфемної системи мови.

На "рівні аналізу" у межах морфемології визначають два розділи:

1) морфотактику, що вивчає «правила формальної, семантичної та стилістичної сполучуваності морфем» [Клименко 1998а: 7];

2) морфонологію, «що вивчає звукові зміни в морфемах, зумовлені їхньою сполучуваністю, а також наголосові ознаки та фонемну структуру морфем» [Клименко 1998а: 7].

Відповідно, об’єктами вивчення у морфемології є не лише морфи, а й морфемні структури слів та морфеми як парадигматичні одиниці мови.

Актуалізація синтагматичних відношень між морфами на обох рівнях організації структури мовного знака приводить до утворення одиниці вищого

рівня – словоформи й, відповідно, репрезентує лінгвістичний вид відношень за ступенем складності – ієрархічні відношення: відношення входження менш складних одиниць (морфів) у більш складні одиниці (словоформи). Основою ієрархічних відношень виступає поняття "комбінаторики" морфів, що визначає тематичне і формальне об'єднання морфів у слово як знакову одиницю вищого рангу. «Здатність до комбінаторики – загальна й обов'язкова властивість одиниць мови, зумовлена загальносистемними й фундаментальними властивостями одиниць мови – дискретністю і неоднорідністю. Ієрархічність і лінійність також належать до фундаментальних властивостей мовних одиниць, які зумовлюють спосіб реалізації комбінаторики» [Солнцев: 1971, 267].

Опис синтагматичних та ієрархічних відношень (між морфами і словоформою) можна здійснити через встановлення дистрибутивно-комбінаторної характеристики морфів на формальному та функціональному рівнях організації морфемних структур слів. Системний дистрибутивний опис морфів ставиться в центрі уваги українських та зарубіжних лінгвістичних досліджень: Н. Клименко та Є. Карпіловської [Клименко 1996], [Клименко 1998], [Клименко 1998а], [Карпіловська 1999], [САМУК 1998]; Я. Горецького [Horecki 1964]; Т. Єфремової [Ефремова 1968], [Ефремова 1970]; Р. Зятковської [Зятковская 1980]; Х. Ковалик [GWJH 1984]; О. Зубань [Зубань 1997б], [Зубань 1998], [Зубань 2000]), у яких описується валентність морфів, їх позиційні характеристики, що зумовлює визначення комплексної одиниці морфемної системи мови – морфної / морфемної структури слова. Типовий характер морфемних структур, регулярна повторюваність і відтворюваність у словах визначає їх онтологічними одиницями структурного характеру морфемного рівня мовної системи. Опис морфструктур здійснюється за допомогою символьних статичних моделей. Лексикографічний досвід такого опису репрезентує «Словник афіксальних морфем української мови» [САМУК 1998], у якому кожен афіксальний морф описаний як окрема синтагматична одиниця в структурних відношеннях із пре- і постпозитивними морфами.

Лексикографічний опис у межах морфотактики ставить завдання описати дистрибуцію кожного морфа, визначивши його валентність у морфемних структурах усіх спільнокореневих чи спільноафіксальних слів. Концепція МБД АСМСА спрямована на виконання цього лексикографічного завдання: формалізоване представлення інформації про морфемну структуру слова в базі даних забезпечує автоматичне формування: 1) реєстру функціональних моделей морфемних структур слів та спільноструктурних класів слів, що описуються певною моделлю; 2) реєстрів префіксів, коренів, суфіксів, інтерфіксів, постфіксів та флексій (система флексій формується обмежено, тільки на матеріалі початкових форм). Одиниці двох реєстрів представляють систему елементів морфеміки української мови.

Визначення комбінаторно-дистрибутивної характеристики морфів дозволяє провести ідентифікацію морфів однієї морфеми й створити

лексикографічний опис морфемної системи мови у термінах морфемних парадигм. Методологічні принципи дослідження морфеми як парадигматичної одиниці, описані у роботах О. Зубань [Зубань 1997], [Зубань 1997а], [Зубань 2001], були покладені в основу моделювання морфеми як інваріантно-варіантної одиниці в описуваній лексикографічній моделі.

Морфема як клас морфів, у яких вона реалізовується в синтагматиці, розглядається в парадигматиці як інварантно-варіантна парадигматична одиниця мови. Інваріантність і варіативність – дві взаємовизначальні характеристики мовної одиниці. Парадигматична цілісність морфеми базується на цих двох ознаках, тому в лексикографічному описі морфема буде представлена як інваріантно-варіантна модель, що передбачає групування в один клас усіх морфів однієї морфеми з урахуванням усіх слів, у яких ці морфи реалізовані.

Явище варіативності одиниці мови зумовлене двома факторами: а) існуванням кожної одиниці у вигляді деякого класу; б) використанням у мовленні завжди одного представника класу. Процедура ідентифікації морфів однієї морфеми враховує інваріантні ознаки морфеми: до однієї морфеми належать морфи, що мають тотожні значення та повну або часткову фонологічну (графемну) подібність плану вираження. Інваріантні властивості морфів (тотожність значення і часткова фонологічна (графемна) подібність плану вираження) виступають інтегральною основою моделювання парадигматичного конструкта мовної системи – морфеми, у якій узагальнюються варіативні риси морфів. Варіативність стосується тільки плану вираження як часткова фонологічна (графемна) відмінність морфів однієї морфеми. Відмінність змісту морфів розглядається не як варіативність, а як актуалізація потенційно закладених у морфемі змістів – інформативність морфеми. Інформативність морфеми – міра змісту одиниці в конкретній реалізації [Гапьперин 1974]. У морфемі інформація виступає лише як потенційно можлива реалізація всього змісту, закладеного в ній. У кожному новому оточенні морфема може змінювати свою форму (реалізуватися в аломорфах), а також розширювати або нівелювати своє значення, яке актуалізується не тільки окремою морфемою, а й всією морфною структурою слова, до якої вона входить. Тому обов’язковим елементом значення морфеми, описуваної в словнику, повинна виступати інформація про структурнопозиційну сполучуваність морфів і про динаміку функціонування значення морфеми в кожній конкретній реалізації.

Визначення значення інваріантної одиниці – морфеми – є найскладнішим завданням. У сучасній лінгвістичній семантиці ставиться завдання описати значення основних мовних знаків: морфем, слів, речень, тестів, проте семантиці морфеми присвячено найменше досліджень. Найбільш глибоко описані словотвірні значення деривaційних та граматичні значення реляційних афіксів, найменше – кореневих морфем. В українській морфемній лексикографії значення афіксальних морфем описані в таких словниках: «Словник українських морфем» [Полюга 2009], «Етимологічний

словник запозичених суфіксів і суфіксоїдів в українській мові» [Селігей 2014], «Словник афіксальних морфем української мови» [САМУК 1998]. Прикладом таких словників у російській лексикографії є «Толковый словарь словообразовательных единиц русского языка» [Ефремова 1996] та «Новый словарь русского языка. Толково-словообразовательный» [Ефремова 2001].

Морфема – це особливий мовний знак, тому що він не зіставляється безпосередньо з референтом (явищем дійсності). Морфема актуалізує своє значення тільки в складі слова, яке виконує номінативну або дейктичну референцію. Тому в мовознавстві побутує думка про те, що морфема має асоціативне значення. Значенню морфеми не можна дати єдиного трактування, тому що кореневі морфеми, деривaційні афіксальні морфеми, реляційні афіксальні морфеми, субморфеми виражають різні типи значень.

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 спільнокореневих 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 МБД АСМСА ставилося завдання не описати значення кореня, а протиставити за значенням корені-омоніми. Такий методологічний підхід ґрунтується на теоретичному постулаті Фердинанда де Соссюра про поняття значущості елемента мовної системи. Розуміючи семантику морфеми, як таку, що має не значення, а значущість, О. Герд [Герд 1983] визначає морфему напівзнаком. У нашому дослідженні це теоретичне положення набуло такого процедурного формулювання: якщо два морфи з однаковим планом вираження відрізняються за семантикою, то ці морфи належать до різних морфем. При цьому значення морфем не описувалось, але в семантичному зіставленні коренів бралась до уваги семантика непохідних слів, у яких кореневі морфи реалізовані[15]. Безумовно, таке визначення значення кореня не може претендувати на семантичний опис, тому що 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 послуговуватися тільки тлумачними словниками, тому приписування кореневій морфемі значення непохідного слова, у якому реалізований один із її морфів, на цьому етапі вивчення морфемної семантики є, на нашу думку, єдиним способом лексикографічного опису семантики кореневих морфем.

Семантика афіксальних морфем традиційно описується як граматичне значення (для формотвірних та реляційних афіксів) або словотвірне значення, що встановлюється через функцію словотвірного форманта у процесі

[15] До кожного слова у МБД подано його тлумачення за словниками української мови.

словотворення або функцію кваліфікативного форманта у непохідних подільних основах *(скрип>к-а)*.

Словотвірне значення афікса встановлюється на базі семантичних відношень між мотивувальною та мотивованою основами слів. Визначення словотвірного значення афікса зумовлюється двома підходами у трактуванні цього поняття. І. Улуханов [Улуханов 1996] визначає словотвірне значення як елемент значення похідного слова, яким воно відрізняється від твірного: словотвірне значення – різниця між мотивованим і мотивувальним у кожній конкретній словотвірній парі. В. Ґрещук [Ґрещук 1995], О. Земська [Земская 1973], Н. Клименко [Клименко 1984], І. І. Ковалик [Ковалик 1971], В. Лопатін [Лопатин 1977] визначають словотвірне значення як загальне категорійне значення слів, об'єднаних в один словотвірний тип: через семантичне зіставлення всіх мотивувальних і похідних одного словотвірного типу, абстрагуючись від конкретних лексичних значень слів цього словотвірного типу. І широке, і вузьке трактування словотвірного значення оперують бінарною перифразою, яка описує зв'язок значення словотвірного форманта (морфеми, декількох морфем) зі значенням твірної основи.

Значення морфеми відрізняється від значення словотвірного форманта, воно встановлюється не тільки в конкретному слові чи словотвірному типі, а в системі морфеміки загалом, тому що в одних словах морфи однієї морфеми можуть виражати значення, а в інших словах нівелювати ці значення, але це не означає, що морфема не має значення. Морфема, як інваріант (парадигматична одиниця), реалізуючись у мовленні в конкретних варіантах – морфах, може змінюватися не тільки в плані вираження, а й у плані змісту.

Морфемний анатомізм слова зумовлює членування його на морфи в плані змісту і плані вираження, проте семантична структура слова не завжди релевантна формальній структурі, що пояснюється асиметричним дуалізмом морфеми як знакової одиниці мови. Є морфи, які виражають декілька значень, наприклад кумулятивні флексії. Є морфи, які не актуалізують свого значення в семантичній структурі конкретного слова: вони для семантичної структури слова є надлишковими семантичними компонентами або формальними семантично спустошеними компонентами. Такі морфи є значущими для формальної структури конкретного слова, де вони виконують основну функцію морфеми – конститутивну. Крім того, ці морфи є значущими й для системи морфеміки загалом, тому що, як варіанти морфеми, ці морфи асоціюються із певним значенням, яке виражають інші морфи цієї самої морфеми в морфемних структурах інших слів.

Значення морфеми встановлюється як системномовне, парадигматичне, а не синтагматичне, причому еталоном у встановленні морфемного значення афіксальної морфеми виступає морф у сильній морфемній позиції: для суфіксів у фінальній позиції морфемної структури основи слова, для префіксів у початковій позиції морфемної структури основи слова. Реалізуючись у цій позиції в похідних словах, афіксальний морф виконує словотвірну функцію – виступає словотвірним формантом мотивованого

слова, тому значення афіксальної морфеми в більшості теорій визначається як словотвірне. Однак значення афіксальної морфеми і словотвірне значення базуються на семантиці різних процесів: словотворення і конструювання слова. Значення афіксальної морфеми має більш узагальнений і абстрагований характер, порівняно із словотвірним значенням, тому що значення морфеми узагальнює всі морфи в одну інваріантну одиницю безвідносно до відношень словотвірної похідності конкретних слів, чи слів одного словотвірного типу. Значення афіксальної морфеми узагальнює семантику всіх слів, морфемна структура яких містить цей афікс, тобто виражає семантику якоїсь семантичної категорії. Саме такий опис семантики суфіксальних морфем представлено у «Словнику афіксальних морфем української мови» [САМУК 1998] як індекс значень омонімічних суфіксальних одиниць. Ці значення було використано в описі семантики суфіксальних одиниць у МБД АСМСА і, на наше переконання, можуть бути використані як інваріантні значення афіксальних морфем у лексикографічній моделі морфеми.

Інваріантно-варіантна лексикографічна модель морфеми передбачає опис формальних варіантних ознак морфеми: часткову фонологічну (графемну, якщо слово подане в графемному записі) відмінність плану вираження морфів, що зумовлює опис варіативності морфеми в термінах аломорфів (морфів, що знаходяться у відношенні додаткової дистрибуції) або варіативних морфів (морфів, що знаходяться у відношенні вільного варіювання). Відповідно, модель морфемної парадигми може бути представлена: 1) тотожними за формою вираження морфами; 2) аломорфами та варіоморфами. Значущі (аломорфічні) варіювання морфів характеризуються морфонологічними правилами: альтернаційний ряд фіксує фонологічну відмінність морфів однієї морфеми, якій у семантичному плані знакової одиниці відповідає відмінність структурно-граматична або структурно-лексична. «Ми вважаємо, що правило може вважатися морфонологічним тільки в тому випадку, коли воно не зумовлюється властивостями звуків і їх комбінаторикою, а також, якщо встановлене чергування, фонологічно протиставляючи два морфи однієї морфеми, одночасно відіграє роль [...] функціонального засобу й закріплене за визначеною позицією» [Кубрякова 1983: 23].

Морфонологічний аналіз спрямований не у сферу синтагматики, а у сферу парадигматики мови: формально-функціональне об’єднання варіантів морфеми утворює визначену парадигму. Альтернаційний ряд фіксує фонологічну (графемну) відмінність морфів однієї морфеми, якій у семантичному аспекті відповідає відмінність структурно-граматичної або структурно-лексичної інформації.

Морфонологічна характеристика морфеми визначає:

1) парадигматичний тип морфеми:

- аломорфічна морфема – всі морфи парадигми знаходяться у відношенні додаткової дистрибуції;

- аломорфічно-варіативна морфема – морфи в одному парадигматичному протиставленні характеризуються додатковою дистрибуцією, а в іншому знаходяться у відношенні вільного варіювання;
- варіативна морфема – морфи в межах парадигми вступають тільки у відношення вільного варіювання;

2) характер морфонологічного варіювання морфеми у синхронії:

- консонантна морфема реалізується в одному морфі або за допомогою декількох морфів, варіативність яких мотивується фонологічними причинами: асимілятивним чергуванням, накладанням, усіченням, інтерфіксацією;
- альтернуюча морфема реалізується в морфах, варіативність яких пояснюється морфологічними причинами: зміною морфологічної позиції у слові.

Морфонологічні варіювання описуються за допомогою морфонологічних альтернаційних рядів, наприклад, а[к-ц'], де: /а/ – стала фонема, [к-ц'] – фонеми, які чергуються (*/вој>ак-а/ – /вој>ац'-і/).* Морфонологічні альтернації виконують функцію структурно-граматичної диференціації морфів: 1) у межах однієї словозмінної парадигми такі альтернації свідчать про граматичну семантизацію афікса; 2) у протиставленні граматичних основ різних частин мови вказують на структурно-позиційні характеристики морфів, наприклад, альтернаційний ряд а[ц'-ч] *(/шв>ач-ø/ – / шв>ац'-к-иј)* свідчать про здатність морфеми реалізовуватися в морфемних структурах відносних прикметників у аломорфі /-ац'-/ .

Зміна структурно-позиційної сполучуваності морфа впливає на лексичну або граматичну інформативність морфеми. Лексична інформативність визначається через можливість / неможливість визначити сему афікса в семантичній структурі словоформи. У певних структурних позиціях морф може нівелювати своє значення (*/риб>ач-к-а/* – морф /-ач-/ втрачає значення ’особа чоловічої статі’) або модифікувати своє значення.

Граматична інформативність морфа визначається через протиставлення аломорфів однієї морфеми у межах морфологічної парадигми однієї частини мови. Така семантизація може зумовлюватися морфемно-структурними модифікаціями двох типів, наприклад:

1) афікс /-ац'-/ (*/вой>ак-а/ – /вой>ац'-і/)* набуває граматичної інформації (значення давального та місцевого відмінків) під впливом контактної постпозитивної флексії. У такому випадку граматична інформація передається в морфемній структурі словоформи двома морфами: морфом-донором /-і/ та морфом-рецепієнтом /-ац'-/, причому форма вираження останнього обов’язково варіює /-а[к-ц']-/;

2) морфи /-ач-/ та /-ч-/ (*хиж>ач-и-ти/ – /хиж>ач-у/*; */жебр>а-ч-и-ти/ – /жебр>а-ч-те/*) набувають граматичної інформації – семи ’процесуальна ознака’ за рахунок утрати основою деяких дієслівних словоформ

теперішнього часу та словоформ наказового способу 2-ої дієвідміни дієслівного словотвірного суфікса.

Лексикографічний опис морфеміки в термінах морфемних парадигм спрямований на вирішення морфонологічних завдань, поставлених М. Трубецьким [Трубецкой 1967]. Для вивчення парадигматичного варіювання морфеми необхідно встановити фонологічну структуру морфеми – це обов'язковий етап у процедурі ідентифікації морфів, який відповідає першому завданню морфонологічної теорії. Два наступних завдання морфонологічної програми М. Трубецького: встановлення комбінаторних чергувань та встановлення звукових (фонемних) чергувань, що виконують морфологічну функцію, логічно виходять із першого завдання й визначають характер варіювання морфеми. У морфемній парадигматиці ці завдання не повинні протиставлятися[16], тому що комбінаторні зміни також виступають засобом передачі морфологічної (структурної) інформації про здатність морфа сполучатися з визначеними морфами у пре- чи постпозиції.

Враховуючи інваріантні та варіативні властивості морфеми як дві характеристики одного явища, лексикографічна модель повинна представити морфему як лінгвістичну комплексну одиницю й описати структурно-системні відношення цієї одиниці в мові:

- синтагматичні – сполучуваність морфів у пре- та постпозиції;
- парадигматичні – протиставлення морфів за формою та функцією в межах однієї морфеми та в морфемній системі мови в цілому;
- ієрархічні: 1) відношення між фонемами (графемами), що репрезентують форму вираження морфів однієї морфеми, та морфемою як інваріантом; 2) відношення між словоформами і морфами, що в системі словоформи можуть набувати різної інформативності.

З опертям на визначені завдання лексикографічна модель морфеми повинна представляти:

1) альтернаційний ряд, що відображає варіативність форми вираження морфеми;

2) повний інвентар морфів описуваної морфеми, зіставлених із тією позицією, яку кожен морф займає в морфемній структурі слів (структурна інформативність), та з тим граматичним чи лексичним квантом значення, яких морф набуває в цій позиції (граматична або лексична інформативність).

Використання методологічних принципів морфонологічного аналізу в укладанні морфемного словника забезпечує системний опис функцій морфеми в синтагматичній реалізації, а також відображає варіювання морфів у процесі конструювання морфемних структур слів, що підтверджується дослідженням морфемних парадигм суфіксальних афіксів дієслівних основ, проведеним О. Зубань [Зубань 1997], [Зубань 1997а].

---

[16] Є. Клобуков [Клобуков 1976] зауважує, що комбінаторні чергування не входять до вивчення морфонології і становлять об'єкт вивчення фонетики.

Лексикографічна концептуальна модель морфемного словника, керуючись методологією взаємовизначення операційної методики та онтологічного упорядження, повинна враховувати онтологічний статус об'єктів морфемної системи, їхні "рівневі" властивості, а відтак, спроектувати завдання морфотактики та морфонології на принципи проектування лексикографічної системи. Складність структурної організації морфемної системи мови визначає принцип інтегральності лексикографічного опису. Керуючись інтегральним принципом моделювання лексикографічного опису морфеми, О. Зубань у статті «Створення морфемної бази даних: принципи опису морфеми як інваріантної одиниці» [Зубань 2001] уклала інтегральну словникову статтю афіксальної морфеми -ак- за такими лексикографічними параметрами: 1) інваріантне значення морфеми; 2) дистрибутивна характеристика морфів морфеми у табличному записі (додаток.1); 3) морфонологічна характеристика морфеми; 4) інформативність морфеми: граматична; лексична. Наведений приклад словникової статті систематизує великий масив лінгвістичної інформації, що, з одного боку, свідчить про глибину лексикографічного опису, а з іншого, спричинює труднощі для сприйняття користувачем.

Розвиток сучасної лексикографічної теорії і практики характеризується тенденцією інтегрального принципу лексикографування, який уперше був обґрунтований Ю. Апресяном у праці «Интегральное описание языка и системная лексикография» [Апресян 1995]. Інтегральний принцип ставить вимогу дати максимально повну, цілісну інформацію про одиницю опису. Традиційно інтегральний принцип опису застосовується до мікроструктури лексичного словника (словникової статті), а одиницями інтегрального опису, як правило, є слова, тому проблема інтегральності в лексикографії пов'язана із системним описом слова в словнику. Концепція інтегрального лінгвістичного опису Ю. Апресяна базується на синтезі словника і граматики, яка розуміється широко як сукупність усіх правил мови, у тому числі й семантичних. «Вони [словник і граматика] [17] повинні бути максимально узгоджені між собою за типами поданої в них інформації і за формальними способами (мовами) її запису. [...] Звідси слідує дуже суттєва вимога до того, як лінгвістична інформація буде розподілена між граматикою і словником, а значить і до того, як повинні працювати граматист і лінгвіст» [ЯКМИСЛ 2006: 42 – 43]. Відповідно, в укладанні словникової статті необхідно дотримуватися двох основних практичних принципів: «1) Будуючи словникову статтю певної лексеми, лінгвіст повинен працювати на всьому просторі граматичних правил і приписати лексемі всі властивості, до яких можуть звертатися правила (налаштування словника на граматику). 2) Будуючи певне правило, лінгвіст повинен працювати на всьому просторі лексем і врахувати всі типи їх поведінки, які не передбачені в словнику (налаштування граматики на словник)» [Апресян 1995: 135].

[17] Інформація додана автором монографії.

Першим інтегральним словником нового типу у слов'янській лексикографії можна по праву вважати «Толково-комбинаторный словарь русского языка» [Мельчук 2016], який був опублікований у 1984 р. Осоливістю розвитку української інтегральної лексикографії є те, що перший інтегральний словник «Активні ресурси сучасної української номінації» [АРСУН 2013:] подає, крім опису неолексем, і системний опис словотвірних афіксів. Одиницями опису в словнику є слова-інновації та словотвірні неоформанти, а словникові статті містять зону "Епідигматичні (деривaційні) відношення реєстрової одиниці (епідигм)". До інтегральних словників нового типу також належить «Етимологічний словник запозичених суфіксів і суфіксоїдів в українській мові» [Селігей 2014], який демонструє інтегральну лексикографічну модель опису етимології афіксальних морфем.

Принцип інтегральності лексикографічного опису задовольняє вимоги повноти й експланаторності наукового лінгвістичного опису, але, як свідчить лексикографічна практика, інтегральні словники є надзвичайно складними для сприйняття інформації користувачами. Наприклад, словникова стаття словника «Активні ресурси сучасної української номінації» може займати до 10-ти сторінок тексту і об'єднує 7 зон, які у свою чергу поділяються на підзони, а в деяких випадках підзони можуть структуруватися на менші підзони до 4-го рівня ієрархії; крім того, словникові статті, зони та підзони мають доповнювальні відношення в представленні лінгвістичної інформації макроструктури словника.

На наше переконання, складна лексикографічна модель інтегрального опису може бути значно спрощена для сприйняття та роботи користувача, якщо текстову інформацію структурувати й представити окремими пошуковими зонами в електронній лексикографічній системі. Це підтверджує досвід конвертації паперового інтегрального словника в електронну версію, який описано О. Зубань у статті «Автоматична конвертація паперового словника «Активні ресурси сучасної української номінації» в електронну лексикографічну систему» [Зубань 2019].

Наукова значущість інтегрального лексикографічного принципу проявляє свої переваги в інтерактивних електронних лексикографічних системах, де кожен лексикографічний параметр може подаватися окремою зоною (екраном) і бути навігаційно пов'язаним з іншими параметрами через гіперпокликання. Ю. Апресян, аналізуючи реалізацію моделі "Смисл↔Текст", до якої входить і тлумачно-комбінаторний словник, в автоматичній системі машинного перекладу ЭТАП, зауважує, що «...сучасне розуміння принципу інтегральності могло визначитися й, дійсно, визначилося тільки після того, як ідеологія моделі "Смисл↔Текст" у достатньо повному обсязі була реалізована на комп'ютері. [...] Тільки після того, як ці граматики і словники стали реально функціонувати на комп'ютері й були багаторазово відредаговані за результатами експерименту на великих масивах текстів, стало зрозуміло, який величезний обсяг граматичних і

лексикографічних результатів виводиться із принципу інтегральності» [ЯКМИСЛ 2006: 42].

У нащому дослідженні пропонується новий підхід до створення інтегральної лексикографічної моделі морфемної лексикографічної системи електронного типу. У цій моделі змінюється аспект принципу інтегральності лінгвістичного опису: вимога інтегральності ставиться не до словникової статті, а до макроструктури лексикографічної системи загалом. Внутрішня лексикографічна модель електронної системи повинна складатися з баз даних, які систематизують різні лексикографічні параметри, а описувана реєстрова одиниця – морфема – наскрізно пов'язана із усіма полями баз даних. Таким чином формується не інтегральна словникова стаття, а інтегральна структура лексикографічної системи, на основі якої будується зовнішня лексикографічна модель, яка передбачає багато входів у цю систему й об'єднує три зони:

1) словник морфів, диференційований за функціональними типами морфів, із представленням повного реєстру слів, у яких реалізований кожен морф;

2) словник морфемних структур слів, типізованих в окремі групи за символьними моделями морфемних структур.

3) тлумачно-морфонологічний словник морфем, у якому морфема описана як інваріантно-варіантний конструкт.

Реалізація такої електронної лексикографічної моделі на великих лексичних і текстових масивах уможливлює доповнення визначених типів словників статистичною інформацією, тому всі три типи словників можуть бути частотними.

Формалізована систематизація лінгвістичних даних у МБД АСМСА забезпечує автоматичне укладання трьох типів словників і конструювання інтегральної електронної лексикографічної системи з морфеміки української мови. У проекції на лексичний масив мови це лексикографічне завдання ще не було виконано, але було розроблено проєкт лексикографічної системи «Морфограф» [Зубань 2017]. У дослідженні українськомовного тексту в Корпусі української мови пропонована лексикографічна модель була частково реалізована у серії електронних частотних морфемних словників [Зубань 2016], [Zuban 2019], [Zuban 2017]. Автоматично були сконструйовані два перші типи словників[18] (словник морфів, словник морфемних структур слів), які представлені в мережі Інтернет як інтерактивні лексикографічні системи [ЧСКУМ 2017].

[18] Принципи даталогічного етапу проектування електронної лексикографічної системи частотних морфемних словників за окремими текстовими вибірками, а також класифікаційні та пошукові можливості цих словників будуть представлені у третьому розділі монографії.

1.5.2. Проєкт інтерактивної комп'ютерної лексикографічної системи «Морфограф»

Лексикографічна система «Морфограф» будується на теоретичних засадах сучасної морфемології й ставить завдання систематизувати та описати основні одиниці морфемної системи української мови – морфи, морфеми та морфемні структури слів. Конструювання системи «Морфограф» здійснюватиметься автоматично за даними МБД АСМСА, що систематизує інформацію про морфемну будову ≈ 200 тис. слів української мови в п'яти БД-таблицях:

1) БД морфемних структур слів ≈ 200 тис. слів;
2) БД аломорфічних коренів (≈ 2500 коренів);
3) БД омонімічних коренів (≈ 3100 коренів);
4) БД афіксальних морфем (у процесі укладання);
5) БД морфонологічних альтернацій (у процесі укладання).

Лексикографічні завдання, які ставляться у створенні системи «Морфограф», зумовлюють необхідність редагування лексичного реєстру АСМСА й формування автономної морфемної бази даних, яка за класифікаційними параметрами лінгвістичної інформації та обсягом лексичного реєстру якісно і кількісно відрізнятиметься від МБД АСМСА. У формуванні лексичного реєстру планується:

- вилучення власних назв та загальних назв з онімними коренями, тому що оніми характеризуються особливою семантикою та морфемною структурою й утворюють велику ономастичну систему, що заслуговує на окремий лексикографічний опис;
- вилучення буквених та звукових абревіатур, тому що семантика та морфемна будова цих слів також потребує окремого теоретичного підходу та принципів морфемного аналізу;
- доповнення регулярними дериватами, які не ввійшли до реєстрів традиційних словників і не представлені в МБД: синтетичними формами ступенів порівняння якісних прикметників, відприкметниковими прислівниками, дієприкметниками, абстрактними іменниками-девербативами із суфіксом -нн-я;
- поповнення словозмінними формами іменників, прикметників, займенників, числівників, дієслів, у кореневих та афіксальних морфемах яких відбуваються морфонологічні альтернації.

Всі завдання планується виконати автоматично, використовуючи ресурси морфемної бази даних і морфологічної бази даних системи АГАТ та алгоритмів автоматичного морфемного синтезу слів.

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 сформованого реєстру чи словникової статті.

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/ти* PRSF; *ви/чит/а/н/ий* PRSSF; *ви/чит/а/н/о* PRSSS. У цій зоні користувач може автоматично групувати реєстрові одиниці в різноманітні вибірки за такими пошуковими параметрами:

1) "кількість коренів": усі слова; прості слова; складні слова;

2) "граматична характеристика": усі 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": усі слова; прості слова; складні слова;

2) "граматична характеристика": усі слова; обрана у випадному списку частина мови;

3) "кількість морфем у слові": усі 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, у яких реалізована кожна модель.

На другому рівні класифікаційних опцій здійснюється:

1) автоматичне формування списку слів за обраною моделлю, наприклад: PRSF *в/тич/к/а*, *в/тіх/оньк/а, в/тул/к/а, в/тіш/ниц/я*;

2) автоматичне формування списку морфів за обраним типом морфа в моделі, наприклад: PRSF P (*без-, в-, ви-, во-, ді-, до-, з-, за-, зав*); R *(-бав-, -бавл-, -багат-, -баж-, -бач-, -бг-, -би-, -бир-, -біг-*);

3) автоматичне формування парадигми моделі з 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 тлумачно-морфонологічний словник кореневих 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": усі слова; прості слова; складні слова; 2) "граматична характеристика": усі 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, у яких реалізований морф, обраний зі списку або заданий користувачем, наприклад, суфікс -ун-: *авіа/двиг/ун/0* RRSF, *баб/ун/я* RSF, *балак/ун/0* RSF, *балак/ун/ств/о* 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.

Тлумачно-морфонологічний словник кореневих морфем укладатиметься на матеріалі трьох баз даних МБД АСМСА: БД морфемних структур, БД омонімічних коренів, БД аломорфічних коренів. Програмний взаємозв'язок цих баз забезпечує автоматичну класифікацію слів української мови в спільнокореневі вибірки з урахуванням кореневої омонімії та аломорфії. На макрорівні словник кореневих морфем будуватиметься автоматично користувачем за такими пошуковими опціями: 1) усі корені; 2) корені обраної у випадному списку частини мови; 3) один корінь, заданий користувачем.

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/аj/уч/ий* визначається дієслівний кореневий морф -берег-.

Словникова стаття має 6 зон:

1) вихідний морф (реєстрова одиниця);

2) непохідне слово із визначеним вихідним морфом;

3) частина мови непохідного слова, яка визначає категорійну семантику морфеми;

4) тлумачення слова: лексичні значення всіх ЛСВ непохідного (першого) слова спільнокореневої групи;

5) варіанти морфеми: аломорфи, варіоморфи для частково тотожних за формою морфів; один морф, що представляє реалізацію морфеми тільки в однакових за формою морфах;

6) спільнокореневі слова, у яких реалізовані варіанти морфеми.

Зони (1), (2), (3), (4), (5) будуть подаватися на першому інтерактивному рівні (рис. 1.10), а зона (6) на другому інтерактивному рівні (рис. 1.11). У першій і п'ятій зонах подаватиметься кількісна характеристика реалізації морфеми в словах української мови.

-берег- (72);
2. *берег-ти;*
3. Граматичне значення дієслово: процесуальність;
4. Семантика 1) не давати витрачатися, 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. Морфемна парадигма -берег- (15), -береж- (39), -беріг- (18).

Рис.1.10. Словникова стаття кореневої морфеми (перший інтерактивний рівень).

На другому рівні інтерактивного конструювання словника користувач працює зі словниковою статтею й може поглиблювати лінгвістичну інформацію про морфему, використовуючи нові класифікаційно-пошукові опції. За вибором варіанта морфеми у зоні (5) формуються списки спільнокореневих слів, у яких реалізований обраний морф (рис.1.11).

Сьогодні ведеться активна робота над редагуванням та розбудовою МБД АСМСА, і тому наш колектив планує реалізувати цей проєкт. Система «Морфограф» поглибить теорію і практику сучасної української комп’ютерної лексикографії як в аспекті автоматичного, так і в аспекті пошуково-класифікаційних можливостей користувача. Ця система матиме інтегральний лексикографічний характер, що визначається поєднанням і взаємодоповненням інформації, дібраної з різних інтерактивних морфемних словників, і виконуватиме функцію ефективного комп’ютерного інструмента, здатного автоматично формувати за вибором користувача саме той лінгвістичний матеріал, який необхідний йому в доборі, обробленні та систематизації лінгвістичної інформації. Лексикографічний опис, який буде зроблено у цій системі, репрезентуватиме великий обсяг лексичного матеріалу й багатоаспектний морфемний аналіз, що ґрунтується на узагальненні провідних теоретичних ідей сучасної морфотактики та морфонології, тому за критерієм лінгвістичної інформативності система «Морфограф» представлятиме повний і цілісний лінгвістичний опис морфемної системи української мови.

-берег-
берег|ти|ся
берег|ти
берег|ин|я
в|берег|ти
в|берег|ти|ся
з|берег|ти|ся
з|берег|ти
о|берег|ти
о|берег|ти|ся
о|берег|Ø|Ø
по|берег|ти
по|берег|ти|ся
при|берег|ти
у|берег|ти
у|берег|ти|ся

-береж-
береж|ен|ий
береж|ен|ий 1
береж|інн|я
береж|к|ий
береж|лив|ий
береж|лив|ість|Ø
береж|лив|о
береж|н|ість|
береж|н|ий 1
береж|н|о
в|береж|ено|
енерг|о|з|береж|енн|я
з|береж|ен|ий
з|береж|ен|ість|Ø
з|береж|енн|я
з|береж|ен|о
ліс|о|з|береж|енн|я
над|о|береж|н|ість|
най|о|береж|н|іш|ий
най|о|береж|н|іш|е
не|з|береж|енн|я
не|о|береж|н|ий
не|о|береж|н|ість|Ø
не|о|береж|н|о
о|береж|н|ий
о|береж|н|ість|Ø
о|береж|н|еньк|о
о|береж|н|іш|ий
о|береж|н|іш|е
о|береж|н|о
по|береж|н|ик|Ø 1
при|береж|ен|ий
при|береж|ен|о
ресурс|о|з|береж|енн|я
сам|о|з|береж|енн|я
тепл|о|з|береж|н|ість|Ø
у|береж|ен|ий
у|береж|енн|я
у|береж|ен|о

-беріг-
в|беріг|а|ти|ся
в|беріг|а|ти
в|беріг|а|нн|я
енерг|о|з|беріг|aj|уч|ий
з|беріг|а|ти|ся
з|беріг|а|ти
з|беріг|а|нн|я
з|беріг|а|ч|
ліс|о|з|беріг|а|нн|я
о|беріг|Ø|Ø
о|беріг|а|ти
о|беріг|а|ти|ся
о|беріг|а|нн|я
при|беріг|а|ти
ресурс|о|з|беріг|aj|уч|ий
у|беріг|а|ти|ся
у|беріг|а|ти
у|беріг|а|нн|я

Рис.1.11. Словникова стаття кореневої морфеми (другий інтерактивний рівень).

Р О З Д І Л 2

АВТОМАТИЗОВАНА СИСТЕМА МОРФЕМНО-СЛОВОТВІРНОГО АНАЛІЗУ СЛІВ УКРАЇНСЬКОЇ МОВИ – БАЗА ЗНАНЬ УКРАЇНСЬКОЇ МОРФЕМОЛОГІЇ ТА СЛОВОТВОРУ

2.1. АСМСА: етапи створення та загальна характеристика структури даних

У час загальної комп'ютеризації змінюються методи і технології лінгвістичного аналізу, і зокрема сучасної лексикографії. Традиційний паперовий словник перестає бути єдиним та ефективним способом представлення знань із двох причин: 1) на сьогодні такий словник не задовольняє потреб користувача, який працює не в бібліотеці, а в Інтернеті; 2) укладання паперового словника вимагає багато часу та людських ресурсів для збирання, оброблення та систематизації матеріалу у вигляді паперових картотек.

Тому в українському мовознавстві на сьогодні нагальними є завдання, спрямовані на укладання параметризованих електронних баз знань та та представлення їх в Інтернеті у формі електронних лінгвістичних словників, оснащених пошуково-класифікаційними програмними аналізаторами для ефективного та оперативного проведення лінгвістичного аналізу.

Автоматизована система морфемно-словотвірного аналізу (АСМСА) – це спеціалізована інтелектуальна система, яка, крім функції інформаційно-довідкової системи з морфеміки та словотвору української мови, початково була спрямована на автоматичний аналіз тексту й може використовуватися в системах автоматичного оброблення тексту у функціях автоматичного морфемного сегментатора текстових слововживань / початкових форм (лем) слів та автоматичного конструктора частотних морфемних словників за текстовими вибірками.

Інтелектуальні системи такого типу є експертними системами, що побудовані на основі баз знань конкретної предметної галузі. У сучасній наукометрії, зокрема в галузях комп'ютерної лінгвістики та інформаційних технологій, використовують два терміни на позначення сукупності даних в інформаційних системах: "бази даних" та "бази знань", які потребують послідовного розмежування або уточнення термінологічного вжитку.

Як зазначає Є. Карпіловська: «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 і 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» [Карпіловська 2006: 34].

Подібне розмежування термінів знаходимо в С. Субботіна: база знань, як компонент експертної системи, поділяється на: 1) статичну частину, у якій «зберігаються довгострокові знання, що описують розглянуту предметну область у вигляді загальних фактів [...] і правил [...], які описують доцільні перетворення фактів цієї області з метою породження нових фактів чи гіпотез" [Субботін 2008: 22]; та 2) динамічну частину – робочу пам’ять (базу даних), «що змінює свій стан під впливом правил, призначена для збереження вихідних даних (...) і проміжних даних розв’язуваної в поточний момент задачі [...]» [Субботін 2008: 22].

В інформаційних технологіях, за стандартом ISO/IEC 2382:2015, база даних (database) визначається як сукупність даних, організованих відповідно до концепції, яка описує характеристику цих даних і взаємозв’язки між їх елементами; ця сукупність підтримує щонайменше одну зі сфер застосування, а база знань (knowledge base) − база даних, що містить правила виведення та інформацію про людський досвід і знання в деякій предметній галузі [ISO/IEC 2015]. Зокрема прийнято вважати, що бази даних інтелектуальних експертних систем є сукупністю особливих даних: законів або принципів предметної галузі, що отримані в результаті практичної діяльності й професійного досвіду, а отже є базами знань. Таким чином, в інженерії програмного забезпечення терміни "база даних" та "база знань" розмежовуються тільки за характером даних, тому база знань також є базою даних, але не кожна база даних є базою знань.

В описах організації структури загальних інформаційних систем, що складається із БД і СКБД (системи керування базою даних), або СУБД (системи управління базою даних) вживається термін "база даних", а терміна "система керування базами знань" не існує: «Компонентами системи баз даних є БД, СУБД і прикладні програми, з якими працюють як розробники, так і користувачі» [Гайна 2005: 14]. «База даних (Database) − організована згідно з певними правилами й збережена в пам’яті комп’ютера сукупність даних, яка характеризує актуальний стан деякої предметної області й використовується для задоволення інформаційних потреб користувачів» [Гайна 2005: 179].

В описі структури і функцій системи АСМСА використовуються терміни:

- "база знань" − сукупність знань із морфемології та словотвору української мови як окремої предметної лінгвістичної галузі;
- "база даних" у широкому загальному трактуванні − сукупність особливих даних, що систематизують знання з морфеміки і є частиною бази знань поряд із системою керування БД;

- "база даних" у вузькому розумінні − окрема БД-таблиця в загальній структурі БД.

Концепція та структура АСМСА зумовлювались її функціональним призначенням:

- здійснювати автоматичну класифікацію лексики (за словником та текстом) у спільнокореневі, спільноафіксальні та спільноструктурні вибірки;
- здійснювати автоматичне формування реєстру афіксальних та кореневих морфем за лексичними вибірками словника та тексту;
- здійснювати автоматичне / автоматизоване укладання словотвірних гнізд за словником.

Структурування системи передбачало послідовність виконання таких завдань (рис. 2.1): укладання морфемної бази даних (1); → розроблення програмного забезпечення – СКБД морфемної бази даних (2), що виконує автоматичну морфемну класифікацію, зокрема автоматичне укладання спільнокореневих вибірок; → автоматичне укладання словотвірної бази даних (3); → розроблення програмного забезпечення – СКБД словотвірної бази (4), що виконує автоматичну / автоматизовану класифікацію лексики за словотвірними ознаками в межах словотвірного гнізда.

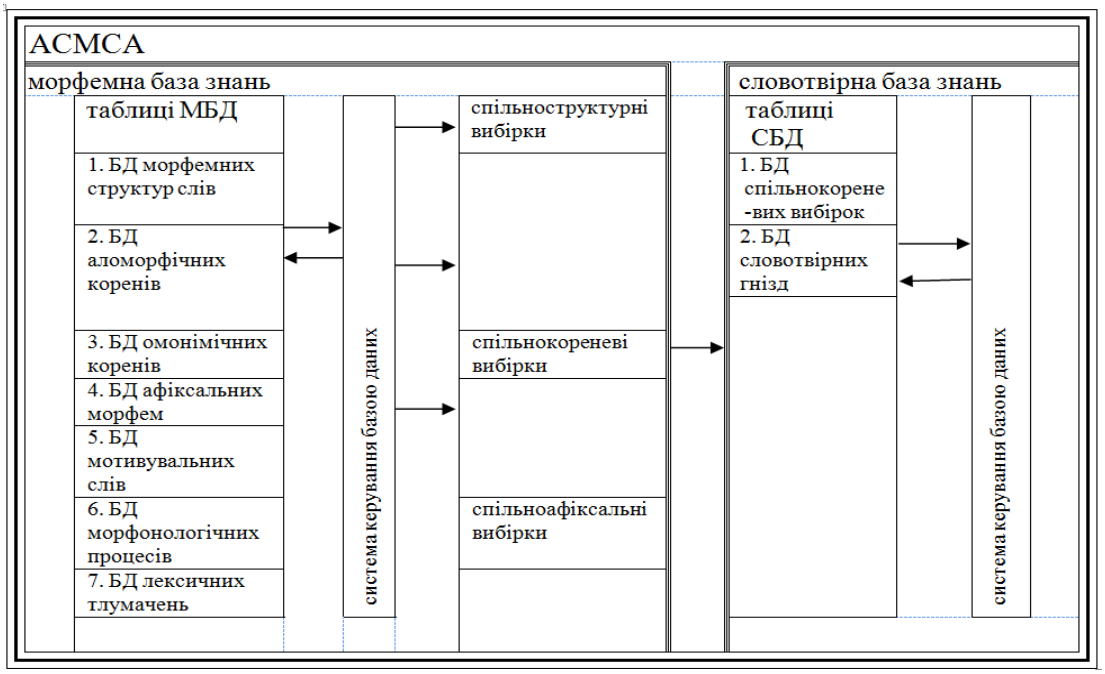


Рис 2.1. Структура АСМСА

Як показує схема, центральним і вихідним компонентом у роботі АСМСА є морфемна база даних (МБД), яка на сьогодні має розгалужену структуру й складається із 7-ми окремих БД-таблиць:

1) БД морфемних структур слів ≈ 200 тис. слів;
2) БД аломорфічних коренів (≈ 2500 коренів);
3) БД омонімічних коренів (≈ 3100 коренів);
4) БД афіксальних морфем (у процесі укладання);

5) БД мотивувальних слів (у процесі укладання);
6) БД морфонологічних процесів (у процесі укладання);
7) БД лексичних тлумаченью

Словотвірна база даних АСМСА (рис.2.1) об'єднує 2 таблиці: (1) БД вибірок спільнокореневих слів та (2) БД словотвірних гнізд.

СКБД обох баз даних (МБД та СБД) працює за допомогою програмних та технічних засобів, що забезпечують:

- оголошення даних – створення, зміну й видалення визначень, які описують організацію кожної БД-таблиці;
- модифікацію даних – додавання даних про морфемні й словотвірні об'єкти та явища, редагування та видалення цих даних;
- отримання даних – надання даних про морфемні й словотвірні об'єкти та явища за запитом у формі, яка дозволяє їх аналізувати й отримувати лінгвістичні знання;
- адміністрування даних – реєстрування та облік дій користувачів, які працюють із АСМСА.

СКБД АСМСА розроблена на основі програми MS Access, яка здійснює конструювання таблиць МБД та збереження даних; а також на основі допоміжних програм, розроблених із використанням об'єктно-орієнтованих мов програмування С++ та С # (програміст В. Сорокін).

За моделлю організації структури даних АСМСА визначається як реляційна БД, організована у вигляді набору формально описаних таблиць, із яких дані про морфемні й словотвірні об'єкти та явища можуть бути повторно дібраними різними способами без необхідності реорганізації таблиць бази даних.

Перераховані типи виконуваних завдань та структурація АСМСА на БД-таблиці демонструють поліфункціональність цієї бази даних і широкий спектр морфемних та словотвірних об'єктів, явищ і процесів, проаналізованих у цій базі, що визначає морфемну базу даних АСМСА базою знань із морфемології та словотвору української мови.

АСМСА має людино-машинний інтерфейс, тому за класифікаційними ознаками комп'ютерних словників, визначених В. Перебийніс [Перебийніс 2009: 24], АСМСА може розглядатися як комп'ютерна лексикографічна система нового типу різнофункціонального призначення:

- за зовнішньою лексикографічною моделлю, це – комп'ютерний (електронний) інтегрований морфемний словник, призначений для користувача в лінгвістичних дослідженнях;
- за внутрішньою лексикографічною моделлю, це – автоматичний словник, що використовуються в комп'ютерних системах автоматичного морфемного аналізу.

Упродовж всього періоду роботи над створенням АСМСА, можна визначити декілька етапів укладання та конструювання, що відображають різні завдання автоматичного морфемного та словотвірного аналізів і поповнення бази даних новою лексикою.

1 етап:

- укладання бази даних морфемних структур слів, де в автоматизованому режимі було просегментовано на морфеми ≈ 170 тис. слів і приписано кожному слову модель морфемної структури;
- укладання бази даних аломорфічних коренів;
- укладання бази даних омонімічних коренів.

Ці завдання виконувались із метою:

1) автоматичного групування слів у спільнокореневі вибірки з урахуванням омонімії та аломорфії коренів і автоматичного / автоматизованого укладання словотвірної бази даних[19];

2) автоматичного сегментування текстових слововживань української мови на морфи в Параметризованій базі даних поетичного мовлення [Алексеенко 2004];

3) автоматичного проведення морфемного аналізу початкових форм слів (лематизованих слововживань) і автоматичного укладання алфавітно-частотних морфемних словників за текстовими вибірками Корпусу української мови [Zuban 2015], [Zuban 2017], [Zuban 2019], [Зубань 2016а].

2 етап:

- укладання БД афіксальних морфем (приписування кожній афіксальній морфемі у складі слова значення та функції);
- укладання БД мотивувальних слів (визначення мотивувального слова та твірної основи);
- укладання БД морфонологічних процесів (приписування афіксальній морфемі морфонологічного явища, яке спричинює аломорфію афікса).

Ці завдання виконуються з метою автоматичного групування лексики в спільноафіксальні вибірки з урахуванням омонімії та аломорфії афіксів і укладання електронного семантичного словника афіксальних морфем.

3 етап:

- поповнення МБД новою лексикою, яка автоматично добиралася як "необроблена" за текстами Корпусу української мови, на сьогодні реєстр МБД складає ≈ 200 тис. слів;
- редагування БД омонімічних коренів та БД аломорфічних коренів (приписування кожній кореневій морфемі з урахуванням омонімії семантичного тлумачення).

Ці завдання виконуються з метою автоматичного укладання електронного семантичного словника кореневих морфем.

Етапи укладання АСМСА демонструють ускладнення функціонального потенціалу та розбудови структури системи, що потребує окремого опису та пояснення концептуальних понять моделювання й операційних принципів конструювання.

[19] БД-таблиці СБД та процедура укладання цієї бази даних буде описана у § 2.3.

2.2. Морфемна база даних АСМСА: структура, функції та процедура укладання

Центральним і вихідним складником МБД є БД морфемних структур слів (рис. 2.1), з укладання якої і розпочалась робота над створенням АСМСА. БД морфемних структур слів створювалась на матеріалі зведеного реєстру (обсягом ≈ 170 тис.) українських слів, взятого з резидентного словника лінгвістичного модуля автоматичної системи перевірки орфографії української мови «РУТА» [Пролинг РУТА 5.0]. Цей словник було укладено на базі відомих українських словників: «Словник української мови» [СУМ 1970–1980], «Словник-довідник з правопису та слововживань» [Головащук 1989], «Словник іншомовних слів» [СІС 1985]; а також доповнено загальними та власними назвами, дібраними з українських текстів у процесі створення цієї системи.

Рутівська БД представляла алфавітний список початкових форм слів із приписаним двосимвольним граматичним кодом (додаток 2), який визначав частину мови та граматичне значення слова, наприклад: *дуло*, ЛИ (де, Л – іменник середнього роду; И – називний відмінок). Пізніше двосимвольний код було змінено на односимвольний (тільки граматична інформація про частину мови й несловозмінні грамеми: *дуло*, Л – іменник середнього роду), тому що інформація про реляційні граматичні значення початкової форми (називний відмінок чи інфінітив) була надлишковою для морфемного аналізу початкових форм. У такому записі морфемна база поповнюється й сьогодні, нові слова потрапляють до МБД після морфологічного анотування текстів у Корпусі української мови, наприклад:

ІТ-дистриб’ютор,Й,
uанет,Й,
авіньйон,й,
автошарж,Й,
автошляховик,Й,
агафонов,й,
аґент,Й,

На першому етапі оброблення сформованого лексичного реєстру (компілятивної БД, записаної у lex-файлі) було автоматично проведено конвертацію графемного запису слів у спрощену фонематичну транскрипцію (див. § 1.4.2).

Другим завданням в укладанні БД морфструктур було проведення морфемного аналізу початкових форм: кожному слову реєстру в автоматизованому режимі за допомогою конструктора, розробленого на основі об’єктно-орієнтованих мов програмування С++ (на початковому етапі) та С # (на нинішньому етапі укладання та редагування БД), приписується модель, яка визначає межі та функціональний тип морфів.

Кожне слово lex-файлу представлялося на окремій інтерфейс-картці morfem.voc (рис. 2.2 та рис.2.4), з якою працює лінгвіст-укладач.

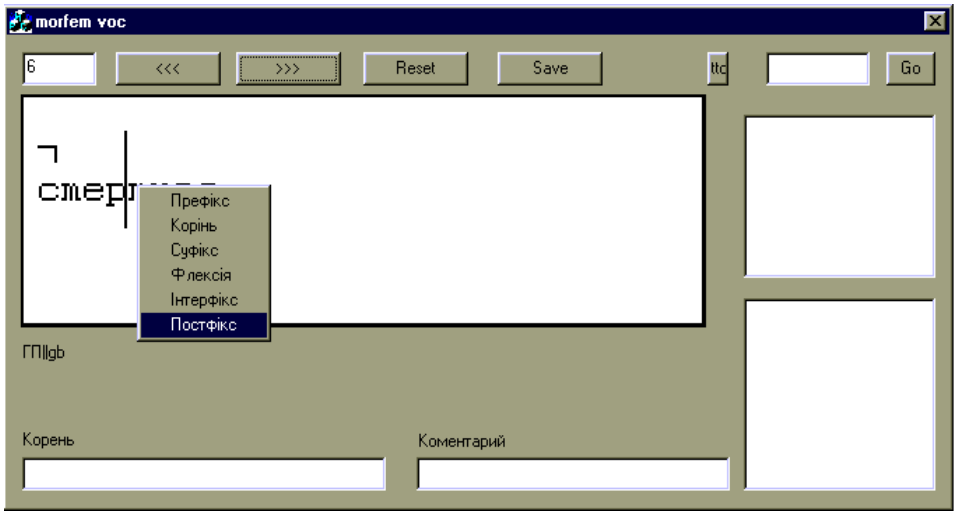
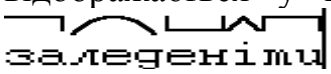


Рис.2.2. Інтерфейс першої робочої картки МБД (тестовий варіант)

Визначення морфемної структури слова здійснювалося в такий спосіб: на екрані є рухомий курсор, який пересувається мишею вліво або вправо по графемному запису слова; навівши курсор на місце морфемного шва, потрібно правою кнопкою миші актуалізувати випадний список, у якому висвітлюється перелік позиційно-функціональних типів морфем. Вибравши курсором потрібний тип морфеми, автоматизовано приписуємо йому позначку у верхньому індексі слова, наприклад, на рис. 2.2 визначено префікс у слові *стерти*. Визначення типу морфеми, що на картці відображається у верхньому індексі аналізованого слова, наприклад –  , супроводжується автоматичним перекодуванням у програмну процедуру: *заледеніти* PCRFSHSIFK (див. § 1.4.2). Одиницею реєстру БД морфемних структур виступає графемний запис слова з програмною процедурою. Реєстр БД має вигляд автоматичного словника:

безсонність,К,PDRGSHSLFM/сон2/
безсонниця,К,PDRGSHSJFK/сон2/
безсонно,Н,PDRGSHSI/сон2/
безсонячний,А,PDRGSISJFL/сон1/
відсоння,Л,PDRGSHFI/сон1/
дисонанс,Й,PCRFSIFJ/сон3/
дисонансовий,А,PCRFSISKFM/сон3/
сонько,Й,RESFFG/сон2/, де кожному слову приписана лінгвістична інформація про: частину мови (додаток 2), наприклад К – іменник жіночого роду; модель програмної процедури сегментації слова на морфеми (PDRGSHSLFM; PDRGSHSI); індекс кореня за умови його омонімії чи аломорфії (/сон1/; /сон2/; /сон3/).

Запис отриманих нових даних здійснювався автоматично у форматі res.ini-файлів (рис. 2.3).

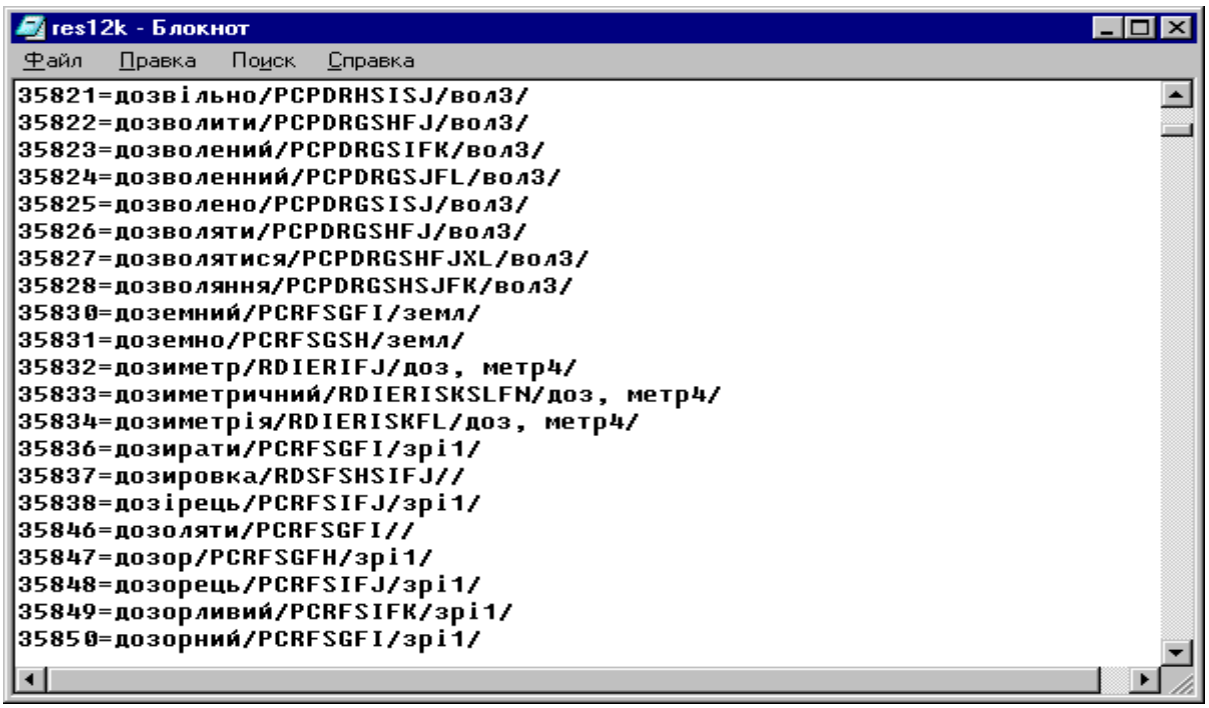


Рис. 2.3. Фрагмент res.ini-файлу

На першому етапі укладання база даних мала локальний характер, і тому СКБД використовувала файлову систему збереження даних: структура запису даних була закладена в прикладній програмі, з якою працював укладач БД. Різні укладачі розрізнено працювали з даними lex-файлу, який був установлений окремо на кожному комп'ютері. Програма забезпечувала доступ до даних для кожного укладача на окремому комп'ютері, але в системі керування були відсутні централізовані методи керування доступом до інформації: укладачі не могли паралельно працювати з даними, які локально зібрані на серверному комп'ютері. Кожен укладач працював зі своєю частиною lex-файлу за допомогою робочої картки morfem.voc. і записував результати морфемного аналізу в окремі res.ini-файли, які об'єднувалися в єдину базу даних на одному комп'ютері завдяки однаковому порядковому номеру слова у lex-файлі та res.ini-файлах.

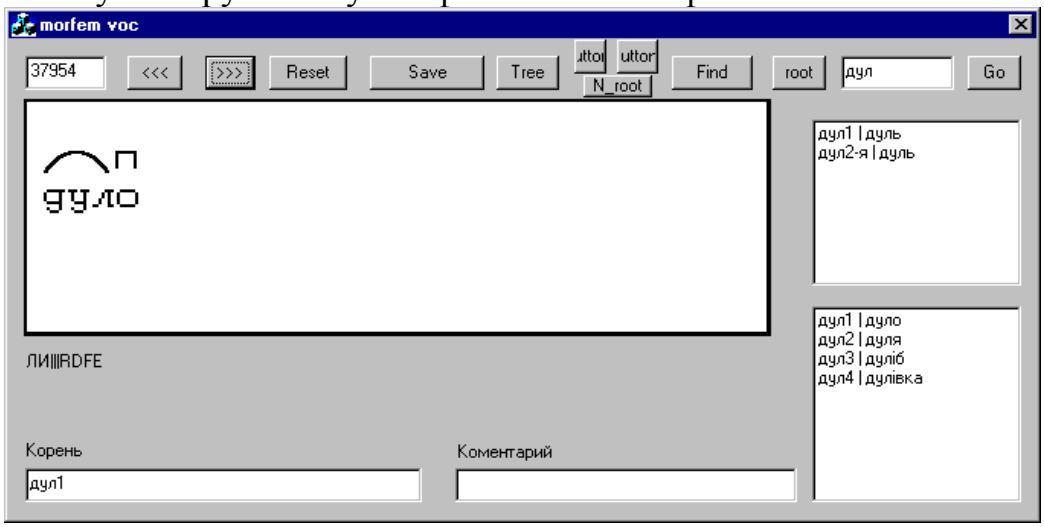


Рис. 2.4. Інтерфейс робочої картки morfem.voc

Кожне поле інтерфейс-картки пов'язано з полями БД-таблиць і має своє функціональне призначення у роботі лінгвіста-укладача:

(1) поле 37954 – *номер слова в МБД*: навігація по lex-файлу;

(2) кнопка <<< – *лівобічна навігація*: повернутися до попереднього слова у МБД;

(3) кнопка >>> – *правобічна навігація*: перейти до наступного слова у МБД;

(4) кнопка *Reset*: витерти приписану морфемну процедуру у полі (5), якщо необхідно виправити помилку;

(5) поле *дуло*: висвітлюється слово, у якому лінгвіст-укладач спеціальними індексами визначає межу морфа та його функціональний тип: ця інформація після активації кнопки (8) *Save* зберігається в res.ini-файлі; після збереження під екраном висвітлюються автоматично сформовано програмна процедура, що моделює морфемну будову слова: RDFE;

(6) поле *Корінь – дул1*: записується інваріантний аломорф кореня або омонімічний код кореня.

(7) поле *Коментар:* записується різноманітна додаткова інформація, необхідна лінгвісту-укладачеві для правильного морфемного аналізу (про лексичне значення омонімів, етимологічне тлумачення запозиченої лексики, пояснення складних випадків морфемного аналізу, тощо).

(8) кнопка *Save*: запам’ятати дані, приписані лінгвістом-укладачем у полі (5), (6) та (7) у res.ini-файлі;

(9) поле *Аналізований корінь – дул* (у правому верхньому кутку картки): автоматично висвітлюється корінь *дул*, межі якого визначені в полі (5), або записується корінь, який укладач хоче знайти в сателітних кореневих базах;

(10) кнопка Go: пошук кореневого морфа, записаного в полі (9), у сателітних кореневих базах, з якими мають зв'язок поле (11) та (12);

(11) поле *Аломорфічні корені – дул1|дуль, дул2-я|дуль*: висвітлюється інформація про аломорфію аналізованого або записаного в полі (9) кореневого морфа через зв’язок із БД аломорфічних коренів.

(12) поле *Омонімічні корені – дул1|дуло, дул2|дуля, дул3|дуліб, дул4|дулівка*: висвітлюється інформація про омонімічні кореневі морфи до аналізованого або записаного в полі (9) кореня через зв’язок із БД омонімічних коренів.

Кнопки *Tree* та *Find* пов’язують МБД із словотвірною базою даних (рис.2.1).

Приписування інформації про омонімію та аломорфію кореня до аналізованого слова здійснювалось із метою автоматичної класифікації лексики в спільнокореневі вибірки та автоматизованого формування словотвірного гнізда в СБД АСМСА. Коренева система української мови

характеризується високою омонімією та аломорфією, тому в межах МБД АСМСА в ручному режимі формувались ще дві сателітні бази:

1) БД аломорфічних коренів (≈ 2500 коренів);

2) БД омонімічних коренів (≈ 3100 коренів.

Ці бази укладались паралельно із БД морфемних структур: кореню аналізованого слова приписувався індекс омонімії, а аломорфу – інваріантнаий морф. Наприклад, коренева морфема -*бал*- реалізовується в українських словах у трьох омонімічних коренях: бал1 *(одиниця виміру)*; бал2 *(танцювальний вечір)*; бал3 – *(балувати)*, які записуються в БД омонімічних коренів (рис. 2.5).

Код	n1	n2	Добавить поле
3956	БАГАТ1	БАГАТИЙ	
3957	БАГАТ2	БАГАТО	
3708	БАЛ1	ОДИНИЦЯ ВИМІРУ	
3709	БАЛ2	ТАНЦЮВАЛЬНИЙ ВЕЧІР	
3710	БАЛ3	БАЛ-УВАТИ	
3970	балк1	балка-яр	
3971	балк2	балка-колода	
3972	бан1	баня	
3973	бан2	банувати-сумувати	
2917	БАНК 1	БАНК	
2918	БАНК 2	БАНКА	
2919	БАНК 3	МІЛИНА	
3733	БАР1	ОДИНИЦЯ ВИМІРУ	
3734	БАР2	РЕСТОРАН	
3735	БАР3	БАР-ИТИСЯ	
3711	БАС1	ГОЛОС	
3712	БАС2	БАС-КИЙ	
3974	бач1	вибачити	
3975	бач2	бачити	
2920	БЕРЕГ 1	БЕРЕГ	
2921	БЕРЕГ 2	БЕРЕГТИ	
4240	бик1	самець корови	
4241	бик2	бики (опори моста)	
4242	бик3	казан	
3842	БІЛ1	БІЛИЙ	
3843	БІЛ2	БІЛ-КА	
3844	БІЛ3	БІЛ-ОК	
2922	БІЛЬ 1	БІЛЬ (страждання)	
2923	БІЛЬ 2	(хвороба рослин)	
2924	БІЛЬ 3	БІЛЬ-Ш-ИЙ	
4040	біс	чорт	
4041	біс1	вигук	

Рис. 2.5. Фрагмент БД омонімічних коренів

Інформація про омонімію висвітлюється у робочій картці в полі (12) за заданим коренем, і тому укладач має доступ до інформації про кореневу омонімію, щоб правильно приписати код омонімії в полі (6), а потім ця інформація приписується до слова в БД морфемних структур, наприклад:

бал,Й,RDFE/бал2/
бал1,Й,RDFE/бал1/
баловий,А,RDSFFH/бал1/
баловство,Л,RDSFSIFJ/бал3/
балощі,И,RDSFFG/бал3/
балуватися,Г,RDSGFIXK/бал3/
балувати,Г,RDSGFI/бал2/
балувати1,Г,RDSGFI/бал3/
балуваний,А,RDSGSHFJ/бал3/
балювати,Г,RDSGFI/бал2/
балюватися,Г,RDSGFIXK/бал2/
збалувати,Г,PBRESHFJ/бал2, бал3/

збалуваний,А,PBRESHSIFK/бал2, бал3/
набалуватися,Г,PCRFSIFKXM /бал3/
небалуваний,А,PCRFSISJFL/бал3/
побалуватися,Г,PCRFSIFKXM/бал3/
побалувати,Г,PCRFSIFK/бал3/
розбалувати,Г,PDRGSJFL/бал3/
розбалуватися,Г,PDRGSJFL/бал3/
розбалуваний,А,PDRGSJSKFM/бал3/
розбалувано,@,PDRGSJSKSL/бал3/

Омонімічні корені -бал1- *(одиниця виміру)* і -бал2- *(танцювальний вечір)* можуть реалізовуватися в спільнокореневих дериватах в аломорфі -баль-. Ця інформація записується в БД аломорфічних коренів (рис. 2.6).

Код	p1	p2	Добавить поле
3	БАВ1	БАВЛ	
488	БАВ2 (добавляти)	БАВЛ	
673	Баварij	БАВАР	
421	БАГАТ1(ий)	БАГАТ, БАГАЧ, БАГАЦЬ	
675	БАЛ1	БАЛЬ	
676	БАЛ2	БАЛЬ	
952	балак	балач	
950	балк1	балоч, балок	
951	балк2	балок, юалоч	
953	балух-и	балуш	
422	БАТЬ(ко)	бат, б.	
677	Башкирij	БАШКИР	
678	БДЖОЛ	БДЖІЛЬ	
11	БЕЗПЕК-А	БЕЗПЕЧ	
692	Бельгij	БЕЛЬГІЙ	
693	Бенгалij	БЕНГАЛЬ	
850	БЕРЕГ1	БЕРЕЖ	
681	БЕРЕГ2	БЕРЕЖ, БЕРІГ	
423	БЕРЕЗЕНЬ	БЕРЕЗН	
773	БЕСТ (РОЗБЕСТИТИ)	БЕЩ (РОЗБЕЩЕНИЙ)	
413	БИ	БІЙ, БЙ, БИЙ, БОЙ, БОJ	
695	БІБЛІJ	БІБЛІЙ	
14	БІГ	БІЖ	
851	БІК	БІЧ, БОК, боц	
954	бiл1	біль	
682	БІЛЬ1	БОЛ	
15	БЛАГ1- О	БЛАЖ	
955	блиск	блис, блищ,	
893	БЛОК1	БЛОЧ	
894	БЛОК2	БЛОЧ	
956	блуд1-ити	блудж	
957	блюв-ати	бльов	

Рис.2.6. Фрагмент БД аломорфічних коренів

Інформація про аломорфію висвітлюється в інтерфейс-картці в полі (11) за заданим коренем, і в такий спосіб укладач має доступ до інформації про аломорфію, щоб правильно приписати інваріантний кореневий морф у полі (6) з урахуванням омонімії та аломорфії, а потім ця інформація приписується до слова в БД морфемних структур, наприклад:

багатобальний,А,RFSGRKSLFN/багат2, бал1/
бальний,А,RESFFH/бал2/
бальний1,А,RESFFH/бал1/
бальність,К,RESFSJFK/бал1
двобальний,А,RCIDRHSIFK/дв, бал1/
дев’ятибальний,А,RGIHRLSMFO/девjать, бал1/
десятибальний,А,RFIGRKSLFN/десять, бал1/
небальний,А,PCRGSHFJ/бал2/

однобальний,А,RDIERISJFL/один,бал/
п’ятибальний,А,REIFRJSKFM/пjать,бал1/
семибальний,А,RDIERISJFL/сім, бал1/
стобальний,А,RCIDRHSIFK/ст, бал1/
трибальний,А,RCIDRHSIFK/тр, бал1/
чотирибальний,А,RFIGRKSLFN/чотир, бал1/
шестибальний,А,REIFRJSKFM/шість, бал1/

Таким чином, усім словам із коренем -*бал*- та -*баль*- приписується інформація про інваріантний ідентифікаційний морф, яка дозволяє автоматично прокласифікувати слова з омонімічними коренями в різні спільнокореневі вибірки, і навпаки, об’єднати в одну спільнокореневу вибірку слова з різними аломорфами однієї кореневої морфеми. Наприклад:

Спільнокоренева вибірка слів із коренем БАЛ1 :

бал1,Й,RDFE/бал1/
баловий,А,RDSFFH/бал1/
багатобальний,А,RFSGRKSLFN/багат2, бал1/
бальний1,А,RESFFH/бал1/
бальність,К,RESFSJFK/бал1
двобальний,А,RCIDRHSIFK/дв, бал1/
дев’ятибальний,А,RGIHRLSMFO/девjать, бал1/
десятибальний,А,RFIGRKSLFN/десять, бал1/
однобальний,А,RDIERISJFL/один,бал/
п’ятибальний,А,REIFRJSKFM/пjать,бал1/
семибальний,А,RDIERISJFL/сім, бал1/
стобальний,А,RCIDRHSIFK/ст, бал1/
трибальний,А,RCIDRHSIFK/тр, бал1/
чотирибальний,А,RFIGRKSLFN/чотир, бал1/
шестибальний,А,REIFRJSKFM/шість, бал1/

Спільнокоренева вибірка слів із коренем БАЛ2:

бал,Й,RDFE/бал2/
балувати,Г,RDSGFI/бал2/
балювати,Г,RDSGFI/бал2/
балюватися,Г,RDSGFIXK/бал2/
збалувати,Г,PBRESHFJ/бал2, бал3/
збалуваний,А,PBRESHSIFK/бал2, бал3/
бальний,А,RESFFH/бал2/
небальний,А,PCRGSHFJ/бал2/

Спільнокоренева вибірка слів із коренем БАЛ3:

баловство,Л,RDSFSIFJ/бал3/
балощі,И,RDSFFG/бал3/
балуватися,Г,RDSGFIXK/бал3/

балувати1,Г,RDSGFI/бал3/
балуваний,А,RDSGSHFJ/бал3/
збалувати,Г,PBRESHFJ/бал2, бал3/
збалуваний,А,PBRESHSIFK/бал2, бал3/
набалуватися,Г,PCRFSIFKXM /бал3/
небалуваний,А,PCRFSISJFL/бал3/
побалуватися,Г,PCRFSIFKXM/бал3/
побалувати,Г,PCRFSIFK/бал3/
розбалувати,Г,PDRGSJFL/бал3/
розбалуватися,Г,PDRGSJFL/бал3/
розбалуваний,А,PDRGSJSKFM/бал3/
розбалувано,@,PDRGSJSKSL/бал3/

2.3. Словотвірна база даних АСМСА: структура, функції та процедура укладання

Схема даних АСМСА (рис. 2.1) показує, що словотвірна база даних (СБД) об'єднує дві БД-таблиці: (1) БД вибірок спільнокореневих слів та (2) БД словотвірних гнізд. Розглянемо процедуру автоматичного / автоматизованого формування таблиць СБД на прикладі укладання словотвірного гнізда із коренем *-голод-*.

БД вибірок спільнокореневих слів укладається автоматично за даними трьох таблиць МБД АСМСА: БД морфемних структур слів; БД омонімічних коренів; БД аломорфічних коренів. В укладанні БД вибірок спільнокореневих слів ставиться завдання згрупувати слова (за реєстром БД морфемних структур слів) у спільнокореневі вибірки з врахуванням омонімії (БД омонімічних коренів) та аломорфії (БД аломорфічних коренів) коренів. Із використанням прикладної програми-конструктора СКБД, написаної мовою програмування С++ (програміст В. Сорокін), робляться запити до трьох БД-таблиць МБД й автоматично формується БД вибірок спільнокореневих слів.

П е р ш и й з а п и т програмується до БД омонімічних коренів, яка систематизує інформацію про омонімію кореневих морфем. За даними цієї бази даних, корінь *-голод-* не є омонімічною морфемою, тобто у БД нема кореневих морфів *-голод1-* або *-голод2-*. Ця інформація є важливою умовою подальшої класифікаційної процедури.

Д р у г и й з а п и т . Якщо корінь не є омонімічним, пошук слів із цим коренем у БД морфемних структур слів здійснюється тільки за одним кореневим морфом, наприклад, за фільтром поля "Тип морфеми" (R) та фільтром поля "Морфема" (задана морфема *-голод-)* БД морфемних структур програма автоматично формує список слів, у яких визначено цей корінь:

1. виголодатися,Г,PCRHSIFKXM
2. виголодніти,Г,PCRHSISJFL
3. виголоднілий,А,PCRHSISJSKFM
4. виголодуватися,Г,PCRHSKFMXO
5. голод,Й,RFFG
16. голодування,Л,RFSISKFL
17. голодуючий,А,RFSHSJFL
18. зголодніти,Г,PBRGSHSIFK
19. зголоднілий,А,PBRGSHSISJFL
20. зголодніло,Н,PBRGSHSISJSK

6. голодний,А,RFSGFI
7. голоденький,А,RFSGSKFM
8. голоднеча,К,RFSGSIFJ
9. голодніший,А,RFSGSIFK
10. голодніше,Н,RFSGSISJ
11. голодно,Н,RFSGSH
12. голодовка,К,RFSHSIFJ
13. голодомор,Й,RFIGRJSKFL/голод, мор2/
14. голодоморний,А,RFIGRJSKFM/голод, мор2/
15. голодувати,Г,RFSIFK
21. ізголодніти,Г,PCRHSISJFL
22. наголодуватися,Г,PCRHSKFMXO
23. найголодніший,А,PDRISJSLFN
24. напівголодний,А,PFRKSLFN
25. неголодний,А,PCRHSIFK
26. неголодно,Н,PCRHSISJ
27. поголодніти,Г,PCRHSISJFL
28. поголодувати,Г,PCRHSKFM
29. приголодніти,Г,PDRISJSKFM
30. проголодатися,Г,PDRISJFLXN
31. проголоднітися,Г,PDRISJSKFMXO
32. проголодувати,Г,PDRISLFN

Якщо корінь омонімічний, то пошук слів з омонімічними коренями здійснюється у два етапи:

1-ий етап: спочатку за тими самими опціями, що й у другому пошуковому запиті, формується лексична вибірка слів, у яких визначено кореневі морфеми без розмежування омонімії, наприклад:

бал,Й,RDFE/бал2/
бал1,Й,RDFE/бал1/
балощі,И,RDSFFG/бал3/

2-ий етап: за даними, дописаними до процедури морфемної сегментації слова через скісну /бал1/, /бал2/, /бал3/, здійснюється автоматична класифікація лексики в межах сформованої вибірки на три списки[20].

Третій пошуковий запит робиться до визначеного кореневого морфа -*голод*- (а в омонімічних коренях до кожного кореневого морфа-омоніма) у БД аломорфічних коренів, яка складається із двох полів: "Морф", "Аломорфи", наприклад:

Морф	**Аломорфи**
голод	голодь, голодж

У результаті другого пошукового запиту визначаються три аломорфи морфеми -ГОЛОД- (-*голод*-, -*голодь*-, -*голодж*-), що свідчить про наявність у БД морфемних структур спільнокореневої лексики із графемно різними кореневими аломорфами. Лексична вибірка з ідентифікаційним аломорфом -*голод*- уже сформована, тому необхідно знайти в МБД слова з аломорфом -*голодь*- та слова з аломорфом -*голодж*-. За двома аломорфами поля "Аломорфи" -*голодь*- та -*голодж*- робиться четвертий пошуковий запит до БД морфемних структур, за яким формуються ще дві лексичні спільнокореневі вибірки:

-голодь-	**-голодж-**
впроголодь,Н,PBPERKSK/голод/	виголоджуватися,Г,PCRISLFNXP/голод
надголодь,Н,PDRJSP/голод/	
обголодь,Н,PCRISI/голод	
упроголодь,Н,PBPERKSK/голод/	

[20] У попередньому параграфі подано повний список слів із омонімічними коренями бал1 *(одиниця виміру)*; бал2 *(танцювальний вечір)*; бал3 *(балувати)* та описано принципи розмежування омонімії.

У такий спосіб формується спільнокоренева вибірка лексики, яка записується у БД вибірок спільнокореневих слів в окреме поле, у якому подано реєстр спільнокореневих слів із збереженою інформацією про: код-номер (ID) слова у МБД, програмну процедуру морфемної будови слова, морфонологічні варіанти кореня, омонімічний код кореня (додаток 2.1).

У межах кожної спільнокореневої вибірки проводиться автоматична класифікація слів за кількістю морфем в основі слова без врахування флексії. Програма-конструктор рахує символьні позначення морфів (перший символ двосимвольного коду) в програмній процедурі (наприклад, *виголодатися*, **P**C**R**H**S**IFK**X**M – 4 морфи; *виголодніти*, **P**C**R**H**SIS**JFL – 4 морфи) і формує в межах вибірки спільнокореневої лексики групи слів з однаковою кількістю морфем (за однауовою кількісно-морфемною моделлю основи слова: 1 морфема, 2 морфеми...).

Враховуючи формальні принципи словотвірної похідності: 1) морфологічні способи словотвору передбачають кількісно-афіксальне зростання морфемної структури мотивованої основи словотвірної пари; 2) інтерфікси не вважаються словотвірними формантами і додаються у процесі словотвору до словотвірних суфіксів та префіксів, було сформульовано робочу гіпотезу процедури автоматичного укладання словотвірного гнізда. Ця гіпотеза визначає правила автоматичної класифікації спільнокореневих слів за тактами словотвірного гнізда: 1) слово з найменшою кількістю морфем становить вершину словотвірного гнізда і належить до нульового такту; 2) кожен наступний словотвірний такт репрезентує слова з кількісно складнішими афіксальними структурами основ, ніж попередній словотвірний такт; 3) складні слова належать до першого такту словотвору.

Кожна група спільнокореневих слів однієї кількісно-морфемної моделі вважається гіпотетичними тактом словотвірного гнізда: кожен наступний такт відображає збільшення кількості морфем в основі слова (табл. 2.1).

Таблиця 2.1. Класифікація спільнокореневої лексики за кількістю морфем в основі слова

0-ий такт: 1 морф основи слова	**1-ий такт: 2 морфи основи слова**	**2-ий такт: 3 морфи основи слова**	**3-ій такт: 4 морфи основи слова**	**4-ий такт: 5 морфів основи слова**	**5-ий такт: 6 морфів основи слова**
голод	голодний голодомор голодоморний голодувати	голодненький голоднеча голодніший голодно голодовка голодування голодуючий напівголодний неголодний поголодувати проголодувати	виголодатися виголодніти виголодуватися голодніше зголодніти ізголодніти наголодуватися найголодніший неголодно поголодніти приголодніти проголодатися	виголоднілий зголоднілий	зголодніло

На матеріалі такої класифікаційної моделі автоматично будується у вигляді дерева залежностей словотвірне гніздо, яке є тільки робочою гіпотезою-моделлю, що вимагає перевірки та редагування (рис. 2.7).

Редагування здійснюється лінгвістом автоматизовано за допомогою інтерфейс-картки morfem.voc, у якій кнопка *Tree* дозволяє зробити перехід від МБД до словотвірної СБД чи навпаки. У полі (9) *Аналізований корінь* уводиться інваріантний кореневий аломорф, а потім натискається кнопка *Find*. Ця кнопка забезпечує зв'язок робочої картки із БД словотвірних гнізд: у полі (5: великий екран) подається у вигляді дерева залежностей робоча гіпотеза-модель словотвірного гнізда (рис. 2.7).

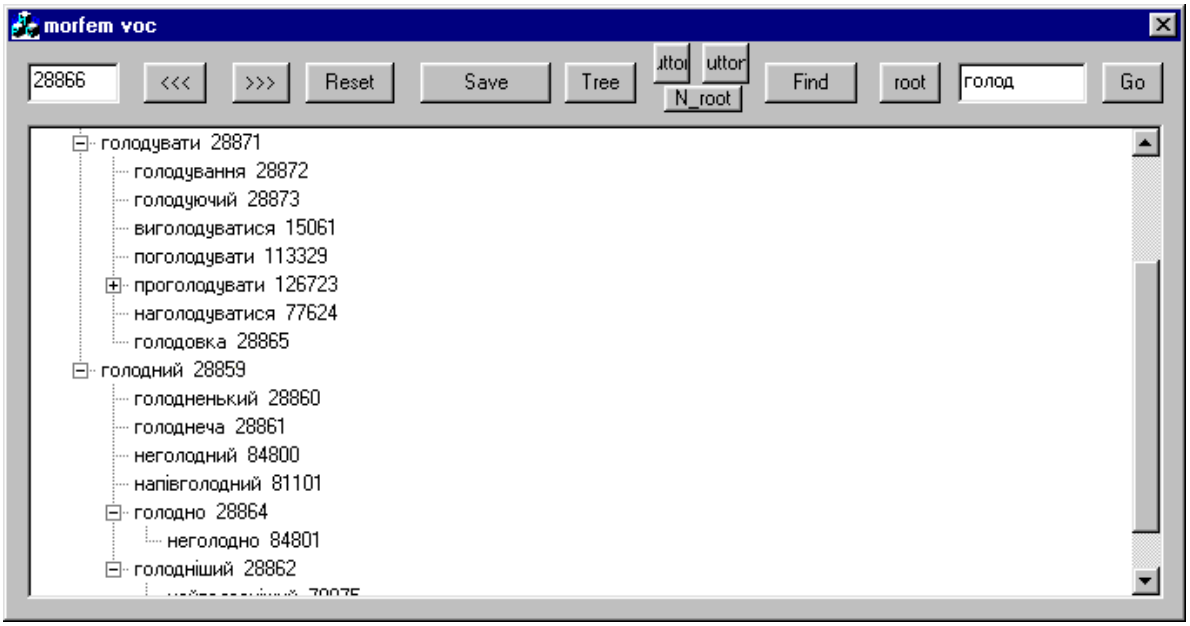


Рис. 2.7. Фрагмент словотвірного гнізда із коренем *голод*

Вихідним словом дерева, що є вершиною словотвірного гнізда, виступає найменше, за кількістю морфем в основі, слово (*голод* – 1 морф), яке є першим вузловим словом (знак <+>), тобто від нього відходить гілка, яку можна розгорнути. Кожна гілка словотвірного дерева відображає зв'язки формальної словотвірної мотивації між вузловим словом і словами, що закінчують гілки цього вузла. Знак <+> вказує, що слово вузлове, тобто від нього відходить гілка, а знак <–> вказує, що ця гілка вже розгорнута.

Автоматична класифікація слів спільнокореневої вибірки за словотвірними тактами на основі формальних словотвірних зв'язків є лише лінгвістичною гіпотезою, яка вимагає перевірки. На цьому етапі робота проводиться лінгвістом, який використовує свої знання і редагує словотвірне гніздо, враховуючи семантичні словотвірні відношення. У разі неправильно визначеної словотвірної похідності, можна "перетягнути" виділене курсором слово з однієї гілки словотвірного гнізда в іншу: перенести його в потрібний вузол. Дані про редаговане словотвірне гніздо зберігаються натисканням кнопки *Save* в робочій картці (рис. 2.7) і автоматично імпортуються в БД словотвірних гнізд, де словотвірні відношення між словами зберігаються через встановлення відповідності між цифровими кодами слів у АСМСА.

Наприклад, словотвірна мотивація між словами *голодувати → голодування* (28871 – 28872), *голодувати → поголодувати* (28871 – 113329) у БД словотвірних гнізд буде мати такий запис:

28871	28872
28871	113329

2.4. АСМСА – морфемна база знань: систематизація лінгвістичної інформації про організацію афіксальної системи української мови

На другому етапі конструювання МБД було поставлено завдання систематизувати лінгвістичні дані про афіксальні морфеми як інваріантно-варіантні знакові одиниці мови: представити парадигматичні та синтагматичні відношення кожного морфа афіксальної морфеми. Це завдання виконувалося з метою автоматичного групування лексики в спільноафіксальні вибірки з урахуванням омонімії та аломорфії афіксів й автоматичного укладання за цими вибірками електронного тлумачно-морфонологічного словника афіксальних морфем.

За розробленою концепцією інфологічної моделі (див. §1.5.1) електронного тлумачно-морфонологічного словника морфем, реєстровою одиницею словника виступає афіксальна морфема, а словникова стаття містить інформацію про:

1) парадигматичну цілісність морфеми, що передбачає групування всіх морфів реєстрового афікса;

2) реалізацію кожного морфа у вибірці спільноафіксальних слів, укладеною за лексичним реєстром БД морфемних структур;

3) значення афіксальної морфеми, що визначається як інтегральна сема семантичного поля слів, у яких реалізовується описувана афіксальна морфема;

4) словотвірну / структурну функцію кожного морфа морфеми;

5) дистрибутивну позицію кожного морфа в морфемній структурі спільноафіксальних слів;

6) морфонологічні процеси, які спричинили аломорфію (варіативність) афіксальної морфеми.

Відповідно до поставлених завдань було визначено технічне завдання про укладання у межах МБД ще чотирьох баз даних (рис.2.1):

- БД афіксальних морфем (у процесі укладання);
- БД мотивувальних слів(у процесі укладання);
- БД морфонологічних процесів (у процесі укладання);
- БД лексичних тлумачень.

На цьому етапі СКБД морфемної бази даних, яка мала раніше локальний характер і файлову систему збереження даних, було змінено на систему керування веб-базою даних: лінгвіст-укладач через систему авторизації на інтернет-порталі mova.info має доступ до всіх даних бази, може змінювати і зберігати дані (крім структури даних), працюючи в мережі

Інтернет. СКБД забезпечує паралельну одночасну роботу з даними багатьох укладачів, централізоване збереження даних, та систему контролю за збереженими даними різними укладачами.

Перші три БД укладаються лінгвістом автоматизовано в режимі on-line за допомогою інтерфейс-картки morfem.exe, структура та функції якої були розширені у зв'язку з поставленими завданнями (див. Рис.2.8).

Вихідними в роботі системи виступають лінгвістичні дані БД морфемних структур, які представлені на рис. 2.8 у 1-ій та 3-ій зоні картки. Зона 1:

(1) випадний список, у якому вибирається функціональний тип морфеми: S – суфікс;

(2) поле – *аналізований морф*, у якому записується аналізований морф: *тв*;

(3) кнопка <<< – *лівобічна навігація*: повернутися до попереднього слова в МБД;

(4) поле – *аналізоване слово*: висвітлюється слово, у якому реалізований, записаний у 2-му полі, морф, за навігацією по реєстру МБД або за записом слова у полі (5): *баштанництво*;

(6) кнопка >>> – *правобічна навігація*: перейти до наступного слова, у якому реалізується вибраний морф;

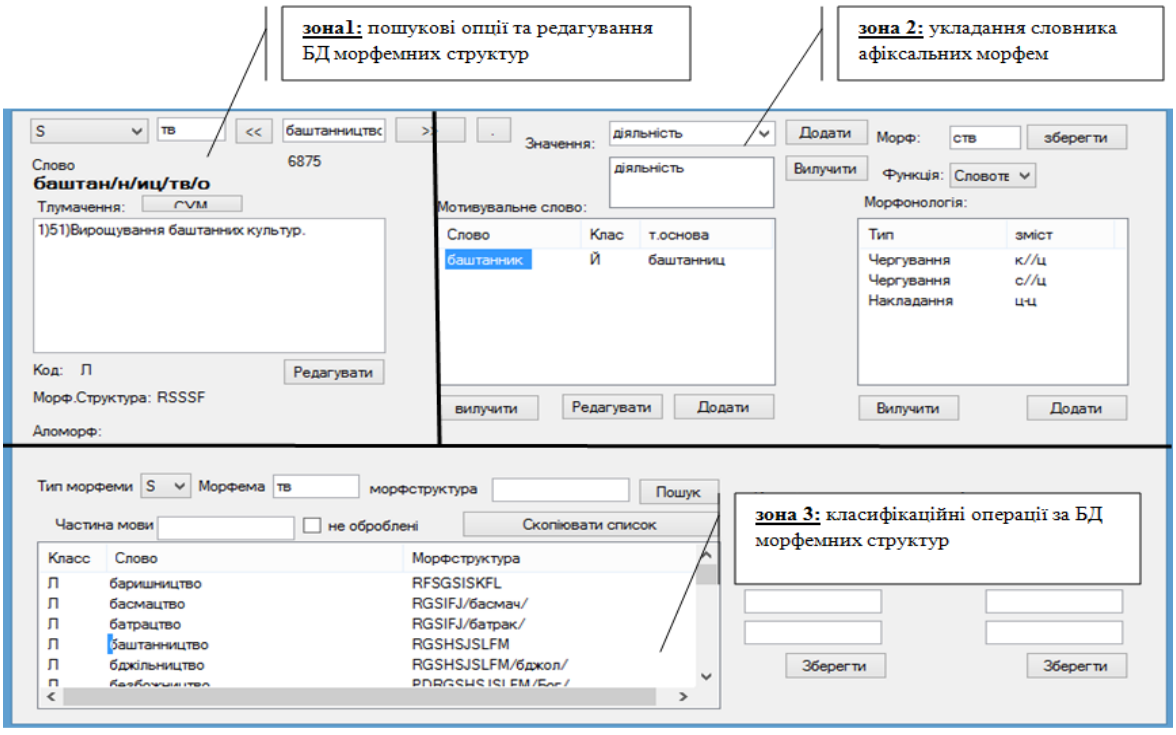


Рис. 2.8. Інтерфейс робочої картки укладання БД електронного словника афіксальних морфем

За заданими пошуковими параметрами (1) і (2) в зоні 1 висвітлюється інформація про:

- номер (цифровий код) аналізованого слова в МБД*: 6875;*

- морфемну структуру аналізованого слова: *баштан/н/иц/тв/о*;
- граматичний код: *Л* – іменник середнього роду;
- функціональну морфемну модель слова: RSSSF;
- аломорф: *інваріантний морф кореневої морфеми, за умови омонімії чи аломорфії кореневої морфеми аналізованого слова*.

(7) кнопка *Редагувати*: активізує окреме діалогове вікно (рис.2.9), у якому можна автоматизовано визначити морфемну будову аналізованого слова, за якою комп'ютер автоматично формує програмну процедуру: ця процедура виконується за умови неправильно визначеної морфемної будови слова або введення до МБД нової лексики;

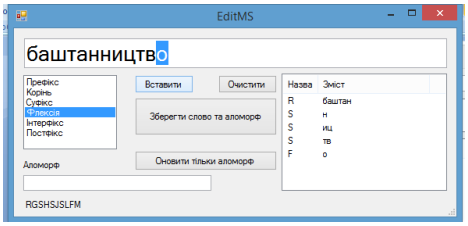


Рис. 2.9. Діалогове вікно редагування морфемної структури слова

У 1-ій зоні робочої картки подається поле (8) *Тлумачення*, у якому висвітлюються тлумачення до аналізованого слова: *Вирощування баштанних культур*. Це поле поєднане із компілятивною БД лексичних тлумачень, реєстр якої укладався за трьома тлумачними словниками: «Великим тлумачним словником сучасної української мови» [ВТССУМ 2005]; «Словником української мови» [СУМ 1970–1980], «Словником іншомовних слів» [СІС 1985]. Уведення цієї БД до структури МБД АСМСА зумовлювалося необхідністю проведення компонентного аналізу лексичного значення аналізованого слова з метою визначення семантики описуваного афікса. Кнопка *СУМ* над полем із тлумаченнями слів здійснює автоматичний зв'язок аналізованого слова МБД із словниковою статтею цього слова в електронній версії «Словника української мови» [ОВСУМ 2018]. Таким чином, укладач має можливість, за потреби, отримати повну інформацію про лексичне значення слова з ілюстративним текстовим матеріалом. БД лексичних тлумачень має інформаційно-довідковий характер і є закритою для укладача: у ній зберігаються оголошені дані, які не можуть змінюватися й не потребують збереження.

Зона 2. Укладання БД афіксальних морфем, БД мотивувальних слів, БД морфонологічних процесів здійснюється за параметрами 2-ої зони робочої картки (рис.2.8).

БД афіксальних морфем укладається за процедурою приписування кожному аналізованому афіксальному морфу в конкретному слові; значення (семантичного індексу); інваріантного ідентифікаційного морфа; функції.

Кожен із цих параметрів приписується лінгвістом автоматизовано за полями робочої картки, а лінгвістичні дані автоматично зберігаються в окремих полях (колонках) БД-таблиці через активацію кнопки (9) *Зберегти*.

У 2-ій зоні робочої картки (рис.2.8) за укладання БД афіксальних морфем відповідають такі опції:

(10) поле *Значення*: у першому вікні обирається значення афікса за випадним списком, сформованим за індексом значень омонімічних суфіксальних одиниць «Словника афіксальних морфем української мови» [САМУК 1998], або записується нове значення; у другому вікні це значення висвітлюється кожен раз, коли активується слово, до якого це значення суфікса приписано; наприклад, суфікс *-тв-* у слові *баштанництво* має значення ’діяльність’, а у слові *клятва* ’опредметнена дія’; кнопки (11) *Додати* та (12) *Вилучити* дозволяють редагувати випадний список;

(13) поле *Морф*: у вікні записується умовно визначений інваріантний аломорф, який індексує морфему як інваріантну одиницю, наприклад, суфіксальні морфи *-тв-* та *-ств-* є аломорфами однієї морфеми -СТВ- ’діяльність’, тому у слові *баштанництво* морфу *-тв-* морфеми -СТВ- у цьому полі приписується ідентифікаційний морф *-ств-*, на форму якого не вплинули морфонологічні процеси; а у слові *клятва* морфу *-тв-* приписується ідентифікаційний морф -ТВ- ’опредметнена дія’;

(14) поле *Функція*: у випадному списку вибирається функція, яку афікс виконує в аналізованому слові (словотвірна або структурна), наприклад, суфікс *-тв-* у слові *баштанництво* виконує словотвірну функцію, а у слові *клятвений* – структурну, тому що це слово утворене за допомогою афікса -ен-.

БД мотивувальних слів та БД морфонологічних процесів взаємопов’язані за логікою проведення аналізу: для правильного визначення морфонологічних альтернацій, які відбулися з морфом у похідному слові, необхідно визначити твірну основу і мотивувальне слово. Тому в робочій картці (рис. 2.8) спочатку заповнюється інформація у полі (15) *Мотивувальне слово***:**

1) визначається мотивувальне слово, на базі якого утворюється аналізоване похідне, наприклад, до слова *баштанництво* записується мотивувальне слово *баштанник*;

2) записується граматичний код мотивувального слова, тому що частиномовна характеристика мотивувального системно визначає тип морфонологічних процесів, наприклад, для слова *баштанник* – Й (іменник чоловічого роду);

3) визначається твірна основа у мотивованому похідному слові, наприклад, *баштанниц-*.

Систематизована у такий спосіб інформація зберігається у БД мотивувальних слів. На основі зіставлення мотивувального слова з твірною основою визначаються типи та моделі морфонологічних процесів, які записуються у полі (16) *Морфонологія*:

(17) кнопка *Додати* активізує вікно (18) *додати морфонологію* (рис.2.10), у якому у випадному списку вибирається тип морфонологічного процесу, а у вільному полі записується морфонологічна модель.

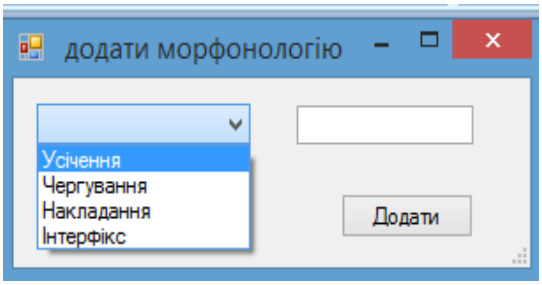


Рис.2.10. Діалогове вікно запису морфонологічних процесів

Записані дані зберігаються у двох колонках таблиці БД морфонологічних процесів і висвітлюються в робочій картці у полі (19) *Морфонологія* (рис.2.8), наприклад, морфонологічні процеси, що зумовлюють утворення аломорфа *-тв-* у слові *баштанництво*: чергування к//ц; чергування с//ц; накладання ц-ц.

Зона 3. У 3-ій зоні робочої картки (рис.2.8) лінгвіст-укладач може здійснювати різноманітні класифікаційні пошуки за параметризованими даними БД морфемних структур і автоматично укладати лексичні вибірки, кнопка (20) *Пошук*, за заданими класифікаційними ознаками:

(21) *Тип морфеми*: за вибраним із випадного списку, функціональним типом морфеми: *S*;

(22) *Морфема*: за записом аналізованого морфа: *тв*;

(23) *Морфструктура*: за записом моделі морфемної структури: *PRSF*;

(24) *Частина мови*: за записом граматичного коду частини мови.

СКБД автоматично формує лексичні вибірки за різними параметрами залежно від кількості обраних класифікаційних ознак. За заданими на рис. 2.8 класифікаційними ознаками: функціональний тип морфеми (S), аналізований морф (тв) – формується вибірка спільноафіксальної лексики обсягом 364 слова (додаток 3), у якій слова подані за інформаційними параметрами даних БД морфемних структур слів, наприклад:

антимистецтво,Л,PERISKSMFN/;
бавовництво,Л,RGSISKFL;
баришництво,Л,RFSGSISKFL;
басмацтво,Л,RGSIFJ/басмач/;
батрацтво,Л,RGSIFJ/батрак/;
баштанництво,Л,RGSHSJSLFM;
бджільництво,Л,RGSHSJSLFM/бджол/;
безбожництво,Л,PDRGSHSJSLFM/Бог/;
бешкетництво,Л,RGSHSJSLFM.

За класифікаційними ознаками: функціональний тип морфеми (S), аналізований морф (*ств***)** – формується вибірка спільноафіксальної лексики обсягом 1146 слів, наприклад:

робочедільство,Л,RDSFIGRKSNFO/роб1, діл2/;
родичівство,Л,RDSFSHSKFL/рід/;

розглагольствувати,Г,PDRKSNSQFS/глагол/;

розглагольствування,Л,PDRKSNSQSSFT/глагол/;

За класифікаційними ознаками: функціональний тип морфеми (S), аналізований морф (*ств***)**, морфемна структура (*PRIRSF)* – формується вибірка спільноафіксальної лексики з 2-ох слів:

інакодумство,Л,PCREIFRISLFM;

неблагородство,Л,PCRGIHRKSNFO//.

За класифікаційними ознаками: функціональний тип морфеми (*S***),** аналізований морф (*ств***)**, морфемна структура (*RSSF*), частина мови (*Г* – дієслово) – формується вибірка спільноафіксальної лексики з 34-ох слів, наприклад:

акторствувати,Г,RFSISLFN;

бузувірствувати,Г,RHSKSNFP;

буйствувати,Г,RDSGSJFL/буй2/;

витійствувати,Г,RFSISLFN.

Таким чином, укладач БД афіксальних морфем має можливість досліджувати дистрибутивно-позиційну та семантичну реалізацію аналізованого афіксального морфа в усіх спільноафіксальних словах, згрупованих за різними класифікаційними ознаками.

Лінгвістичні дані, збережені в трьох базах даних (БД афіксальних морфем, БД мотивувальних слів, БД морфонологічних процесів), можуть бути автоматично згенеровані в єдину БД електронного словника афіксальних морфем. БД електронного словника суфікса -СТВ- (-тв-;-ств-) у значенні ’абстрактне поняття / діяльність’ систематизує лінгвістичні дані, представлені у таблиці 2.2.

Таблиця 2.2. Фрагмент БД словника суфікса -СТВ- (-тв- /-ств-) у значенні ’абстрактне поняття / діяльність’

морф	**слово**	**морфструктура**	**значення**	**мотивувальне слово**	**усічення**	**накладання**	**чергування**
тв	бавовництво	RSSF	діяльність	бавовник		ц-ц	к//ц, с//ц
тв	баришництво	RSSSF	діяльність	баришник		ц-ц	к//ц, с//ц
тв	батрацтво	RSF	діяльність	батрак		ц-ц	к//ц, с//ц
тв	баштанництво	RSSSF	діяльність	баштанник		ц-ц	к//ц, с//ц
тв	бджільництво	RSSSF	діяльність	бджільник		ц-ц	о//і, к//ц, с//ц

За даними БД морфемних структур слів суфіксальний морф -тв- (без розведення омонімії) реалізовується у 364 словах (додаток 3), за даними БД електронного словника афіксальний морф -тв- морфеми -СТВ- у значенні ’абстрактне поняття / діяльність’ реалізовується у 125 словах (додаток 4). БД електронного словника систематизує лінгвістичні дані про:

- слово, у якому реалізовується суфікс -*тв*-;
- морфемну модель слова;
- значення суфікса;

- мотивувальне слово;
- морфонологічні процеси, які спричинили аломорфію морфа *-тв-*: чергування і його моделі, накладання і його моделі.

На сьогодні описані бази даних перебувають на стадії формування і накопичення знань про кожний афіксальний морф у кожному слові МБД. Ведеться робота над параметризацією суфіксальної зони, а префіксальна зона ще не оброблялася. Досвід укладання автономної БД словника афіксальних морфем за параметрами систематизації лінгвістичних даних у МБД АСМСА розкриває вагомість цих даних для лінгвістичної предметної галузі, їх значущість для опису морфемної та словотвірної системи української мови й визначає систему АСМСА як лінгвістичну базу знань.

2.5. АСМСА – морфемна база знань: систематизація лінгвістичної інформації про організацію кореневої системи української мови

Третій етап створення МБЗ АСМСА був спрямований на виконання двох завдань: 1) уведення нової лексики до БД морфемних структур; 2) систематизації даних про кореневі морфеми з метою створення тлумачно-морфонологічного словника кореневих морфем[21].

Початково БД морфемних структур укладалася за реєстром українських слів обсягом ≈ 170 тис. одиниць (див. § 2.1). Лексичний реєстр початкових форм резидентного словника морфологічного автоматичного аналізу тексту, який використовується в Корпусі української мови, постійно доповнювався новою лексикою, яка систематизувалась у МБД в окремому lex-файлі необробленої лексики, що складала алфавітний список ≈ 35 тис. слів. Активація лексичного реєстру необробленої лексики здійснювалась у робочій картці за допомогою кнопки "необроблена" (див. опис зони 3 § 2.4). До цього списку потрапляли неологізми; регулярні деривати, не подані системно у словниках; запозичення; власні назви; скорочення; загальновживана лексика, що з різних причин не ввійшла початково до реєстру; а також слова із помилками. Відповідно постало завдання – вибрати із цього списку лексику, яка необхідна для повного опису морфемної системи української мови, і приписати цим словам інформацію про морфемну структуру за попередньо визначеними параметрами (див. § 2.2). Уведення необробленої лексики до БД морфемних структур здійснювалось автоматизовано за робочою карткою morfem.exe (див. опис зони 1 § 2.4). На сьогодні обсяг БД морфемних структур слів АСМСА становить ≈ 200 тис. слів.

З метою автоматичного укладання електронного словника кореневих морфем було поставлено завдання – описати семантику кореневих морфем та верифікувати дані трьох баз даних: БД морфемних структур, БД омонімічних коренів, БД аломорфічних коренів, структура яких описана у § 2.2. Оскільки

[21] Зовнішня лексикографічна модель інтерактивного інтегрального тлумачно-морфонологічно словника кореневих морфем описана у §1.5.2.

на першому етапі укладання МБД зберігання даних здійснювалося за файловою системою і укладачі не мали централізованого доступу до баз даних, то приписування ідентифікаційної кореневої морфеми словам з омномічними та аломорфічними коренями було проведено з порушенням принципу системності: не було дотримано послідовності цифрового кодування коренів-омонімів, а відповідно, і коренів-аломорфів; не всім словам БД морфемних структур був приписаний інваріантний індекс кореневої морфеми.

За концепцією інфологічної моделі електронного словника кореневих морфем (див. § 1.5.1), коренева морфема представляє парадигматичний клас морфів, визначений за інваріантно-варіантними ознаками морфеми як цілісної знакової одиниці мови, що враховує спільне значення класу формально тотожних або частково тотожних за формою морфів. Таке розуміння морфеми зумовлює поділ словникової статті на шість зон:

1) реєстрова одиниця словника – вихідний кореневий морф, той, який реалізований у непохідному слові спільнокореневих слів, наприклад, вихідним морфом у спільнокореневих словах *берег/ти, береж/інн/я, енерг/о/з/беріг/aj/уч/ий* визначається дієслівний кореневий морф -берег-;

2) непохідне слово із визначеним вихідним морфом: *берегти* ;

3) частина мови непохідного слова, яка визначає категорійну семантику морфеми: Г;

4) тлумачення непохідного слова: лексичні значення всіх ЛСВ *берегти*;

5) варіанти морфеми: аломорфи, варіоморфи;

6) спільнокореневі слова, у яких реалізовані варіанти морфеми.

Автоматичне конструювання словникової статті планується здійснити за даними трьох баз даних МБД (БД морфемних структур слів, БД омонімічних коренів, БД аломорфічних коренів), які програмно взаємопов’язані й систематизують лінгвістичні дані для формування кожної зони словникової статті.

З метою розбудови і редагування баз даних було створено новий інтерфейс робочої картки morfem.exe (рис. 2.11), з якою лінгвіст-укладач працює в автоматизованому on-line режимі. До робочої картки додались три нові зони (2), (4) і (5), а зони (1) та (3) структурно і функціонально не змінились (див. § 2.4).

Зона (2) призначена для укладання та редагування БД омонімічних коренів та БД аломорфічних коренів, а (4) та (5) – для користування цими базами даних: пошук за записом у полі *Перші літери аломорфа або омоніма* кореня, частини кореня чи початкової букви.

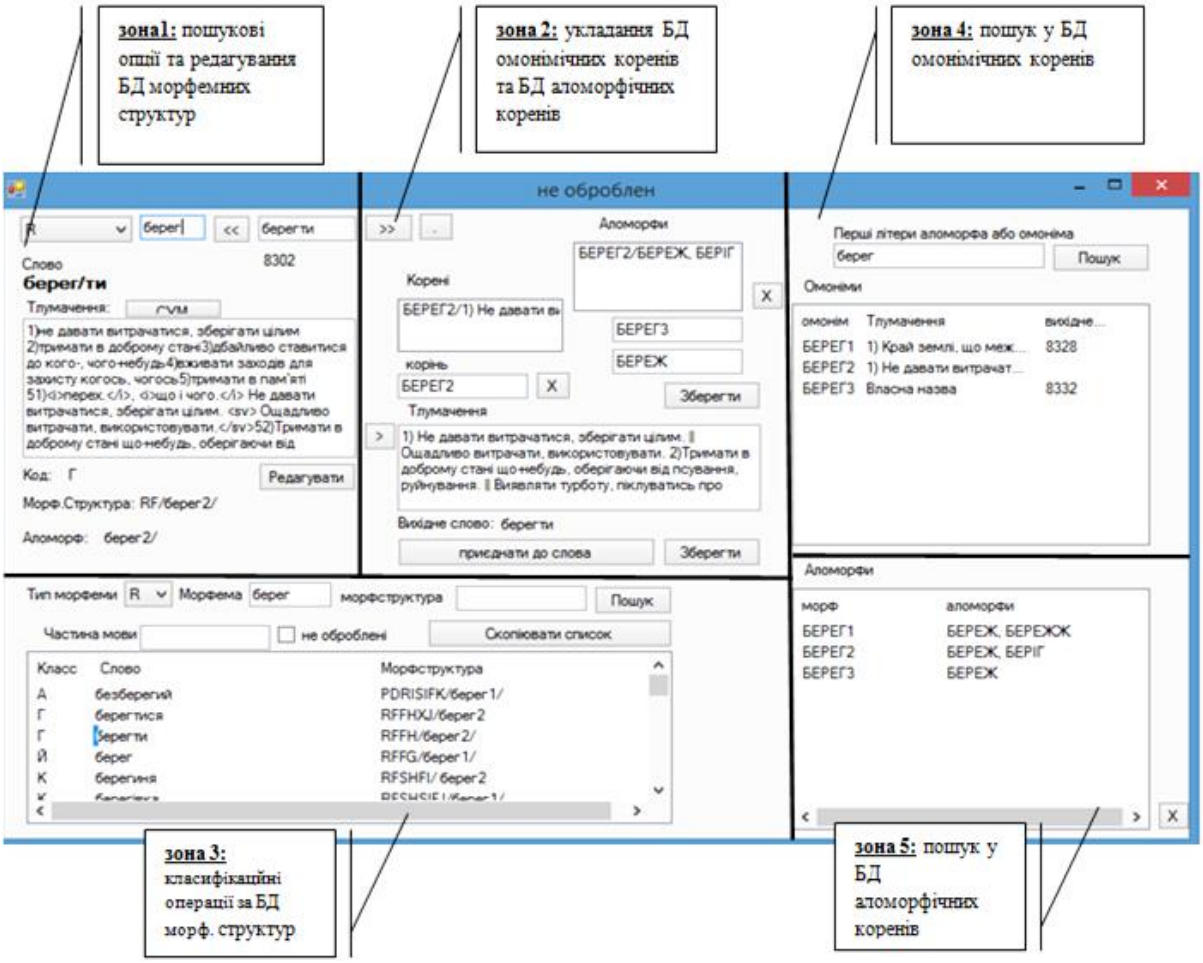


Рис. 2.11. Інтерфейс робочої картки редагування БД омонімічних коренів та БД аломорфічних коренів

На рис. 2.11 показано результати пошуку кореня -берег-. За даними МБД, морфемна система української мови має три кореневі морфеми -берег-:

1. БЕРЕГ1 – семантика: *1) Край землі, що межує з поверхнею річки, озера, моря і т. ін. 2) Суходіл, територія, що прилягає до річки, озера, моря і т. ін. 3) Те, що по краях смугою обрамовує що-небудь*; аломорфи: БЕРЕГ, БЕРЕЖ, БЕРЕЖЖ;

2. БЕРЕГ2 – семантика: *1) Не давати витрачатися, зберігати цілим. || Ощадливо витрачати, використовувати. 2)Тримати в доброму стані що-небудь, оберігаючи від псування, руйнування. || Виявляти турботу, піклуватись про кого-небудь; доглядати. || Оберігати, охороняти кого-, що-небудь від чогось. 3) Тримати в пам'яті, пам'ятати. || Старанно підтримувати, цінити, шанувати (якусь рису, властивість, почуття і т. ін.);* аломорфи: БЕРЕГ, БЕРЕЖ, БЕРІГ;

3. БЕРЕГ3 – семантика: *Власна назва*; аломорфи: БЕРЕГ, БЕРЕЖ.

У процесі укладання електронного словника кореневих морфем ставиться те саме завдання, що й в укладанні БД словотвірних гнізд (див. § 2.3), автоматично згрупувати спільнокореневу лексику із розведенням кореневої омонімії і водночас зведенням кореневої аломорфії, щоб до спільнокореневої вибірки кореня БЕРЕГ 1 потрапили слова із кореневими

морфами *-берег-, -береж-, -бережж-* у значенні ’берег’; до спільнокореневої вибірки кореня БЕРЕГ 2 потрапили слова із кореневими морфами *-берег-, -береж-, -беріг* у значенні ’берегти’; до спільнокореневої вибірки кореня БЕРЕГ 3 потрапили слова із кореневими морфами *-берег-, -береж-* у значенні ’власна назва’. Для цього необхідно до кожного слова із цими коренями дописати формальний маркер, за яким буде проведено автоматичну класифікацію. Таким маркером виступає інваріантний індекс кореневої морфеми із цифровим кодом, наприклад, БЕРЕГ 2 приписується всім словам, у яких реалізовується сема ’берегти’ та в яких реалізовуються кореневі морфи *-берег-, -береж-, -беріг*. Графемний запис БЕРЕГ виконує функцію об’єднання всіх слів із кореневими морфами *-берег-, -береж-, -беріг, -бережж-*, а цифровий код 2 виконує функції розмежування кореневої омонімії: за ним вибираються лише слова із семою ’берегти’. Для послідовного приписування такої інформації необхідно мати алфавітний список омонімічних та аломорфічних коренів, який і представляють описувані в цьому параграфі бази даних (БД омонімічних коренів та БД аломорфічних коренів).

У к л а д а н н я Б Д о м о н і м і ч н и х к о р е н і в . Перед лінгвістом-укладачем стоїть завдання – встановити значення кореня у словах, об’єднаних в одну вибірку за спільною формою вираження кореневого морфа, приписати кожному слову з омонімічним коренем ідентифікаційний інваріантний індекс із цифровим кодом і визначити кількість кореневих морфем з однаковою формою вираження в цій вибірці. У МБД така вибірка формується за записом морфеми у зоні 3 робочої картки (рис. 2.11). За пошуковими опціями: *Тип морфеми:* R; *Морфема:* -берег- – автоматично формується вибірка із 30 слів (додаток 5), у морфемній структурі яких можуть реалізовуватися кореневі морфи в трьох значеннях:

безберегий,А,PDRISIFK/берег1/;

берегтися,Г,RFFHXJ/берег2;

берегівський,а,RFSHSKFM/берег3/.

Приклад демонструє уже оброблену вибірку спільнокореневих слів, у якій кожному слову приписано ідентифікаційний морф із цифровим кодом за опціями зони 1 (рис.: 2.8, 2.11), але попередньо необхідно було розмежувати омонімію коренів: записати в БД омонімічних коренів список коренів-омонімів із цифровим кодом, лексичним значенням непохідного слова із цим коренем та саме непохідне слово. Ці дані лінгвіст-укладач вносить до зони 2 (рис.2.11):

- якщо виділити мишею непохідне слово із аналізованим коренем, наприклад *берегти*, у списку слів зони 3, то у полі *Тлумачення* зони 1 з’являються тлумачення цього слова;
- у зоні 2, у полі *Корінь*, лінгвіст-укладач записує ідентифікаційний інваріантний індекс із цифровим кодом – БЕРЕГ2;
- кнопкою (>) лінгвіст-укладач копіює тлумачення із зони 3 у поле *Тлумачення* зони 2, де редагує запис лексичних значень;

- кнопкою *Приєднати до слова* лінгвіст-укладач приєднує до морфа-індекса із цифровим кодом номер слова у МБД;
- кнопкою *Зберегти,* у правому нижньому куті зони 2, лінгвіст-укладач імпортує введені дані до БД омонімічних коренів, які висвітлюються у зоні 4 (рис. 2.11) кожен раз, коли укладач має потребу отримати інформацію про омонімію кореня.

У к л а д а н н я Б Д а л о м о р ф і ч н и х к о р е н і в . Морфи трьох кореневих морфем БЕРЕГ1, БЕРЕГ2, БЕРЕГ3 можуть зазнавати морфонологічних альтернацій, які зумовлюють аломорфію в межах кожної кореневої морфеми. Завдання лінгвіста-укладача визначити потенційно можливі аломорфи кожної кореневої морфеми, перевірити їх за БД морфемних структур, автоматично сформувати вибірки слів із аломорфічними кореневими морфами й розмежувати кореневу омонімію в межах кожної аломорфічної вибірки, приписуючи ідентифікаційний інваріантний індекс із цифровим кодом (за БД омонімічних коренів) кожному слову з аломорфічної вибірки. У МБД такі вибірки формуються за записом морфеми у зоні 3 робочої картки (рис. 2.11).

За пошуковими опціями: *Тип морфеми:* R; *Морфема:-береж-* – автоматично формується вибірка із 78 слів (додаток 6), у морфемній структурі яких можуть реалізовуватися кореневі морфи трьох кореневих морфем у трьох значеннях (БЕРЕГ1, БЕРЕГ2, БЕРЕГ3), наприклад:

безбережність,К,PDRISJSNFO/берег1/;

бережанський,А,RFSHSKFM/берег3/;

бережений,А,RFSHFJ/берег2/.

За пошуковими опціями: *Тип морфеми:* R; *Морфема:-беріг-* – автоматично формується вибірка із 18 слів (додаток 7), у морфемній структурі яких можуть реалізовуватися кореневі морфи тільки однієї кореневої морфеми БЕРЕГ2 у значенні ’берегти’, наприклад:

вберігання,Л,PBRGSHSJFK/берег2/;

енергозберігаючий,А,RFIGPHRMSOSQFS/енерг, берег2/;

зберігатися,Г,PBRGSHFJXL/берег2/;

За пошуковими опціями: *Тип морфеми:* R; *Морфема:-бережж-* – автоматично формується вибірка із 8 слів (додаток 8), у морфемній структурі яких можуть реалізовуватися кореневі морфи тільки однієї кореневої морфеми БЕРЕГ1 у значенні ’берег’, наприклад:

безбережжя,Л,PDRJSJFK/берег1/;

крутобережжя,Л,REIFRLSLFM/крут1, берег1/;

лівобережжя,Л,RDIERKSKFL/лів1, берег1/.

Приклади демонструють результат автоматичної класифікації слів за аломорфічними коренями морфем БЕРЕГ1, БЕРЕГ2, БЕРЕГ3 за даними БД морфемних структур слів, у якій кожному слову уже приписано ідентифікаційний інваріантний індекс із цифровим кодом за опціями зони 1, але попередньо необхідно було розмежувати омонімію коренів у межах кожної аломорфічної вибірки за даними, збереженими у БД аломорфічних

коренів, таблиця якої структурується на два поля (рис. 2.11, зона 5): 1) запис кореня-омоніма: БЕРЕГ1, який також є одним із аломорфів (ідентифікаціним) морфеми БЕРЕГ1; 2) запис аломорфів цього кореня: БЕРЕЖ, БЕРЕЖЖ. Ці записи лінгвіст-укладач здійснює у правому куті зони 2 (рис. 2.11).

Параметризація лінгвістичних даних про організацію кореневої системи української мови у трьох базах даних АСМСА (БД морфемних структур, БД омонімічних коренів, БД аломорфічних коренів) демонструє значущість систематизованої інформації для предметної лінгвістичної галузі, що дозволяє визначити систему АСМСА лінгвістичною базою знань із морфемології та словотвору української мови.

Р О З Д І Л 3
КОМП'ЮТЕРНА ПАРАМЕТРИЗАЦІЯ УКРАЇНСЬКОМОВНОГО ТЕКСТУ НА МОРФЕМНОМУ РІВНІ СТРУКТУРИ

3.1. Корпус української мови – інформаційна експертна система лінгвістичного аналізу українськомовних текстів

В останні десятиріччя в центрі наукових досліджень різних сфер гуманітарних знань знаходиться текст як засіб передачі інформації, збереження знань і культури, організації соціальної комунікації, а у філології як об'єкт літературознавчих та лінгвістичних студій. Текст як основна форма збереження й передачі інформації, будучи результатом мовленнєвого акту, є інвентарем мовних одиниць, які комбінуються в ньому за законами мовної норми і утворюють систему тексту зі структурою, що відображає рівневу стратифікацію мовної системи: фонетичний рівень тексту, морфемний рівень тексту, лексичний рівень тексту, синтаксичний рівень тексту.

Всі етапи розвитку української мови, які характеризуються динамічними процесами кількісних і якісних змін мовних явищ, відображені в текстах різних періодів й різних функціональних стилів. Особливої інтенсивності ці зміни набули сьогодні, у період інформаційного суспільства, який позначений посиленням диференціації сфер функціонування мови, розвитком електронних технологій комунікації і великим накопиченням різноманітних текстів української мови в паперових варіантах та в комп'ютерних версіях. Вивчення українських текстів цікавить сьогодні на тільки філологів, а й істориків, соціологів, психологів, фахівців інших галузей сучасної науки, а також спеціалістів інформаційних технологій. Тому уже на перших етапах вивчення текстів постають питання: де взяти необхідну кількість текстів, дібрану за потрібними хронологічними чи функціональними ознаками? Як швидко та ефективно вилучити з текстів необхідну для дослідження інформацію? Адже традиційна форма збирання та збереження мовних фактів (так звана, ручна паперова чи ручна електронна картотека) є надзвичайно складним завданням і не задовольняє потреб сучасних дослідників.

У сучасній комп'ютерній лінгвістиці відповідь на ці питання дає нова галузь – корпусна лінгвістика, яка займається створенням інформаційних ресурсів – корпусів текстів, що забезпечують доступ до репрезентативного текстового матеріалу та ефективний автоматичний аналіз текстів.

Відомі мовознавці, аналізуючи тенденції розвитку сучасної лінгвістики, визначають особливий статус корпусної лінгвістики в лінгвістичній парадигмі XXI ст. Зокрема В. Плунгян, зазначає, що «сучасна лінгвістика – лінгвістика корпусів», і перераховує такі пріоритетні завдання:

«...1) увага не до слова чи речення, а до тексту (дискурсу), тобто до реального інструмента комунікації в цілому, а не до його окремих фрагментів;

2) увага до квантитативного компонента мови, врахування насамперед найчастотніших елементів, порівняно з менш частотними, визнання квантитативних відношень суттєвим фактором у мовній еволюції і структурі мовних правил;

3) увага до синхронічної варіативності мови, тобто визнання того, що не існує єдиної системи засобів вираження змісту, а існують її різні реалізації, зокрема й залежні від психологічних, біологічних і соціальних факторів;

4) увага до діахронічної варіативності мови, тобто визнання того, що мова постійно змінюється в часі, і повністю відсторонитися від цієї нестабільності не можливо, у кожен момент часу в мові співіснують «прогресивні» і «консервативні» ділянки;

5) зміна відношення до поняття мовної норми і мовної правильності, тобто межа між «помилкою» та «маргінальним варіантом» визнається більш рухомою і хиткою» [Плунгян 2008: 7].

Також дослідники визначають зміну об'єкта лінгвістичного дослідження в корпусній лінгвістиці, тому що пріоритетним у цій галузі стає дослідження не тільки системи мови, а й системи мовлення. «У деякому розумінні корпусна лінгвістика змінює пріоритети дослідження: об'єктом дослідження стає мовлення, яке не зводиться до мовної абстракції, норм літературної мови, висновків про правильність / неправильність у мові, які ґрунтуються винятково на інтуїції освіченого дослідника. Другим важливим теоретичним наслідком корпусних досліджень можна вважати те, що соссюрівська дихотомія langue-parole замінюється уявленням про первинність мовленнєвої діяльності з плавною шкалою генералізації від мовленнєвого штампу до граматичного правила» [Копотев 2008: 12].

Українська лінгвістка В. Жуковська, докладно аналізуючи теорію та практику сучасної корпусної лінгвістики, наголошує на ефективності використання корпусу в лінгвістичному дослідженні: «Все вище сказане чітко окреслює дослідницьку програму корпусної лінгвістики, яка, будучи суто емпіричною дисципліною, при аналізі лінгвального матеріалу покладається на реальне функціонування мови з метою встановлення правил та вивчення особливостей продукування мови людиною, на відміну від тих досліджень, які опираються на вигадані приклади чи інтроспекцію. Застосування комп'ютерів дозволяє миттєво обробити величезний обсяг мовного матеріалу й відібрати всі можливі в конкретному корпусі приклади вживання необхідних для аналізу одиниць. У розпорядження лінгвіста надаються об'єктивні кількісні дані, забезпечуючи досягнення більш ґрунтовних та переконливих висновків. Корпусна лінгвістика дозволяє вченим підтвердити або спростувати гіпотези про функціонування мови, а також окреслити нові напрями дослідження, які до застосування корпусних методів не попадали до фокусу уваги дослідників» [Жуковська 2013: 13].

Наголошуючи на перспективності та важливості корпусних досліджень, статус корпусної лінгвістики в системі прикладних

лінгвістичних наук мовознавці визначають по-різному: одні вчені (Є. Карпіловська [Карпіловська 2006], Н. Дарчук [Дарчук 2010а], О. Зубань [Зубань 2018] та ін.) вважають корпусну лінгвістику частиною комп'ютерної лінгвістики, інші окреслюють її в окрему прикладну галузь, що визначається власним предметом, об'єктом, завданнями та методами дослідження (А. Баранов [Баранов 2001], О. Демська-Кульчицька [Демська 2005] та ін.). «Процедура корпусного аналізу включає три кроки: 1) ідентифікація мовних даних за допомогою категорійного аналізу; 2) співвідношення мовних даних за допомогою статистичних методів; 3) інтелектуальна інтерпретація результатів. Якщо перші два кроки повинні бути найбільшою мірою автоматизованими, то останній вимагає людської інтелектуальної діяльності, адже будь-яка інтерпретація є актом застосування розумових здібностей, а тому не може бути переведена в алгоритмічну процедуру. Саме у цьому проявляється головна відмінність між корпусною і комп'ютерною лінгвістикою, що зводить мову до набору процедур» [Teubert 2007: 113].

Розмежування комп'ютерної та корпусної лінгвістики, зроблене О. Демською-Кульчицькою, на наше переконання, не достатньо обґрунтоване: «Для корпусної лінгвістики застосування комп'ютерних інструментів не є визначальним критерієм, але, на відміну від комп'ютерної лінгвістики, корпусна лінгвістика покликана не моделювати функціонування мови в різних умовах, ситуаціях, проблемних галузях та послуговуватися цими моделями, а лише фіксувати всі аспекти функціонування мови[22], зберігаючи як інтра-, так і екстралінгвістичну специфіку, забезпечуючи оптимальну адекватність лінгвальних даних» [Демська 2005: 14].

Розуміння статусу корпусної лінгвістики залежить від того, яка параметризація закладена при організації корпусу мови. Справа в тому, що багато корпусів мови є *ілюстративними*, вони ставлять завдання: зібрати тексти, укласти словник-конкорданс за цими текстами й параметризувати, у кращому випадку, морфологічну інформацію (створити лематизатор) і /або лише метатекстову інформацію. У такому розумінні корпус мови виконує функцію фіксації текстів і пошуку текстових прикладів (як правило речень) за словоформою або лемою. *Дослідницькі* корпуси текстів покликані забезпечити лінгвістичний автоматичний аналіз зібраних текстів, тобто мати морфологічну, синтаксичну, семантичну, просодичну та анафоричну розмітки. Без комп'ютерного моделювання об'єктів і явищ тих мовних рівнів організації тексту, на яких проводиться лінгвістична анотація корпусу, не можливо здійснити лінгвістичну розмітку, а також автоматичний пошук та класифікацію текстового матеріалу. Тому особливої уваги сьогодні заслуговують ті корпусні лінгвістичні продукти, які спрямовані на аналіз тексту й мають статус динамічних пошукових систем, які здатні в автоматичному або автоматизованому режимі вилучати інформацію про мовні одиниці з будь-якого параметризованого тексту.

[22] Підкреслено автором монографії.

Слушною, на наше переконання, є думка, обґрунтована М. Копотєвим, та А. Мустайокі: «Зазначимо, що вживання цього терміна [корпусна лінгвістика][23] вимагає окремого обговорення. Справа в тому, що сам по собі він має два значення. Це, по-перше, теорія і методика створення корпусів і, по-друге, корпусні дослідження, тобто дослідження мови за допомогою корпусних методів. Проте, чіткої межі між ними не існує, і практично всі, хто займається створенням корпусів проводять водночас і лінгвістичні дослідження. У цілому, корпусна лінгвістика в першому значенні більш технологічна й передбачає спільну роботу лінгвістів і спеціалістів із комп'ютерних технологій, тоді як друге завдання – справа лінгвістів, зокрема й фахівців зі статистичного оброблення мови. Якщо говорити про російську корпусну лінгвістику, частіше розуміється друге значення, але необхідно пам'ятати, що використання терміна в першому значенні широко розповсюджене у світі й інституалізовано у формі великої кількості дослідницьких центрів і спеціалізованих журналів (див., наприклад, журнал Corpus Linguistics), і без першого, строго кажучи, не існувало б і другого» [Копотев 2008:11 – 12] .

Вихідним положенням у нашому дослідженні є диференційований підхід до розуміння терміна "корпусна лінгвістика": 1) завдання і дослідження, які спрямовані на створення такого інформаційного продукту, як корпус мови, безумовно належать до комп'ютерної лінгвістики, і в такому трактуванні корпусна лінгвістика є одним із напрямів комп'ютерної лінгвістики; 2) завдання, які виконує лінгвіст у дослідженні мови за допомогою інформаційного продукту – корпусу мови, належать до корпусноорієнтованих лінгвістичних розвідок різних галузей мовознавства.

У корпусноорієнтованих лінгвістичних дослідженнях значущість, вагомість результатів, безумовно, визначається насамперед репрезентабельністю матеріалу дослідження: чим більше мовних фактів, тим достовірніші спостережувані закономірності, але без об'єктивного точного автоматичного / автоматизованого визначення у великому текстовому масиві мовних одиниць і явищ ефективність проведення такого дослідження й встановлення достовірних висновків не є можливим. Тому в сучасній українській корпусній лінгвістиці особливої актуальності набувають глибоко параметризовані корпуси текстів, оснащені пошуково-класифікаційними програмними аналізаторами, що забезпечують:

- ефективне та оперативне проведення автоматичного лінгвістичного аналізу;
- можливість автоматичного оброблення великих текстових масивів;
- отримання точних формальних та реляційно-функціональних характеристик мовних одиниць і явищ різних рівнів текстової організації.

Українська корпусна лінгвістика, як галузь комп'ютерної лінгвістики, розпочинає свій розвиток на початку XXI ст. [Бобкова 2014], тоді ж

[23] Вставка автора монографії.

утверджуються терміни "корпус текстів" та "корпусна лінгвістика". У вільному доступі в мережі Інтернет сьогодні представлені такі корпуси текстів української мови: Корпус української мови (КУМ) [КУМ 2019], Генеральний регіонально анотований корпус української мови (ГРАК) [ГРАК 2018], Корпуси текстів української мови [КТУМ 2018], Браунський корпус української мови (БрУК) [БрУК 2018]. Закритими для доступу широкого користувача є Український національний лінгвістичний корпус [Широков 2011], який використовується Українським мовно-інформаційним фондом НАН України для текстової ілюстрації при укладанні різноманітних словників; Комп'ютерний фонд інновацій (КФІ) [АРСУН 2013], який активно використовується у дослідженнях колективу відділу лексикології, лексикографії та структурно-математичної лінгвістики Інституту української мови НАН України [Карпіловська 2019].

Серед перерахованих корпусів найглибшу параметризацію має Корпус української мови [КУМ 2019], який створюється колективом комп'ютерних лінгвістів Інституту філології Київського національного університету імені Тараса Шевченка під керівництвом доктора філологічних наук Н. Дарчук.

У Корпусі української мови можна визначити три взаємопов'язані структурно-функціональні зони:

1) модуль-текст, у якому в електронній формі представлені українські тексти;

2) модуль-аналізатор, який має програмне забезпечення автоматичного пошуку мовних явищ;

3) модуль-словник, у якому результати автоматичного аналізу тексту систематизуються в електронних словниках, представлених в Інтернеті для користувача. Тобто, тільки 3-ій модуль, як результат роботи всіх систем бачить користувач.

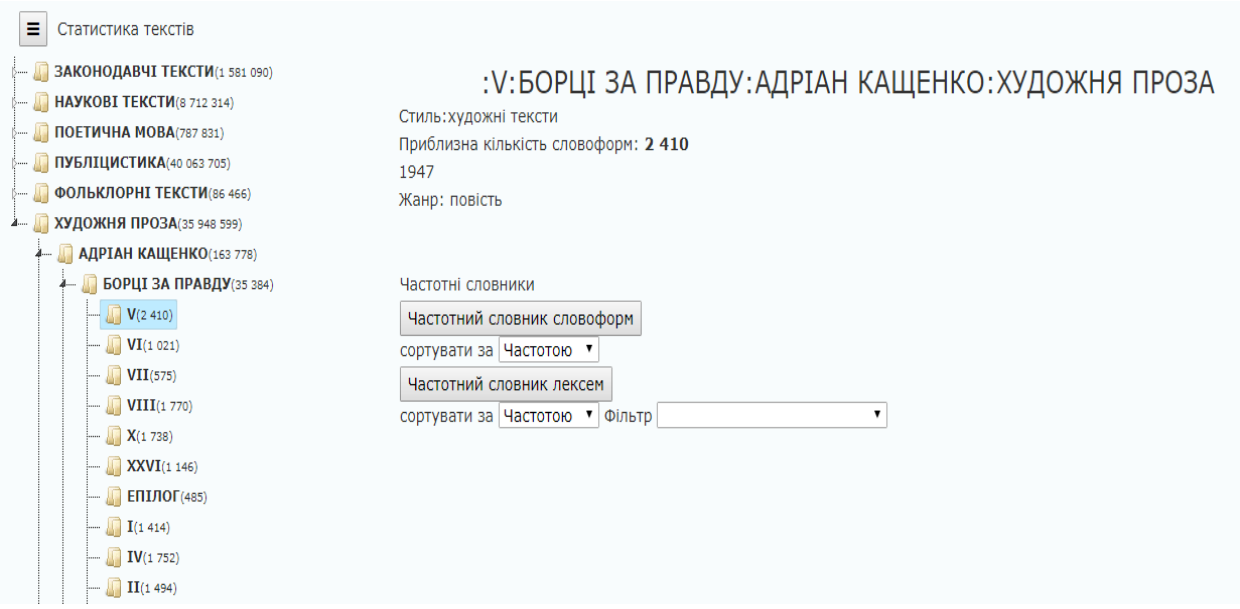


Рис. 3.1. Стильові підкорпуси Корпусу української мови (за даними 2018 р.)

М о д у л ь - т е к с т . На сьогодні Корпус української мови представляє зібрання текстів обсягом ≈ 100 млн. слововживань. Маркування метаінформації текстів корпусу здійснюється, насамперед за стилем.

За стильовими ознаками формується шість підкорпусів (рис.3.1): законодавчі тексти – 1 581 090 слововживань; наукові тексти – 8 712 314 слововживань; поетична мова – 784 831 слововживання; публіцистика – 40 063 705 слововживань; фольклорні тексти – 86 466 слововживань; художня проза - 35 948 599 слововживань. У межах кожного стильового підкорпусу формуються за ієрархічним принципом підкорпуси за різноманітними ознаками (галузь, тема, періодичне видання, автор та ін.). Кінцевою ланкою ієрархії є заголовок конкретного тексту (наприклад, художня проза: Андріан Кащенко: Борці за правду: V частина). За умови активації конкретного твору чи частини твору, до нього додається інформація про видавництво, місце видання, жанр тексту та деяка інша інформація. Як показує статистика, корпус текстів вимагає стилістичного збалансування, над чим сьогодні працює колектив викладачів та студентів.

М о д у л ь - а н а л і з а т о р . Модуль-аналізатор – інструмент лінгвістичних досліджень великих текстових масивів. Програмно-операційне середовище модуля-аналізатора може виконувати такі функції:

1) забезпечує зв’язок модуля-тексту із лінгвістичними базами даних: морфологічною, морфемною, синтаксичною та семантичною;

2) здійснює автоматичному режимі пошук і класифікацію лексики за різними параметрами, формує реєстри лексем (початкових форм) та словоформ за текстовими вибірками;

3) проводить автоматичний морфологічний, морфемний, синтаксичний, семантичний і статистичний аналізи;

4) формує словник-конкорданс до одиниць усіх рівнів лінгвістичного аналізу, що здійснюється у КУМ.

Ефективність пошуку в корпусі залежить від глибини параметризації текстів. Залежно від того, на яких рівнях тексту (морфологічному, морфемному, синтаксичному, семантичному) проведено анотацію, такі мовні одиниці в текстах корпусу й можна шукати: морфеми, граматичні форми слів, лексеми, словосполучення, моделі структури речень, семантичні класи слів.

Тексти Корпусу української мови параметризуються у модулі-аналізаторі за чотирма рівнями анотації [Дарчук 2016]:

1) морфологічна анотація – базовий етап для всіх наступних рівнів – визначає морфологічні параметри слова: частиномовну належність і граматичні ознаки кожного слововживання тексту (працює автоматично), а також здійснює лематизацію слововживань в одну лексему;

2) морфемна анотація сегментує леми тексту на морфеми й визначає функціональний тип морфеми (працює автоматично);

3) синтаксична анотація визначає словосполучення й приписує кожному з них інформацію про тип і вид синтаксичного зв’язку (працює

автоматично); а також будує дерева структур речень (працює автоматично / автоматизовано);

4) семантична анотація приписує кожному слову код семантичного поля таксономічної класифікації (працює автоматично / автоматизовано).

Анотування текстів відбувається через зв'язок із великими лінгвістичними базами даних (наприклад, морфологічна БД – 3,5 мільйони словоформ, морфемна БД – 200 тисяч слів), які укладалися за розробленою методикою комп'ютерного моделювання одиниць різних мовних рівнів – комп'ютерною граматикою української мови: «Для автоматичного аналізу українського тексту нами створено комп'ютерну граматику, яка є ієрархічним комплексом комп'ютерних моделей: морфемно-словотвірної, морфологічної, синтаксичної моделі, побудованих на основі формальних, точних й однозначних правил. Ці моделі можна вважати дослідницькими, тому що закладені в граматики алгоритмічні правила призводять до виявлення того чи іншого мовного явища (морфів, словоформ з їх частиномовними й категорійними характеристиками, словосполучень, дерев залежностей речень тощо). Алгоритмічно зімітовано діяльність лінгвіста – а саме забезпечено перехід від сукупності текстів до системи, яка лежить в їх основі, встановлено елементарні одиниці і класи елементарних одиниць. Розроблені моделі є моделями аналізу, індуктивними, несемантичними і детерміністськими (структурними)» [Дарчук 2013: 28]. Дослідницький Корпус текстів – це лише один спосіб застосування комп'ютерної граматики. Вона може бути використана в різних автоматичних системах оброблення тексту ненаукового спрямування: чатових діалогових системах, системах реферування текстів, системах визначення тематики текстів, пошукових онтологіях, системах перевірки авторства текстів та в інших завданнях, які потребують роботи з текстовими масивами. У такому використанні комп'ютерна граматика виступає складовою NLP-систем.

У корпусній лінгвістиці поняття анотації визначається за Дж. Лічем: «процес анотування корпусних даних – це додавання інтерпретованої, лінгвістичної інформації до електронного корпусу усного чи/або писемного мовлення» [Leech 1997: 2]. Лінгвістична розмітка в Корпусі української мови проводиться двома способами:

1) анотація всіх слововживань за введеними текстами: суцільна анотація тексту на всіх рівнях розмітки й формування великої анотованої текстової бази даних;

2) анотація словника початкових форм (лем) та текстових слововживань, укладеного за обмеженими текстовими вибірками: вибіркова анотація і формування автономних баз даних.

Укладачі корпусу свідомо відмовились від першого способу анотації, тому що розмітка мільйонного масиву тексту на всіх рівнях анотації вимагає дуже потужного технічного забезпечення, інакше робота з корпусом стає надзвичайно повільною. О. Ляшевська [Ляшевская 2016: 15], аналізуючи анотацію Національного корпусу російської мови, наводить фрагмент XML-

представлення розмітки фрагмента тексту, у якому три слововживання (*Цены в них*) анотовані 79 рядками розмітки: лексико-граматичні теги (lex и gramm) і лексико-семантичні теги (sem), не враховуючи метарозмітки (додаток 9). Цей приклад наглядно демонструє, який обсяг інформації систематизує сформована у такий спосіб база даних анотованих текстів.

Анотація текстових слововживань у КУМ здійснюється на двох рівнях текстової розмітки: морфологічному та синтаксичному. Автоматичне приписування кожному слововживанню тексту граматичного коду та автоматична лематизація відбувається при введенні тексту в корпус. Морфологічно анотовані тексти утворюють БД модуля-тексту, на якому відбувається автоматичне формування конкорданса та укладаються резидентні словники словоформ і лем. На всіх інших рівнях анотації параметризація відбувається автоматично / автоматизовано за обмеженими текстовими вибірками: на морфемному [24] та семантичному рівні параметризується словник початкових форм (лем); на синтаксичному рівні – речення, із приписаними граматичними кодами кожному слововживанню, наприклад:

Трагедія(КИ) середини(КР) XIX(U) ст.(ББ) стала(ГЙ) для(ПР) ірландського(АР) народу(ЙР) історичним(АТ) рубежем:(ЙТ) після(ПР) «великого(АР) голоду»(ЙР) незалежність(КИ) острівної(АЗ) країни(КР) стала(ГЙ) лише(Ь0) питанням(ЛТ) часу.(ЙР).

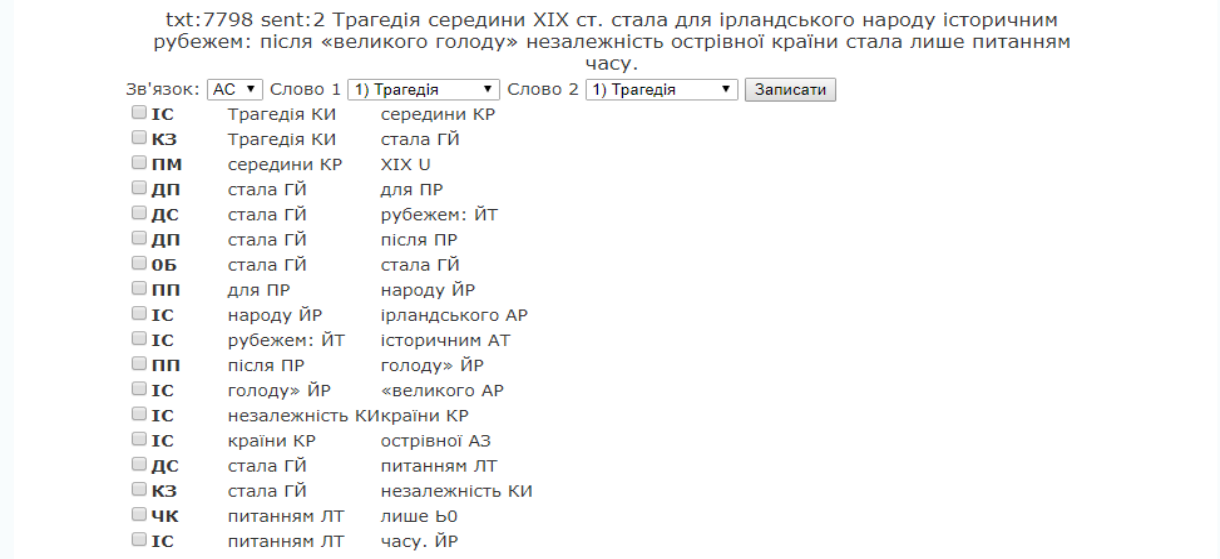


Рис. 3.2. Робоча картка перевірки анотації синтаксичних зв’язків

Синтаксичні зв’язки в реченнях визначаються автоматично[25], а потім в автоматизованому режимі лінгвіст релагує, визначені машиною, синтаксичні зв’язки (рис. 3.2).

[24] Процедура автоматичного морфемного аналізу описується у наступних параграфах.
[25] Глибокий аналіз синтаксичної анотації зроблено Н. Дарчук у монографії «Комп’ютерне анотування українського тексту: результати і перспективи» [Дарчук 2013].

За базою даних синтаксичних зв’язків машина будує дерево залежностей, яке також перевіряється й редагується лінгвістом автоматизовано (рис. 3.3).

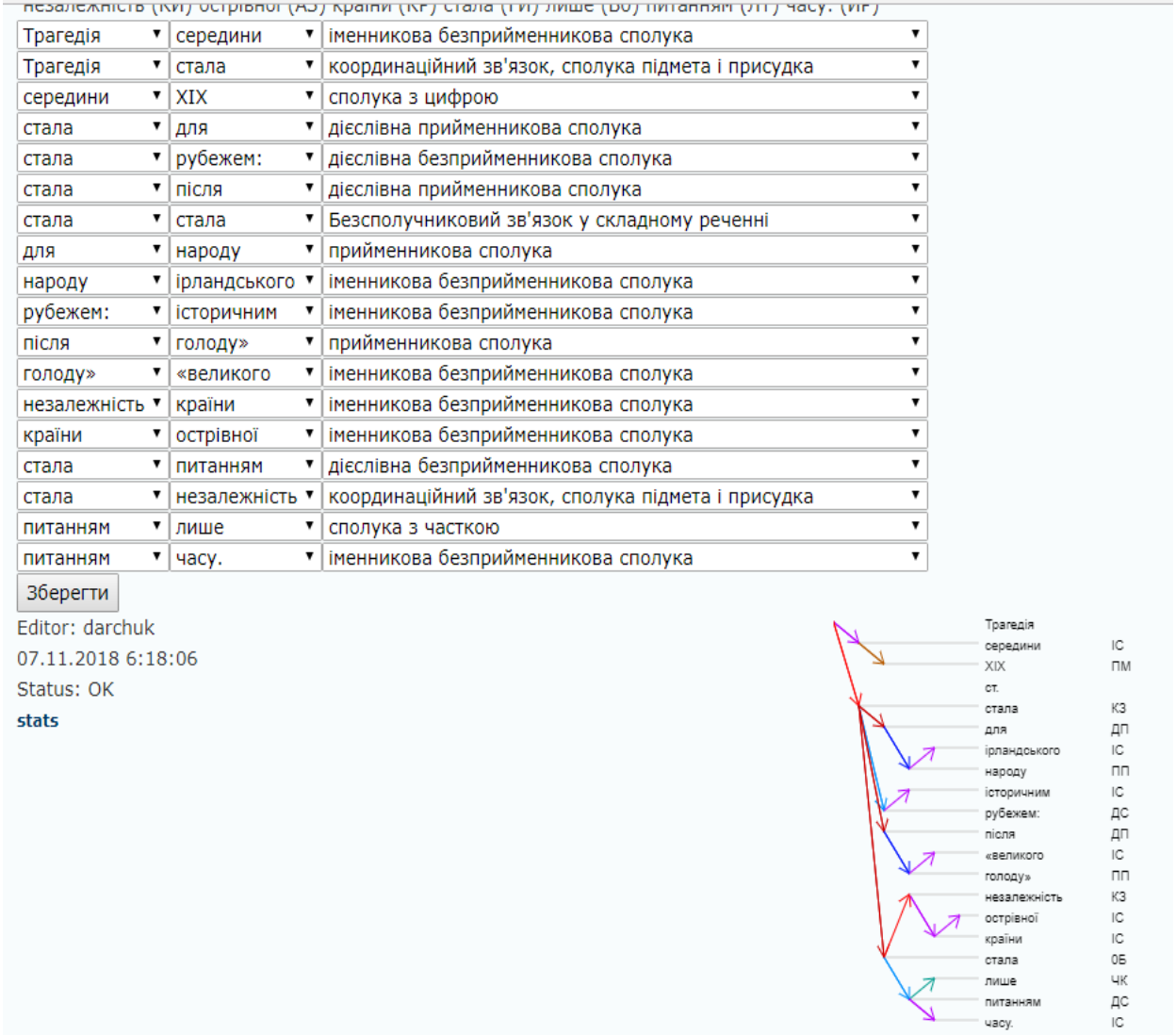


Рис. 3.3. Робоча картка автоматизованого редагування синтаксичного дерева

На семантичному рівні анотація відбувається на реєстрі словника початкових форм, укладеного за обмеженою текстовою вибіркою, автоматично: кожній лемі тексту приписується код семантичного класу за БД семантичних таксонів [Дарчук 2016].

Базовою одиницею кожного рівня анотації є слово. Такий підхід ґрунтовано на лінгвістичному постулаті: слово – центральна одиниця мови, яка на різних рівнях мовної системи характеризується різними типами структур, або ж як структурна одиниця входить до одиниць вищих мовних рівнів – словосполучення та речення. Кожне слововживання, яке в технічному середовищі визначається як послідовність графем між пробілами тексту, через посередництво морфологічного модуля має постійний зв’язок із лемою – початковою формою слова в морфологічній БД та номером речення

у текстовій БД. У базах даних рівневого аналізу робоча одиниця (лема або слововживання) анотується: приписується код комп'ютерної розмітки, що моделює лінгвістичну інформацію (морфемну, морфологічну, семантичну та синтаксичну).

Модуль-аналізатор – це складна система автоматичного аналізу тексту, яка складається із баз даних та СКБД, що забезпечують:

1) проведення автоматичного / автоматизованого лінгвістичного аналізу;

2) автоматичне укладання за результатами цього аналізу автономних баз даних, за якими укладаються різноманітні електронні словники, що представлені у Корпусі української мови у рубриках "Пошук у корпусі", "Частотні словники", "Статистика".

М о д у л ь - с л о в н и к . Модуль-словник систематизує лінгвістичну інформацію, отриману в результаті автоматичної лінгвістичної анотації текстів. У поєднанні роботи двох модулів – модуля-текстів та модуля-аналізатора – за запитом користувача автоматично укладаються різні типи словників: словники-конкорданси; алфавітно-частотні словники слів (лексем), словоформ, морфем, морфемних структур слів, словосполучень, семантичних таксонів. Це інтерактивні словники, які автоматично класифікують, групують досліджувані одиниці за різними ознаками та подають усі можливі контексти вживання цих одиниць за вибірками КУМ.

Інфологічна модель кожного словника та його структура визначалися специфікою електронного характеру та лінгвістичними особливостями одиниць реєстру [Зубань 2016а].

Словники-конкорданси автоматично укладаються за опцією "Пошук у корпусі" (рис.3.4) у межах підкорпусу текстів обраного стилю (перше зверху діалогове вікно, опція "Виберіть зону пошуку"). Укладання може здійснюватися за такими пошуковими параметрами:

1) пошук контекстів до одного слова за заданою конкретною лексемою (всі словоформи лексеми) або словоформою – перше в другому рядку діалогове вікно;

2) пошук контекстів до всіх словоформ стилю за обраною морфологічною характеристикою (друге діалогове вікно в другому рядку, опція "Морфологічні ознаки").

Словник-конкорданс лексеми / словоформи можна автоматично укласти, записавши обране слово в перше знизу діалогове вікно: на рис. 3.4 задано пошук контекстів до словоформи *Україною* (орудний відмінок однини) в межах підкорпусу художньої прози);

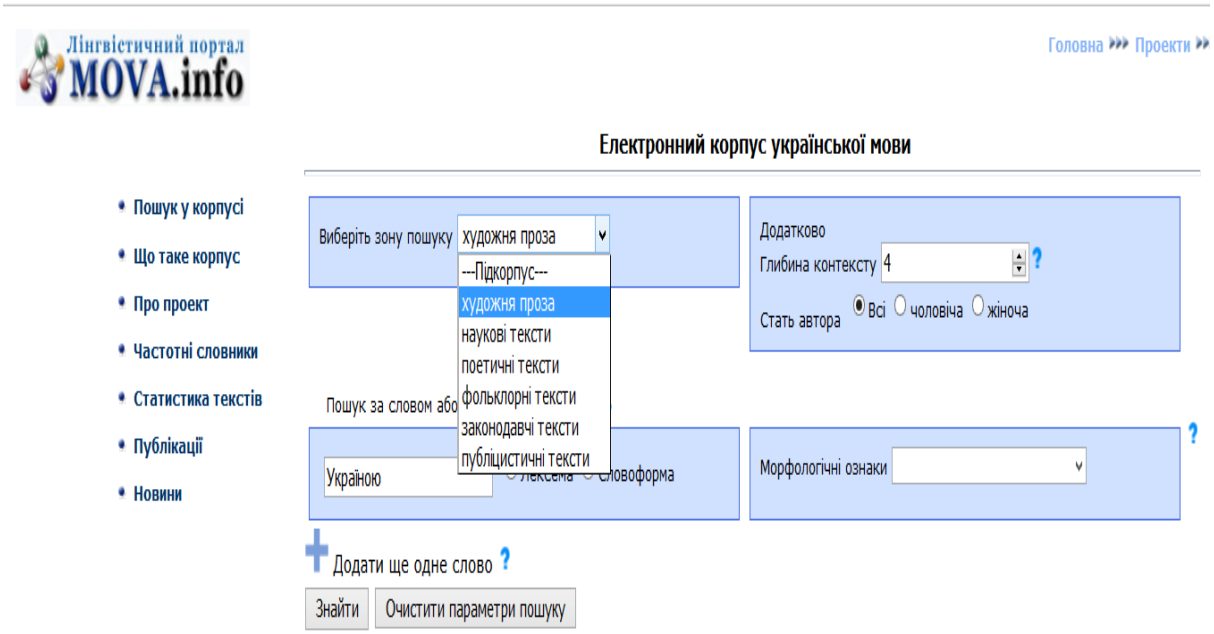


Рис.3.4. Параметри автоматичного укладання словника-конкорданса словоформи *Україною*

За умови активації кнопки "Знайти", автоматично будується конкорданс до заданої словоформи (рис. 3.4) з урахуванням двох додаткових параметрів: 1) глибини контексту (кількості слововживань правобічного та лівобічного оточення словоформи у реченні); 2) статі автора.

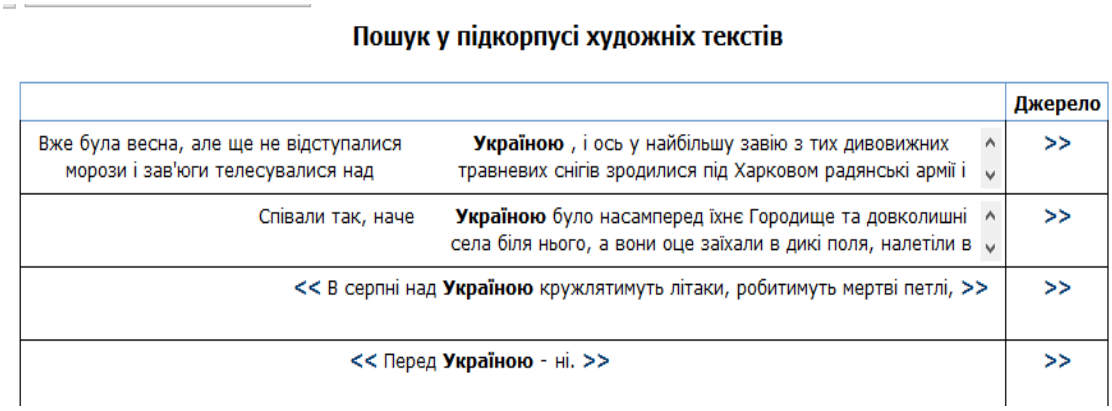


Рис.3.5 Фрагмент словника-конкорданса словоформи *Україною*

Як показує приклад словника-конкорданса словоформи *Україною* (рис. 3.5), контекст до аналізованої словоформи може бути розширений у двох напрямках дистрибуції слова до межі речення за допомогою активації позначок – «; – ». До кожного текстового слововживання подається індекс джерела (рис. 3.6), з якого взято текстовий фрагмент, наприклад, перше речення – *Вже була весна,..* – взято із роману «Диво» (глава: «1966 рік Весна. Київ») П. Загребельного. Навігація до джерела здійснюється автоматично за допомогою позначки - » - в колонці "Джерело".

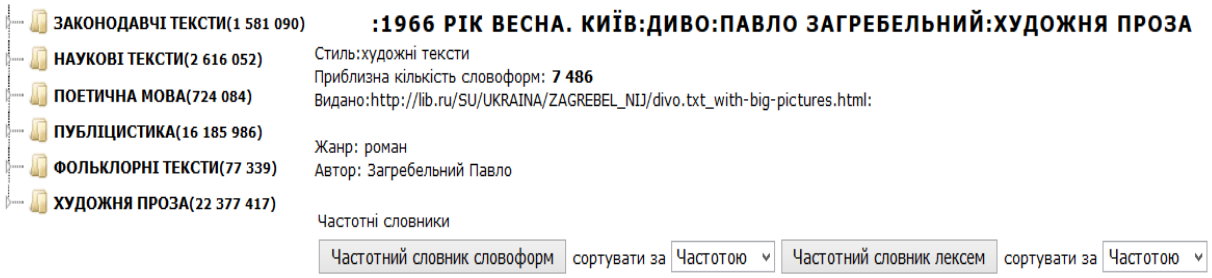


Рис. 3.6. Визначення джерела текстового фрагмента

Виведення результатів пошуку в побудові конкорданса можливе й у режимі цитування (рис. 3.7).

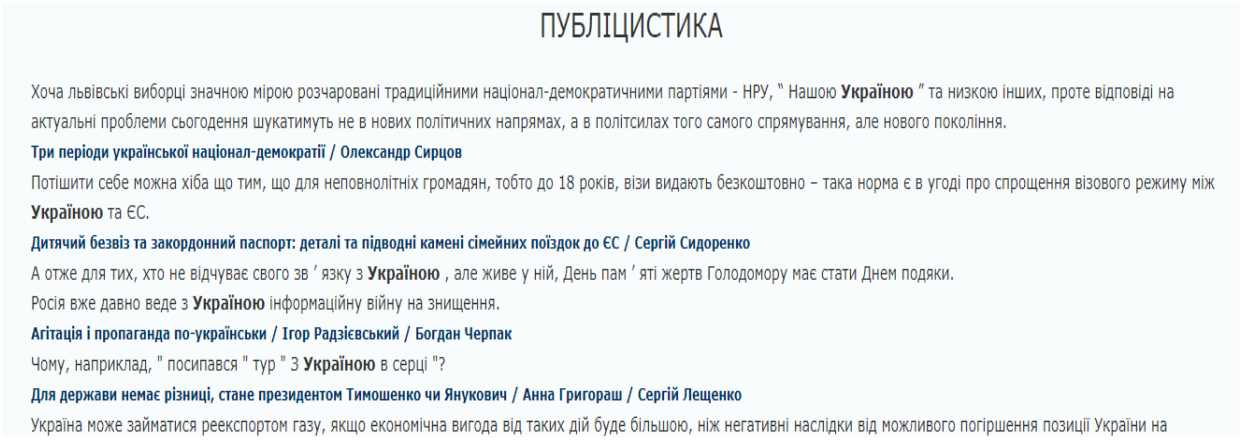


Рис. 3.7. Фрагмент словника-конкорданса у режимі цитування

Словник-конкорданс за параметром пошуку контекстів до всіх словоформ стилю за обраною морфологічною характеристикою (друге діалогове вікно в другому рядку – рис. 3.4) будується за вибором морфологічних ознак кожної частини мови у випадному списку опції "Морфологічні ознаки".

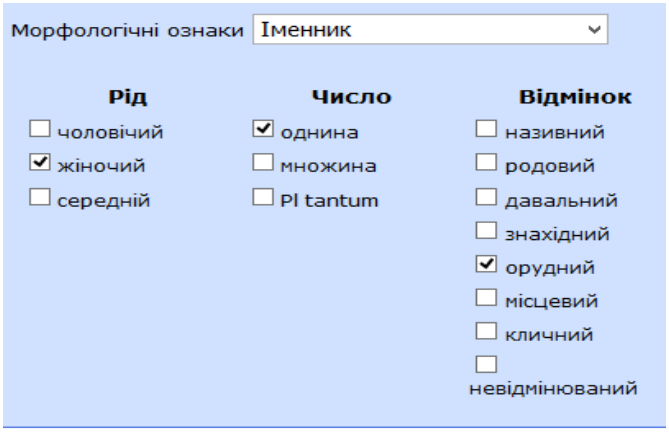


Рис.3.8. Параметри автоматичного укладання конкорданса за обраною морфологічною характеристикою

На рис. 3.8 задано пошукові параметри: частина мови – іменник, рід – жіночий, число – однина, відмінок – орудний**.** За цими параметрами пошуку автоматично укладається конкорданс до всіх іменників жіночого роду орудного відмінка однини, які вживаються в текстах художньої прози (рис.3.9). Навігація до текстового джерела здійснюється автоматично через позначку - » -.

Пошук у підкорпусі художніх текстів

	Джерело
Дівчата хоча притомлені, та водночас і вдоволені щойно пережитою **напругою** , ніби й справді їм вдалося когось порятувати.	>>
<< доводиться в духотняві, яма налита **спекою** . >>	>>
<< З **місією** Червоного Хреста в далекій південній країні була >>	>>
<< , головою поводить, стежить за **танцівницею** , що зовсім близько перед нею теж >>	>>
<< Липка, задушна ніч, повалені **холерою** люди стогнуть за брезентом твого намету, >>	>>
<< Костянтинівна, що, посріблена тепер **сивиною** , з поглядом пригаслим, сидить серед >>	>>

Рис. 3.9. Фрагмент словника-конкорданса іменникових словоформ жіночого роду орудного відмінка однини на базі текстів художньої прози

Рубрика "Статистика текстів" відкриває діалогове вікно зі стилістичною параметризацією корпусу за підкорпусами стилів (рис.3.1). Розгортаючи дерево кожного підкорпусу до кінцевої ланки – конкретного тексту, користувач на базі цього тексту може автоматично укласти алфавітно-частотні словники лексем та словоформ кожного тексту корпусу із визначенням абсолютної частоти вживання.

:1966 РІК ВЕСНА. КИЇВ:ДИВО:ПАВЛО ЗАГРЕБЕЛЬНИЙ:ХУДОЖНЯ ПРОЗА

Стиль:художні тексти
Приблизна кількість словоформ: **7 486**
Видано:http://lib.ru/SU/UKRAINA/ZAGREBEL_NIJ/divo.txt_with-big-pictures.html:

Жанр: роман
Автор: Загребельний Павло

Частотні словники

Частотний словник словоформ сортувати за Алфавітом Всього словоформ:3234

Словоформа	Абс.частота
Але	88
альпіністами	1
алюмінієвий	1
аніж	1
анатомія	1
Андре	1
ансамбль	1

Частотний словник лексем сортувати за Частотою Всього лексем:2410

Лексема	Абс.частота
навіть	23
та	23
б	23
час	23
ми	23
колоти	22
бузина	22

Рис. 3.10. Фрагмент частотного словника лексем та словоформ глави "1966 рік Весна. Київ" роману «Диво» П. Загребельного

На сьогодні у Корпусі української мови в режимі on-line автоматично укладаються ≈ 40 тис частотних словників лексем (лем) / словоформ за окремими текстовими вибірками. На рис.3.10. показано фрагмент двох частотних словників: частотного словника лексем та частотного словника словоформ. Словники укладаються за вибором параметра формування реємтру одиниць: алфавітом або частотою (абсолютна частота вживання у вибірці тексту). За параметром частоти вживання інвентар одиниць може формуватися за спадом частот або за ростом частот при активації опції "Абс. частота".

За цією самою текстовою вибіркою також можливе автоматичне укладання семантичного словника, у якому до кожної реєстрової одиниці подано категорійну сему таксона, до якого належать ЛСВ лексеми реєстру (рис. 3.11)

Лексема	Клас	Абс.частота	Сема
віно	Л	46	Іменник НЕПРЕДМЕТНІ ІМЕНА фінанси, гроші
до	Л	43	Іменник НЕПРЕДМЕТНІ ІМЕНА звук
отава	К	43	Іменник ПРЕДМЕТНІ ІМЕНА рослини, сорти рослин
йога	К	16	Іменник НЕПРЕДМЕТНІ ІМЕНА метод (спосіб, напрям, прийом)//Іменник НЕПРЕДМЕТНІ ІМЕНА погляд
жінка	К	15	Іменник ПРЕДМЕТНІ ІМЕНА особа//Іменник ПРЕДМЕТНІ ІМЕНА СЛОВОТВІРНІ МІТКИ nomina feminina//Іменник ПРЕДМЕТНІ ІМЕНА особа імена родинності
тога	К	15	Іменник ПРЕДМЕТНІ ІМЕНА інструменти і пристрої одяг, взуття, прикраси
етюд	Й	14	Іменник ПРЕДМЕТНІ ІМЕНА тексти (книги, документи) зображення//Іменник ПРЕДМЕТНІ ІМЕНА тексти (книги, документи)//Іменник НЕПРЕДМЕТНІ ІМЕНА музика
виставка	К	13	Іменник НЕПРЕДМЕТНІ ІМЕНА захід
люди	И	13	Іменник ПРЕДМЕТНІ ІМЕНА МЕРЕОЛОГІЯ множини і сукупності об'єктів
Софія	к	13	Іменник ВЛАСНІ ІМЕНА топоніми//Іменник ВЛАСНІ ІМЕНА імена
художник	Й	12	Іменник ПРЕДМЕТНІ ІМЕНА особа
рука	К	12	Іменник ПРЕДМЕТНІ ІМЕНА тексти (книги, документи)//Іменник ПРЕДМЕТНІ ІМЕНА МЕРЕОЛОГІЯ частини частини тіла і органи людини//Іменник ПРЕДМЕТНІ ІМЕНА МЕРЕОЛОГІЯ частини частини тіла й органи тварин
слово	Л	12	Іменник НЕПРЕДМЕТНІ ІМЕНА мовлення//Іменник ПРЕДМЕТНІ ІМЕНА тексти (книги, документи)

Рис. 3.11. Фрагмент семантичного словника лексем за текстовою вибіркою глави «1966 рік Весна. Київ» роману «Диво» П. Загребельного

Частотні словники за стилями, розділами, авторами, збірками тощо з метою оптимізації пошуку на базі великого обсягу текстової інформації не укладаються в режимі on-lin. Такі словники укладені на замовлення користувача й викладені в рубриці "Частотні словники" (рис. 3.4) як автономні електронні лексикографічні системи, з якими також можна працювати в режимі on-lin. На сьогодні укладено 20 електронних частотних словників [ЧСКУМ 2017].

Наприклад, в електронному словнику мови Тараса Шевченка [ЕСМТШ 2017] користувач може автоматично укласти такі частотні словники:

1) алфавітно-частотний словник словоформ за заданою буквою або словом;

2) алфавітно-частотний словник усіх словоформ усіх частин мови або за вибраною морфологічною характеристикою;

3) алфавітно-частотний словник лексем усіх частин мови або за вибраною морфологічною характеристикою;

4) алфавітно-частотний словник словосполучень;

5) алфавітно-частотний словник морфем (префіксів, коренів, суфіксів, інтерфіксів) усіх слів або за вибраною морфологічною характеристикою;

6) алфавітно-частотний словник морфемних структур слів (початкових форм) усіх частин мови або за вибраними морфологічними ознаками.

Кожен електронний словник структурується на три зони: інвентар одиниць; статистичні дані; конкорданс.

Розглянемо алфавітно-частотний словник словосполучень, укладений автоматично за вибіркою поетичних текстів Т. Шевченка. Словник будується в інтерактивному режимі за 9-ма лексикографічними параметрами (рис. 3.12):

1) уведеним словосполученням – "Уведіть повністю словосполучення";

2) початковою формою головного слова (опція - "Головне слово"), залежного слова (опція - "Залежне слово") або обох слів словосполучення;

3) прийменником у прийменниковому словосполученні (опція – "Прийменник");

4) морфологічними характеристиками (випадні списки морфологічних ознак в опції "Частина мови") головного слова, залежного слова або обох слів словосполучення;

5) типом синтаксичного зв’язку (опція – "Тип зв’язку": усі, координація, підрядний ядровий, підрядний ад’юнктний, сурядний);

6) грамемою залежного слова (опція – "Граматична форма залежного слова");

7) морфологічним типом словосполучення (опція – "Тип словосполучення": усі, іменникове, дієслівне, прикметникове, займенникове, числівникове).

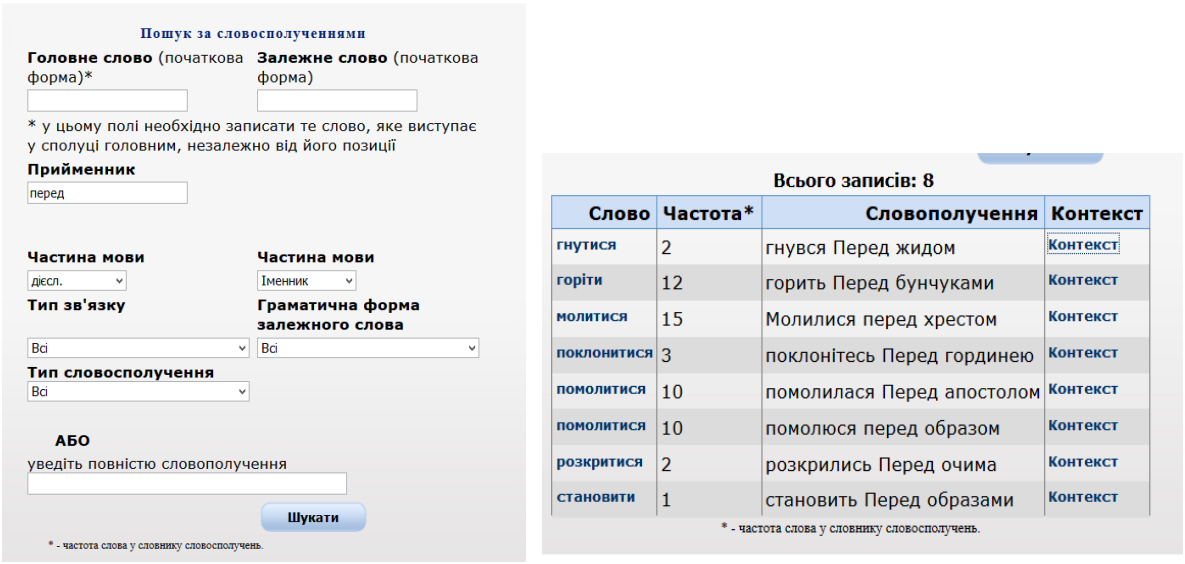


Рис. 3.12. Параметри укладання частотного словника словосполучень

На рис. 3.12 задано такі параметри пошуку: головне слово – дієслово; залежне слово – іменник, зв'язок прийменниковий за допомогою прийменника *перед*. За цими параметрами при активації опції "Шукати" автоматично укладається алфавітно-частотний словник восьми словосполучень, автоматично визначених у текстах Т. Шевченка. Як і в попередніх типах частотних словників, цей словник має зв'язок із контекстами, у яких реалізовується кожне словосполучення, а контексти мають зв'язок із джерелом (конкретним твором), із якого взято речення.

Два типи словників у Корпусі української мови – конкорданси та частотні – поєднані між собою взаємозворотньою й інформаційно доповнювальною навігацією:

1) конкорданс → частотний словник: словники-конкорданси (рис. 3.4, рис. 3.5.) через опцію "Джерело" поєднуються із алфавітно-частотними словниками тексту, з якого взято речення (рис. 3.6.);

2) частотний словник → конкорданс: якщо користувач працює із частотними словниками стилів, авторів, збірок, то через опцію "Контекст" або активацію конкретного слова (рис. 3.12) він може перейти до конкорданса обраного слова або словосполучення.

Електронні частотні словники та словники-конкорданси у дослідницькій лінгвістичній системі Корпус української мови – надзвичайно ефективні і раціональні інструменти лінгвістичних досліджень, тому що вони передбачають різноманітні автоматичні класифікації з одиницями українського тексту в інтерактивному режимі. Отримана статистична інформація про організацію українських текстів на різних рівнях структури текстів дає можливість вивчати закономірності функціонування мовних одиниць у різних стилях, комплексно досліджувати мовні особливості ідіостилів українських поетів та письменників.

Багатоаспектна систематизація лінгвістичної інформації в Корпусі української мови, встановлення статистичних закономірностей функціонування мовних одиниць у різних типах текстів формують лексикографічну інтегральну систему нового покоління, яка розглядається як універсальна довідкова система з української мови для учителя, журналіста або пересічного користувача, а для філолога-дослідника, викладача – як лінгвістична база знань.

3.2. Автоматичний морфемний аналіз у Корпусі української мови

3.2. 1.Чи потрібний у корпусі текстів морфемний аналіз?

Корпус української мови [КУМ 2019] має дослідницький характер і за своїм призначенням орієнтований на вирішення широкого кола лінгвістичних завдань, зокрема в галузі морфеміки і словотвору, хоча більшість корпусів слов'янських мов не мають параметризації текстів на морфемному рівні. За нашими даними, крім Корпусу української мови, пошук за морфемами мають Національний корпус російської мови [НКРЯ

2018], Комп'ютерний корпус текстів російських газет кінця XX-ого століття [ККТРГ 2018], який пізніше ввійшов до Полістильового корпусу текстів сучасної російської мови [Кукушкина 2005]. Виникає ряд питань: Чи потрібна в корпусі морфемна (словотвірна) анотація тексту? Які дослідницькі перспективи відкриває перед лінгвістом корпуснoорієнтірований автоматичний морфемний і словотвірний аналіз?

Відповіді на ці питання знаходимо у роботах відомих вітчизняних та зарубіжних лінгвістів: Є. Карпіловської [Карпіловська 2019], Н. Клименко [Клименко 2008], Н. Дарчук [Дарчук 2013], А. Полікарпова [Поликарпов 2013], О.В. Кукушкіної [Кукушкина 2006а], О. Ляшевської [Ляшевская 2016], А. Токтонова [Токтонов 2006] та ін..

Одним із векторів корпуснoорієнтованого морфемно-словотвірного аналізу є вивчення неологічних процесів у сучасному лексиконі, що формує нову галузь мовознавчих досліджень – неологію [Попко 2007] та нову галузь лексикографії – неографію [Дубичинский 2000].

В українському мовознавстві лексикографічне моделювання мовних інновацій не обмежується лише фіксацією нових слів та їх тлумаченням: неографічні та неологічні дослідження спрямовані на вивчення деривaційних, граматичних, стилістичних та інших процесів у проекції на динаміку сучасного українського лексикону. Л. Кислюк [Кислюк 2017] відзначає синтез кодифікаційних словотвірних моделей та мовної практики у вивченні українських новотворів: «Для визначення нормативності новотвору важливими є системна та комунікативна підтримка такої одиниці. Українські дослідники впровадили й успішно застосовують для визначення стабілізації інновації поняття функціонального потенціалу як сумарного показника парадигматичних (ієрархічних, гіперо-гіпонімічних), епідигматичних (деривaційних та асоціативних) і синтагматичних (позиційних і комбінаторних) відношень інновації в системі мови й у тексті, її номінаційної та комунікативної активності». Методологічне поняття "функціональний потенціал інновації" використовують у своїх дослідженнях Л. Кислюк [Кислюк 2017], Є. Карпіловська [Карпіловська 2007, 2008: 6], А. Таран [Таран 2011] та ін. Зокрема Є. Карпіловська зазначає, що «... через нове слово як стрижневу номінативну одиницю дослідник мовної динаміки дістає змогу виявити й інші типи мовних інновацій: правописні, так звані неографізми чи орфографічні новації, морфемні (неоморфеми), словотвірні (неоформанти й неооснови, нові моделі словотворення), граматичні (морфологічні й синтаксичні), стилістичні. Встановлення зовнішніх і внутрішніх детермінант розвитку сучасної української мови, тобто пріоритетів у вимогах суспільної практики до її оновлення та пріоритетів у ресурсах мови для їх задоволення, уможливлює відповіді на такі конкретні питання: а) які чинники суспільної практики становлять зовнішню детермінанту оновлення української мови і сприяють появі в її лексиконі активних і стабільних інновацій; б) які питомі й запозичені мовні ресурси формують внутрішню детермінанту оновлення української мови; в) як і якою мірою вторгнення нових номінацій порушує

рівновагу в системі мови, впливає на стратифікаційну організацію лексикону; г) які регулятори сприяють підтриманню стійкої рівноваги в лексиконі та відновленню його нестійкої рівноваги» [Карпіловська 2008: 149].

Вивчення розвитку сучасного українського лексикону в аспекті взаємозв'язку тексту, як вияву комунікативного процесу, і системи мови є методологічною тенденцією низки глибоких мовознавчих розвідок, проведених колективом відділу лексикології, лексикографії та структурно-математичної лінгвістики Інституту української мови НАН України, серед яких монографія «Динамічні процеси в сучасному українському лексиконі» [Клименко 2008] та інтегральний словник «Активні ресурси сучасної української номінації: ідеографічний словник нової лексики» [АРСУН 2013]. Ці фундаментальні лінгвістичні розвідки проведено за матеріалами комп'ютерного фонду інновацій у сучасній українській мові [Карпіловська 2007, 2008] та морфемно-словотвірного фонду. Мета словника «Активні ресурси сучасної української номінації: ідеографічний словник нової лексики» – «описати нову українську лексику, що виявляє системотвірні ознаки, здатність до творення лексичних об'єднань: словотвірних гнізд і рядів, синонімічних рядів, антонімічних опозицій (пар або тріад), демонструє широкий спектр словосполук, а отже, виявляє розгалужені парадигматичні, синтагматичні й епідигматичні (деривaційні) відношення в тексті та в системі мови. Така лексика виказує ознаки усталення в мовній свідомості носіїв української мови та в мовній діяльності сучасного українського суспільства, а засоби її творення можна вважати активними ресурсами сучасної української номінації» [АРСУН 2013: 8].

Л. Кислюк перераховує низку актуальних питань у галузі сучасної української словотвірної номінації, які мають вирішення тільки у корпусноорієнтованих розвідках: «Актуальність вивчення словотвірної номінації як провідного способу оновлення лексикону сучасної української мови зумовлено пошуками відповіді на питання: 1) як та які словотвірні ресурси підтримують типологічні риси українськомовної номінації в нових суспільно-політичних обставинах побутування української мови, забезпечують її нові функції в статусі мови держави; 2) якою мірою в словотворенні діють механізми захисту самобутності української номінації в умовах сучасних процесів євроінтеграції та глобалізації; 3) які соціальні, когнітивні й комунікативні чинники впливають на вибір і реалізацію словотворчих ресурсів системи української мови в сучасній колективній, загальній (узус) та індивідуальній (ідіолект) мовній практиці; 4) як сучасна мовна практика – загальна та індивідуальна – впливає на систему мови (словотвір) та сучасну українську словотворчу норму» [Кислюк 2017: 10 – 11].

Вивчення словотвірної номінації на матеріалі корпусу текстів розширює вектор дослідження неологічних та динамічних процесів у системі мови статистичними даними, які є обов'язковою умовою системного опису функціонування мови в синхронії та діахронії. Враховуючи досвід

комп’ютерного моделювання морфемної структури слова при укладанні «Хронологічного морфемно-словотвірного словника російської мови» [Поликарпов 1998], колектив лабораторії загальної і комп’ютерної лексикології і лексикографії філологічного факультету МДУ розробив методику автоматичної морфемної анотації у Полістильовому корпусі російської мови [Кукушкина 2005], яка була використана авторами також у комплексній тексто-аналітичній системі «СтилеАналізатор-2» [СтилеАнализатор-2 2014]. Застосування автоматичного морфемного аналізу у процедурі сегментації слововживань російського тексту дозволяє кожному слововживанню приписати корінь слова і його афіксальну модель, а також автоматично укласти частотні словники коренів, афіксів та афіксальних моделей. На матеріалі даних цих частотних словників було проведено глибокі лінгвістичні дослідження морфемної та словотвірної будови російської лексики, що відзначаються комплексним аналізом взаємодії чинників формування нової лексики та системністю використання статистичних даних [Кукушкина 2006], [Кукушкина 2006а].

Ґрунтуючись на методології системного підходу Г. Мельникова [Мельников 1978] у визначенні цілісності мовної системи[26], А. Полікарпов формулює поняття "моделі життєвого циклу знака" і визначає механізми лексико-семантичного і словотвірного процесів важливими складники мовної комунікації в синхронії і діахронії: «Дослідження мови на макрорівні її організації передбачає врахування механізмів взаємного узгодження дії всіх тих циклічних процесів, які відбуваються на мікрорівні життя мови. Зокрема, це передбачає розгляд закономірностей ансамблевої поведінки в часі знаків тієї чи іншої мови, розгляд історичної динаміки лексичного, а також морфемного і фразеологічного інвентарів мови як цілісних підсистем, що розвиваються взаємообумовлено» [Поликарпов 2013: 680]. Досліджуючи життєвий цикл мовного знака на матеріалі корпусу газетних текстів, А. Полікарпов робить висновок, що «...основне джерело словотвірного процесу є похідним від більш фундаментального процесу базових семантичних змін будь-якої лексеми, морфеми і фраземи, що ймовірно супроводжують кожен акт їх вживання» [Поликарпов 2013: 701]. Системне вивчення мовних явищ різних рівнів мовної системи, зокрема морфемних і деривaційних процесів, крізь призму функціональної цілісності організації системи мови можливе лише «...на основі аналізу спеціально спланованих, особливим чином структурованих і керованих на основі спеціалізованих баз даних комп’ютерних корпусів текстів. Саме корпусна організація матеріалу дозволяє усвідомлено планувати й тримати під контролем відібраний матеріал (як тексти, так і їх одиниці)...» [Поликарпов 2013: 702].

[26] Г. Мельников у монографії «Системология и языковые аспекты кибернетики» [Мельников 1978] визначає цілісність мовної системи не тільки як структуру цілого, а і як взаємозв’язок властивостей елементів, що спричинений самим фактом існування системи-цілого.

Перспективи корпуснооріентованого підходу у вивченні словотвірних процесів представлені розробниками автоматичного морфемно-словотвірного аналізу в Національному корпусі російської мови [НКРЯ 2018]. «До цього часу продуктивність словотворення в тексті й мові вивчалася в основному в стилістичному аспекті (Виноградова 1984 та ін.). Однак, як здається, цікаво було б проаналізувати, як реалізація словотвірних моделей у тексті пов'язана з реалізацією інших конструкцій; як одні словотвірні моделі поєднуються з іншими; як відрізняється поведінка дієслівних й іменних коренів; яким чином однокореневі слова задіяні у встановленні кореференції; яка частотність тієї чи іншої моделі в корпусі в цілому або в тому чи іншому жанрі; також цікаво було б простежити мікрозміни в процесах словотворення (наприклад, яка швидкість залучення до словотворення нових слів та ін.). Усі ці можливості може надати словотвірна розмітка корпусу, реалізована із використанням електронного морфемно-словотвірного словника і забезпечена пошуковою системою» [Ляшевская 2016: 211]. На матеріалі НКРМ було проведено лінгвостатистичні дослідження у галузі морфеміки та словотвору, зокрема розвідки С. Татевосова [Татевосов 2009], А. Пазельської [Пазельская 2009а], [Пазельская 2009], які демонструють ефективний добір матеріалу дослідження та можливість вивчення словотвірних і морфемних процесів у кореляції із синтаксичною валентністю та семантикою описуваних лексем.

Параметризація тексту на морфемному рівні в корпусах також відкриває широкі можливості для нових стилеметричних досліджень. Традиційно стилістичні дослідження зосереджували увагу переважно на продуктивності словотворення [Виноградова 1984], хоча уже в 1967 р. у монографії «Статистичні параметри стилів» [СПС 1967], було встановлено частоту префіксів як статистичний параметр стилю. За даними електронних частотних морфемних словників у Корпусі української мови було проведено низку стилеметричних досліджень ідіостилю українських поетів на морфемному рівні організації тексту [Зубань 2014], [Зубань 2014а], [Зубань 2016], [Zuban 2019a]. Ці дослідження підтверджують значущість у стилеметричній моделі авторського стилю статистичних даних як морфемних одиниць (коренів, префіксів, суфіксів), так і морфемних структур слів (морфемна довжина слова, морфемна модель слова).

Крім корпуснооріентованих лінгвістичних досліджень, автоматичний морфемний аналіз тексту відкриває нові перспективи у створенні систем автоматичного розпізнавання змісту тексту. Колектив науковців РосНДІ штучного інтелекту під керівництвом О. Пацкіна [Пацкин 2002], [Пацкин 2004] створив електронний гіперсловник «Аріадна», який систематизує морфемну, словотвірну та словозмінну інформацію на базі лексичного обсягу «Грамматического словаря русского языка» [Зализняк 1977] та списків морфем «Словаря морфем русского языка» [Кузнецова 1986]. Словник «Аріадна» пов'язаний із системою представлення знань «Абріаль-2» і призначений для побудови семантичної мережі глибокого семантичного

аналізу російського тексту. Створення резидентного словника системи на базі морфем зумовило значне скорочення його обсягу й підвищення ефективності роботи системи. «Врахування словотвірних зв'язків і морфемної структури російських слів може дати значні переваги при побудові семантичних мереж в обсязі всієї мови. Обсяг необхідних семантичних описів і, відповідно, роботи комп'ютерних лексикографів, за умови пріоритетного опису морфем перед лексемами, може скоротитися на порядок, ураховуючи середню продуктивність російського кореня рівну приблизно 13 (за А.М. Тихоновим). А з урахуванням того факту, що найбільш частотні корені одночасно і найбільш продуктивні (потужність "нести" дорівнює 540), важко переоцінити ефективність пріоритетного семантичного опису морфем, для отримання компактного опису семантики, використовуваної в російському ПМ-аналізі» [Пацкин 2002].

У сучасних системах автоматичного оброблення тексту можна визначити два рівні автоматичного морфемного аналізу:

1) формальне опрацювання морфемної будови слова без врахування значення морфем: такий аналіз розроблено в Корпусі української мови, Національному корпусі російської мови та Полістильовому корпусі текстів сучасної російської мови;

2) семантичне опрацювання морфемної будови слова: побудова семантичної мережі на основі входження морфем у класи онтологічної мережі (система «Абріаль-2»).

Активне використання корпусних морфемних та словотвірних даних у лінгвістичних розвідках різного спрямування засвідчує необхідність морфемно-словотвірної анотації в дослідницьких корпусах мов. Поглиблення морфемно-словотвірної параметризації в корпусах текстів інформацією про семантичну організацію морфемної структури слова відкриває нові перспективи для корпусноорієнтованих досліджень у галузі морфеміки та словотвору, зокрема в завданнях розпізнавання змісту морфем і моделювання логіко-семантичних відношень між ними в структурі слова, системі мови та в тексті. Описати семантику слів через значення морфем можна більш точно і більш системно, ніж приписувати кожному слову його тлумачення чи таксономічну характеристику, адже узуальна семантика слів частково експлікована морфемами, які виступають носіями семантичних множників семеми й репрезентують семи абстрагованої узагальненої семантики.

3.2.2. База даних частотних морфемних словників: структура та процедура укладання

Система АСМСА з початку свого створення орієнтувалася на проведення автоматичної морфемної сегментації текстових слововживань. Використання АСМСА в процедурі автоматичної морфемної сегментації слововживань у параметризованій базі даних поетичної мови [Алексеенко 2004], створеної колективом лабораторії комп'ютерної лінгвістики в 2004 р.,

демонструє досвід автоматичної морфемної анотації текстових слововживань.

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

Рис. 3.12(1). Фрагмент БД морфемної сегментації текстових слововживань

На рис. 3.12(1) показано морфемну сегментацію слововживань текстового фрагмента *...за щитом смарагдових лісів*. Кожне слововживання автоматично сегментується на морфеми (4-та колонка) за процедурою автоматичного приписування двох програмних процедур (5-та колонка): перша процедура – модель морфемної сегментації початкової словоформи лексеми (леми) за МБД АСМСА; друга процедура – модель морфемної сегментації слововживання текстового фрагмента. Друга модель автоматично формується через зіставлення даних МБД АСМСА із даними морфологічної БД системи АГАТ.

У процесі створення автоматичного морфемного аналізу в Корпусі української мови з метою оптимізації пошуку на великих текстових масивах ми відмовилися від методу морфемної анотації текстових слововживань. Морфемна розмітка не проводиться, тексти Корпусу виступають тільки матеріалом для укладання різних частотних морфемних словників за допомогою системи АСМСА, яка виконує функцію морфемного модуля-аналізатора в Корпусі української мови.

На сьогодні укладено 20 електронних ЧМС, які входять до автономних лексикографічних систем, що представляють частотні словники одиниць різних рівнів структури тексту: початкових форм (лем), словоформ, морфем, морфемних структур слів, а деякі ще й словосполучень. Як уже зазначалось, найглибшу параметризацію має словник мови Тараса Шевченка [ЕСМТШ 2017. БД частотних морфемних словників, укладених за текстовою вибіркою Т. Шевченка, буде розглянуто в цьому параграфі.

В. Перебийніс та В. Сорокін, описуючи процес автоматичного укладання частотного словника, обґрунтовують необхідність розроблення еталонної схеми вихідної вибірки даних частотного словника, яка забезпечить здійснення різноманітних статистичних досліджень на базі параметризованого тексту: «При сучасному розвиткові комп’ютерної техніки

укладання ЧС може стати основою багатофункціональної науково-дослідної інтерактивної системи, яка уможливить дослідження не тільки статистичних особливостей лексичного складу обстежуваної вибірки, але й морфологічних категорій, синтаксичних структур, семантичного складу вибірки. Іншими словами, слід уніфікувати підхід до побудови баз даних таким чином, щоб одержати дослідницький комплекс, здатний надати різноманітну інформацію з однієї вибірки й порівнювати різні вибірки» [Перебийніс 2009: 60].

У процедурі конструювання лексикографічної системи електронних частотних словників наш колектив керувався саме цим концептуальним завданням – створити даталогічну модель вихідної вибірки текстових даних, яка б систематизувала текстові дані у такий спосіб, щоб на її основі можна було автоматично конструювати частотні словники лексем (лем), словоформ, морфем, морфемних структур, словосполучень та семантичних таксонів. Для кожної текстової вибірки було створена БД "temp_freq" (рис. 3.13), яка систематизує вихідні текстові дані для автоматичного укладання кожного типу словників і входить до схеми даних внутрішньої даталогічної моделі кожного словника.

Для укладання та роботи електронних ЧС лексем, словоформ, морфем і морфемних структур слів за поетичними текстами Т. Шевченка було створено єдину для всіх ЧС лексикографічну БД, яка має реляційний характер і складається із 10 БД-таблиць (рис. 3.13).

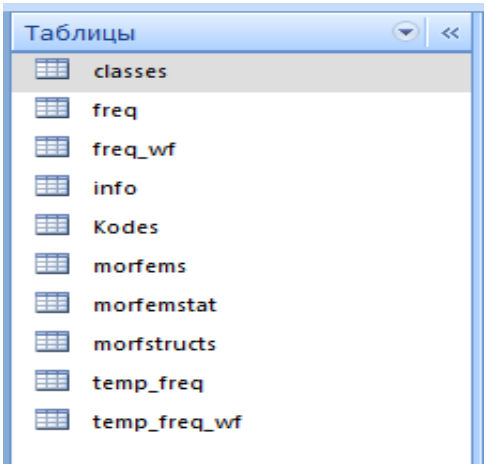


Рис.3.13. Структура лексикографічної БД частотних словників

Парметризація вихідної вибірки текстових даних у процесі укладання частотних словників у Корпусі української мови здійснюється автоматично на етапі проведення морфологічної анотації тексту. На виході модуля автоматичного морфологічного аналізу кожному слововживанню тексту приписується двосимвольний граматичний код (додаток 2), лема та номер речення в аналізованому тексті (додаток 10). Наприклад, текстовий сегмент *Реве та стогне Дніпр Широкий* у результаті проведення морфологічної анотації буде мати таку структуру даних:

```
Реве,ГЯ,ревти,1
```

```
та,СС,та,1
стогне,ГЯ,стогнати,1
Дніпр,йИ,Дніпр,1
широкий,АИ,широкий,1
",",,",",1
```

У такий спосіб параметризується вибірка текстів Т. Шевченка обсягом ~ 60920 слововживань. На основі цієї вибірки автоматично укладаються дві БД-таблиці: табл. 3.1 систематизує вихідні текстові дані слововживань, а табл. 3.2. текстові дані лем.

Таблиця 3.1. Фрагмент таблиці текстових слововживань "temp_freq_wf"

Code	wrd	cls	vib	tid	sentnum
1	реве	ГЯ	1	10295	1
2	та	СС	1	10295	1
3	стогне	ГЯ	1	10295	1
4	Дніпр	йИ	1	10295	1
5	широкий	АИ	1	10295	1

Таблиця 3.2. Фрагмент таблиці початкових форм (лем) текстових слововживань "temp_freq"

temp_freq					
Code	wrd	cls	vib	tid	sentnum
1	ревти	Г	1	10295	1
2	та	С	1	10295	1
3	стогнати	Г	1	10295	1
4	Дніпр	й	1	10295	1
5	широкий	А	1	10295	1

У процесі автоматичного укладання електронних морфемних словників за вихідними даними двох БД (табл. 3.1 та табл 3.2) можна визначити таку послідовність етапів роботи:

1) систематизація лексичних даних текстової вибірки з метою подальшого визначення одиниць підрахунку та здійснення статистичних обчислень у частотних морфемних словниках;

2) укладання бази даних частотного словника початкових форм;

3) виділення у початкових формах одиниць підрахунку (морфем та морфемних структур слів) і визначення функціональної характеристики виділених морфем: префікс, корінь, суфікс, інтерфікс, флексія;

4) обчислення частотних характеристик морфем та морфемних структур слів;

5) укладання алфавітного та рангового списку частотних словників морфем та морфемних структур слів;

6) створення людино-машинного інтерфейсу ЧМС та пошукових запитів до бази даних ЧМС.

Етапи 1 – 5 спрямовані на автоматичне укладання лексикографічної БД ЧМС, яка представляє внутрішню даталогічну лексикографічну модель. Завдання (6) спрямоване на використання укладеної лексикографічної БД ЧМС і виконується за схемою зовнішньої даталогічної лексикографічної моделі. Роботу двох типів електронних ЧМС (ЧС морфем та ЧС морфемних структур слів) забезпечують 6 таблиць лексикографічної БД (рис. 3.13). Ці таблиці укладаються автоматично за послідовністю визначених етапів і утворюють схему даних, представлену на рис. 3.14.

За моделлю організації даних БД частотних морфемних словників визначається як реляційна реляційна база даних. СКБД бази даних ЧМС розроблена на основі програми MS Access, яка здійснює конструювання таблиць БД та збереження даних; а також система керування використовує програмне забезпечення (мова програмування C #), розроблене В. Сорокіним.

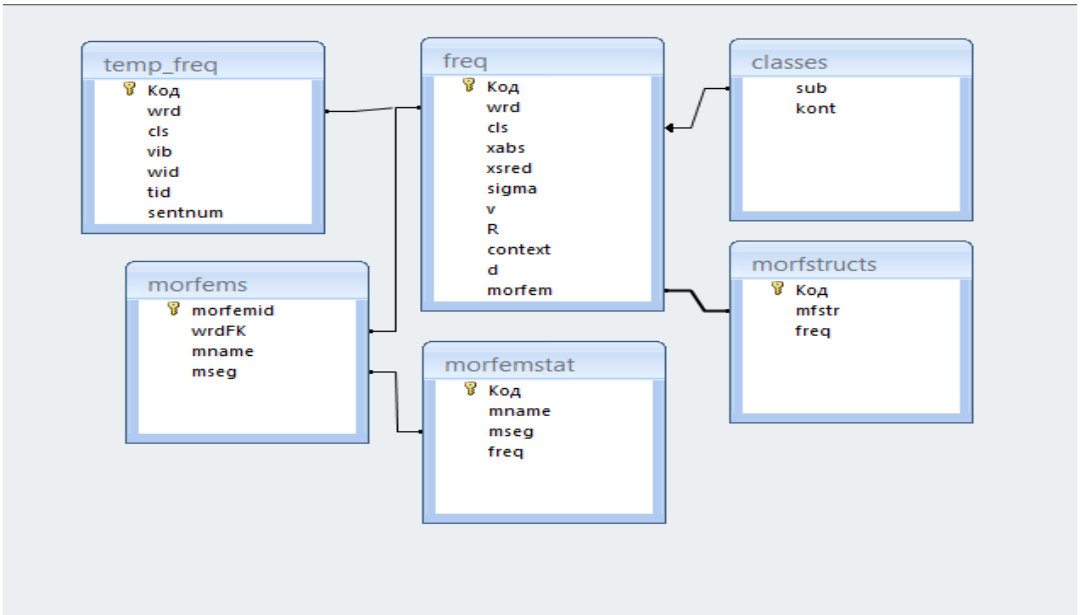


Рис. 3.14. Реляційна діаграма схеми даних частотних морфемних словників

(1) Таблиця "temp_freq" виконує функцію систематизації лексичних та текстових даних текстової вибірки для подальшого автоматичного оброблення: визначення одиниць підрахунку та здійснення статистичних обчислень. Таблиця "temp_freq" (додаток 11, табл. 3.2) автоматично укладається за запитами до морфологічно анотованого тексту і систематизує таку інформацію:

1) поле "Код" – номер слова у таблиці;

2) поле "wrd" – лема кожного текстового слововживання (*ревти*), записана в окремий рядок таблиці за послідовністю цього слововживання у реченні;

3) поле "cls" – перший символ двосимвольного граматичного коду, який визначає частину мови (Г – дієслово);

4) поле "vib" – номер текстової підвибірки (вся вибірка поетичних текстів Т. Шевченка автоматично, за першим полем таблиці, була поділена на 61 підвибірку обсягом 1000 слововживань кожна, слово *ревти* належить до 1-ої підвибірки);

5) поле "tid" – 10295 (номер тексту в Корпусі української мови);

6) поле "sentnum" – номер речення в тексті твору.

Використовуючи текстовий фільтр поля "wrd", можна для кожної леми автоматично сформувати за полями таблиці картку (окрему таблицю), яка використовується для статистичних обчислень наступної таблиці "freq" (табл. 3.3). Лема записується у таблиці стільки разів, скільки слововживань цієї леми зустрілось у текстовій вибірці.

Таблиця 3.3. Систематизовані дані за лемою *ревти* таблиці "temp_freq"

temp_freq					
Код	**wrd**	**cls**	**vib**	**tid**	**sentnum**
1	ревти	г	1	10295	1
4079	ревти	г	5	10306	21
7209	ревти	г	8	10311	49
8054	ревти	г	9	10314	82
8057	ревти	г	9	10314	83
10844	ревти	г	11	10317	60
13702	ревти	г	14	10323	121
13962	ревти	г	14	10325	9
14027	ревти	г	15	10325	17
14122	ревти	г	15	10325	29
14174	ревти	г	15	10325	37
14325	ревти	г	15	10325	55
14235	ревти	г	15	10325	44
14238	ревти	г	15	10325	44
16256	ревти	г	17	10334	6
23311	ревти	г	24	10373	2
23736	ревти	г	24	10376	9
30083	ревти	г	31	10418	3
30085	ревти	г	31	10418	4
29266	ревти	г	30	10406	6
44939	ревти	г	45	10521	1
59391	ревти	г	60	14793	37
∑	22				

Необхідність укладання реєстру початкових форм (лем) текстових слововживань зумовлена онтологічними особливостями системно-структурної організації мови: морфеми утворюють морфемну структуру слова й поза словом не існують. Тому для визначення одиниць підрахунку

(морфем та морфемних структур слів) у ЧМС необхідно визначити одиниці вищі за рангом ієрархії – слова, які розглядаються як контекст для реалізації морфеми, розуміння її значення та розмежування морфемної омонімії, адже ЧМС укладаються для морфем, визначених на формальному рівні без врахування значення морфеми.

Частотний морфемний словник можна укладати за морфами, визначеними як у слововживаннях тексту, так і в початкових формах слова, що представляють лексему як одиницю мови – традиційний методологічний принцип морфемної лексикографії. Оскільки автоматичне визначення морфем у ЧМС передбачає використання МБД АСМСА, яка систематизує морфемні структури початкових форм слів, то інфологічна й даталогічна моделі проектування бази даних ЧМС визначає пріоритетним методологічний підхід морфемної сегментації лексики текстової вибірки на базі початкових форм текстових слововживань. Тому вихідними даними в укладанні ЧМС служить таблиця "temp_freq", що систематизує початкові форми в послідовності їх фіксації в параметризованих текстах.

(2) Таблиця "freq" (додаток 12, табл.3.4) систематизує алфавітний реєстр початкових форм текстових слововживань та їх статистичні характеристики.

Таблиця 3.4. Фрагмент даних таблиці "freq" для леми *ревти*

Code	wrd	cls	xabs	xsred	sigma	v	R	d	morfem
5501	ревти	Г	22	0,360655737704918	0,923869036301992	2,56163687338279	14	6,692941	RF

Таблиця "freq" (рис. 3.14) автоматично укладається за даними таблиці "temp_freq" (відношення між таблицями здійснюється через поле "wrd" обох таблиць) і систематизує такі дані:

1) поле "Код" – номер слова в таблиці;

2) поле "wrd" – алфавітний реєстр початкових форм, укладений за процедурою автоматичної ідентифікації однакових лем текстових слововживань таблиці "temp_freq" (60920 лем текстових слововживань представлено 6889 початковими формами реєстру);

3) поле "cls" – перший символ граматичного коду, який визначає частину мови;

4) поле "xab" – абсолютна частота слова за вибіркою поетичних текстів Т. Шевченка: f обраховується автоматично за кількістю однакових

лем таблиці "temp_freq" (наприклад, *ревти* має абсолютну частоту 22, це означає, що лема *ревти* зустрічається у полі "wrd" таблиці "temp_freq" 22 рази (табл 3.3);

5) поле "xsered" – середня частота слова за вибіркою поетичних текстів Т. Шевченка (*ревти* $\overline{x}$ =0,360655737704918), що обраховується автоматично: визначена у попередньому полі абсолютна частота ділиться на кількість підвибірок, на які розбита вибірка текстів у таблиці "temp_freq", наприклад, для обчислення середньої частоти слова *ревти* (0,360655737704918) абсолютна частота 22 ділиться на 61;

6) поле "sigma" – середнє квадратичне відхилення, яке обчислюється за формулою $\sigma = \sqrt{\frac{\sum(x_i-\bar{x})^2 \cdot n_i}{\sum n_i}}$: формула програмується за показником середньої частоти попереднього поля та даними поля "vib" у картці кожної леми таблиці "temp_freq" (табл.3.3), яке систематизує дані про розподіл абсолютної частоти леми у підвибірках текстової вибірки. За цим полем визначаються варіанти абсолютної частоти (x_i) та кількість підвибірок (n_i), у яких зустрічається ця варіанта. Наприклад, для леми *ревти* за даними поля "vib" встановлюються такі дані:

(x_i)	0	1	2	3	4	5	6
(n_i)	48	8	4	0	0	0	1

Значення σ для леми *ревти* становить 0,923869036301992.

7) поле "v" – коефіцієнт варіації, який обчислюється автоматично за формулою: $V = \frac{\sigma}{\bar{x}}$, де статистичні показники σ та $\overline{x}$ беруться із попередніх полів. Значення *V* для леми *ревти* становить 2,56163687338279;

8) поле "R" – кількість творів текстової вибірки корпусу, у яких зустрілась лема (лема *ревти* за вибіркою поетичних текстів Т. Шевченка зустрічається в 14 творах);

9) поле "d" – коефіцієнт стабільності, який обчислюється автоматично за формулою: $D = 1 - \frac{V}{\sqrt{n-1}}$, де статичний показник *V* береться із поля "v", а n – кількість підвибірок (61). Значення *D* для леми *ревти* становить 6,692941;

10) у полі "morfem" за даними МБД АСМСА (див. Розділ 2) через зіставлення початкової форми (леми) таблиці "freq" із початковою формою МБД записується модель морфемної структури слова (для леми *ревти* – RF).

Як свідчить фрагмент таблиці "freq" (додаток 12), у полі "morfem" модель морфемної структури слова визначена не для всіх слів реєстру, наприклад, такі слова, як *ревти-завивати, регот, реєстер, решотка, Ржавиця, рибалонька, ридати-молитися* та ін. не мають моделі морфемної структури. За фільтром поля "morfem" таблиці "freq" автоматично формується вибірка таких слів (як правило, це низькочастотна лексика), і вони, за потреби, додаються до реєстру МБД. За текстовою вибіркою поетичних творів Т. Шевченка було автоматично сформовано таку лексичну вибірку обсягом 1560 слів, тоді як загальний реєстр початкових форм лексикону Т.Шевченка становить 6889 слів. Отже, 22,6 % лексичного масиву

текстової вибірки Т. Шевченка не обробляються автоматично в морфемному аналізаторі корпусу текстів і потребують окремого морфемного аналізу. Це свідчить про те, що ці слова не були зафіксовані словниками української мови і є або помилками, або архаїзмами, або діалектизмами, або неологізмами, або оказіоналізмами, або потенційними словами, або регулярними формами (демінутивами, аугментивами, ступенями порівняння), або власними назвами, або абревіатурами, або з якихось інших причин не ввійшли до реєстру МБД. Вибірка таких слів є надзвичайно важливим матеріалом для редагування МБД. У додатку 13 подано вибірку (А-Б) слів із невизначеною морфемною будовою.

Як показує реляційна діаграма схеми даних (рис. 3.14), таблиця "freq" є центральним об'єктом БД ЧМС: вона має зв'язки з полями інших таблиць, тому ця таблиця є фінальною систематизацією вхідних текстових даних, за якими конструюються морфемні частотні словники за такими правилами процесу укладання даталогічної внутрішньої моделі:

1) морфемне сегментування слів лексичної вибірки текстів Т. Шевченка здійснюється за алфавітним реєстром початкових форм слів, який порівняно із таблицею "temp_freq" набагато зменшився: 60920 лем текстових слововживань таблиці "temp_freq" представлено 6889 початковими формами реєстру таблиці "freq", з яких для наступного етапу укладання ЧМС вибрано тільки 5329, отже, слово сегментується тільки один раз, а не стільки разів, скільки воно зустрілось у тексті;

2) обчислення статистичних характеристик морфем та морфемних структур слів здійснюється за встановленими статистичними даними слів, тому що абсолютна частота вживання морфеми у тексті дорівнює сумі абсолютної частоти вживання всіх слів, у яких визначена ця морфема.

(3) Таблиця "morfems" (додаток 14) автоматично укладається за даними поля "wrd" таблиці "freq" та МБД АСМСА. Відношення між таблицями здійснюється через зв'язок між полями "wrdFK" таблиці "morfems" та "Код" таблиці "freq" (рис. 3.14).

Таблиця 3.5. Фрагмент даних таблиці " morfems " для леми *ревнути*

morfems			
morfemid	**wrdFK**	**mname**	**mseg**
12341	5078	R	рев
12342	5078	S	ну
12343	5078	F	ти

Таблиця "morfems" (табл. 3.5) систематизує дані про морфемну сегментацію слів:

1) поле "morfemid" – порядковий номер визначеного у початковій формі слова морфа (за вибіркою поетичних текстів Т.Шевченка 6889 початкових форм слів просегментовані на 16258 морфів);

2) поле "wrdFK" – номер слова у полі "Код" таблиці "freq": номер переноситься разом із початковою формою слова й записується в полі "wrdFK" стільки разів, скільки морфем визначається в початковій формі.

3) поле "mname" – функціональний тип морфеми: P – префікс, R – корінь, S – суфікс, F – флексія, I – інтерфікс, X – постфікс;

4) поле "mseg" – морфемна сегментація початкових форм слів (подані за алфавітом), яка записується в таблиці вертикально: кожен визначений морф – окремий рядок таблиці. Інформація про належність морфів до одного слова подана через номер слова в полі "wrdFK".

Наприклад, морфемна сегментація 3-ох слів – *ревнути, ревти, ревучий* – подана у таблиці 8-ома рядками:

morfemid	wrdFK	mname	mseg
12341	5078	R	рев
12342	5078	S	ну
12343	5078	F	ти
12344	5079	R	рев
12345	5079	F	ти
12346	5081	R	рев
12347	5081	S	уч
12348	5081	F	ий

Три рядки (12341-1243) представляють 3 морфи *-рев-, -ну-,* та *-ти-* (4 поле), із яких 1-ий морф – R (корінь), 2-ий морф – S (суфікс), 3-ій морф – F (флексія). У 2-ому полі "wrdFK" морфам *-рев-, -ну-,* та *-ти-* відповідає однаковий номер 5078, який записано тричі: це свідчить, що визначені морфи утворюють одне слово *ревнути*. Номер 5079 об'єднує два морфи слова *ревти*. Номер 5081 об'єднує 3 морфи слова *ревучий*.

Морфемна сегментація початкових форм здійснюється автоматично за допомогою розробленого програмного забезпечення мовою програмування C #; фрагмент коду програми:

```
bool bitenwrd(string wrd, string morfem, string wrdid, OleDbConnection con)
{
bool w = false;
int st = 0, nd = 0;

OleDbCommand cmd = new OleDbCommand("", con);
for (int i = 0; i < morfem.Length; i += 2)
{
if (morfem[i] == '/' || morfem[i] == '|')
{
break;
}
nd = (int)morfem[i + 1] - (int)'A';
if (nd <= wrd.Length)
{
string seg = wrd.Substring(st, nd - st);

if (seg == "") seg = "0";
cmd.CommandText = "INSERT INTO morfems (wrdFK,mname,mseg) SELECT " + wrdid + ",'" + morfem[i] + "','" +
seg.Replace("'", "''") + "'";
cmd.ExecuteNonQuery();
}
st = nd;
}
return w;
}
```

Із реєстру початкових форм таблиці "freq" за кожним порядковим номером береться слово, яке зіставляється із цим самим словом морфемної бази даних АСМСА:

5078 ревнути ↔ ревнути,Г,RDSFFH;

5079 ревти ↔ ревти,Г,RDFF;

5081 ревучий ↔ ревучий,А,RDSFFH.

За моделлю програмної процедури сегментації слова на морфи, взятої із МБД АСМСА, відбувається сегментація графемного запису слова (з урахуванням перекодування йотованих я, ю, є, ї у дві графеми). Модель програмної процедури представляє комбінацію двосимвольних кодів, у яких перший символ – функціональний тип морфеми, а другий символ – кількісно-графемна правобічна межа морфа у слові, наприклад, програмна процедура RDSFFH:

RD – R (корінь), D (3-тя графема слова *рев-нути*);

SF – S (суфікс), F (5-та графема слова *ревну-ти*);

FH – F (флексія), H (7-ма графема слова *ревнути*).

Перший символ коду записується в окремий рядок поля "mname", за другим символом коду програма, попередньо порахувавши графеми у слові, розділяє його на частини за порядковим номером графеми, відділення графемного ланцюжка відбувається у постпозиції до визначеної за номером графеми. Кожна частина поділеного слова є морфом і записується в окремий рядок поля "mseg":

morfemid	**wrdFK**	**mname**	**mseg**
12341	5078	R	рев
12342	5078	S	ну
12343	5078	F	ти

Якщо слово має нульовий суфікс чи флексію, які закінчують слово, то вони упускаються при автоматичній морфемній сегментації. Наприклад, у слові *ревнитель*,Й,RESJFJ, за програмною процедурою у МБД дописується нульова флексія як додатковий символ у графемному записі слова, що кодується двосимвольним кодом FK. Коли слово *ревнитель* сегментується при укладанні таблиці "morfems", ця процедура не здійснюється, тому що 10-ї графеми у слові *ревнитель* нема, і тому програма визначає кінець слова на 9-ій графемі, а отже, це слово буде мати морфемну сегментацію на два морфи (корінь і суфікс), а флексії не буде:

morfemid	**wrdFK**	**mname**	**mseg**
12336	5076	R	ревн
12337	5076	S	итель

Це великий недолік автоматичної морфемної сегментації, над виправленням якого колектив сьогодні працює.

Якщо нульовий суфікс стоїть у позиції перед фонемно вираженою флексією, наприклад, *без/верх//ий* (безверхий,А,PDRHSHFJ), то нульовий суфікс дописується за процедурою морфемної сегментації, тому що графемний ланцюжок слова не закінчився, і програма продовжує морфемну сегментацію:

morfemid	wrdFK	mname	mseg
200	149	P	без
201	149	R	верх
202	149	S	0
203	149	F	ий

У такий спосіб у таблиці "morfems" у полі "mseg" формується реєстр морфів, на які поділена кожна початкова форма слова, взята із поля "wrd" таблиці "freq" із збереженим номером цього слова в полі "wrdFK". До реєстру відбираються лише ті початкові форми таблиці "freq", яким приписана модель морфемної будови, тобто із 6889 початкових форм таблиці "freq" до таблиці "morfems" входять 5329 початкових форм слів, а 1560 слів не переносяться.

(4) Таблиця "morfemstat" (додаток 15) представляє алфавітно-частотний та ранговий список морфем за поетичними текстами Т. Шевченка. Таблиця укладалася автоматично за даними таблиці "morfems" та таблиці "freq". Відношення між таблицями (рис. 3.14) здійснюється через послідовність таких зв'язків: 1) поле "mseg" таблиці "morfemstat" зв'язано з полем "mseg" таблиці "morfems"; 2) поле "wrdFK" таблиці "morfems" зв'язано із полем "Код" таблиці "freq"; 3) поле "Код" таблиці "freq" зв'язано із іншими полями цієї таблиці, зокрема і з полем "xab", у якому подано інформація про абсолютну частоту слова у вибірці поетичних текстів Т. Шевченка.

Таблиця 3.6. Фрагмент вибірки кореневих морфем таблиці "morfemstat"

morfemstat			
Code	**mname**	**mseg**	**freq**
1006	R	раз	19
1007	R	ранн	2
1008	**R**	**рев**	**41**
1009	R	регот	12

Таблиця "morfemstat" (табл. 3.6) систематизує такі дані:

1) поле "Код" – порядковий номер рядка таблиці;

2) поле "mname" – функціональний тип морфеми: P – префікс, R – корінь, S – суфікс, F – флексія, I – інтерфікс, X – постфікс;

3) поле "mseg" – алфавітний реєстр морфів, визначених у початкових формах слів і записаних вертикально: кожен визначений морф – окремий рядок таблиці;

4) поле "freq" – абсолютна частота морфа у вибірці поетичних текстів Т.Шевченка: за фільтром цього поля формується ранговий список морфем за ростом або спадом абсолютної частоти;

Поля "mname" та "mseg" укладаються за аналогічними полями таблиці "morfems". Морфи у полі "mseg" таблиці "morfemstat" записуються не стільки разів, скільки вони зустрілись у просегментованих початкових формах, а тільки один раз за послідовністю їх появи у словах в алфавітному порядку,

наприклад, якщо корінь -*ангел*- зустрівся перший раз у слові *ангеляточ*ко, то він більше не буде повторюватися у полі цієї таблиці, тоді як у таблиці"morfems" він записаний у полі "mseg" двічі: у словах *ангел* та *ангеляточ*ко. Морфи з однаковою формою вираження ідентифікуються в одну одиницю й записуються один раз. Відповідно, реєстр морфів у таблиці "morfemstat" становить лише 2681 одиницю, тоді як у таблиці "morfems" 16258 одиниць.

Абсолютна частота кожного морфа встановлюється за послідовністю таких операцій: кожен морф таблиці "morfemstat" пов'язаний із вибіркою ідентичних за формою вираження морфів таблиці "morfems", яка автоматично укладається за фільтром поля "mseg". Через номер у полі "wrdFK" відбувається зв'язок із таким самим номером поля "Код" таблиці "freq", де кожному номеру (початковій формі слова) у полі "xab" відповідають показники абсолютної частоти, які сумуються, а сума автоматично записується у поле "freq" таблиці "morfemstat". Наприклад, обчислення для вибірки кореневого морфа -рев-:

morfemid	wrdFK	mname	mseg
4516	1902	R	рев
4520	1903	R	рев
12333	5075	R	рев
12341	5078	R	рев
12344	5079	R	рев
12346	5081	R	рев

таблиця "morfems"

Код	wrd	xabs
1902	заревіти	1
1903	заревти	10
5075	ревіти	3
5078	ревнути	3
5079	ревти	22
5081	ревучий	2
		∑41

таблиця "freq"

Обчислення абсолютної частоти морфеми здійснюється за правилом: скільки разів зустрілись у тексті слова із визначеним морфом (сума різних слів з одним морфом), стільки разів зустрівся в тексті й визначений морф: сума слів із кореневим морфом -рев- 41, отже кореневий морф -рев- має абсолютну частоту 41.

(5) Таблиця "morfstructs" (додаток 16, табл. 3.7) автоматично укладається за таблицею "freq" і систематизує дані про абсолютну частоту морфемних структур слів.

Таблиця 3.7. Фрагмент таблиці "morfstructs"

morfstructs		
Код	**mfstr**	**freq**
152	PPRFX	9
153	PPRSF	204
154	RSRF	3
155	**PPRF**	**27**
156	RIF	3
157	PRIRS	4

1) поле "Код" – порядковий номер рядка таблиці;

2) поле "mfstr" – модель морфемної структури слова, у якій кожен морф представлений через символ функціонального типу морфеми: P – префікс, R – корінь, S – суфікс, F – флексія, I – інтерфікс, X – постфікс;

4) поле "freq" – ранговий список (за спадом або ростом) абсолютної частоти моделі морфемної структури слова у вибірці поетичних текстів Т.Шевченка.

Поле "mfstr" пов’язано із полем "morfem" таблиці "freq" (рис. 3.14). У цьому полі за фільтром вибору моделі морфемної структури слова формується вибірка всіх слів, у яких визначається обрана модель, наприклад, для моделі PPRF формується вибірка із 12 слів (табл. 3.8).

Таблиця 3.8. Вибірка з таблиці "freq" за моделлю PPRF

freq									
Код	wrd	cls	xabs	xsred	sigma	v	R	d	morfem
860	вспомин	Й	2	3,27868852459016E-02	0,253966121062781	7,74596669241483	1	1,110223E-15	PPRF
1329	донедавний	А	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	PPRF
1884	заповідь	К	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	PPRF
1886	запопасти	Г	3	4,91803278688525E-02	0,216244359971687	4,39696865275764	3	4,323538	PPRF
2039	здобути	Г	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	PPRF
2127	знайти	Г	5	8,19672131147541E-02	0,328687502553499	4,00998753115268	4	4,823128	PPRF
3424	недосвіт	Й	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	PPRF
5188	роздобути	Г	2	3,27868852459016E-02	0,1780783687082	5,43139024560011	2	2,988105	PPRF
5250	розповити	Г	2	3,27868852459016E-02	0,1780783687082	5,43139024560011	2	2,988105	PPRF
5766	сповити	Г	5	8,19672131147541E-02	0,274314762798058	3,3466401061363	5	5,679506	PPRF
5770	сповідь	К	2	3,27868852459016E-02	0,1780783687082	5,43139024560011	2	2,988105	PPRF
6204	увійти	Г	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	PPRF

У полі "xabs" цієї вибірки систематизовані дані про абсолютну частоту кожного слова. Ці дані сумуються, а сума записується у поле "freq" таблиці "morfstructs", наприклад, модель PPRF має абсолютну частоту 27 (додаток 16, Табл. 3.7).

(6) Таблиця "classes" (додаток 17) систематизує дані про значення граматичних кодів поля "cls" таблиці "freq" і пов’язана із цією таблицею через 1-ше поле "sub" (рис.3.14), у якому записані коди частин мови, а також коди деяких інших класів слів (? – невизначена граматична характеристика; Ф – неукраїнський алфавіт; Э – абревіатура); 2-ге поле "kont" подає значення кодового символу. Ця таблиця має дуже малий обсяг (22 рядки, 2 поля) і укладається вручну, функціонально вона призначена для формування випадного списку в інтерфейсі частотного словника, за яким користувач здійснює пошук морфем та морфемних структур слів за частиною мови.

На останньому етапі укладання електронних ЧМС створюється людино-машинний інтерфейс [27] у вигляді веб-додатка ASP.Net.:

[27] Робота інтерфейсу ЧМС буде розглянута у наступному параграфі.

розробляються запити SQL до таблиць лексикографічної БД з урахуванням прогностичних запитів користувача.

Кількісна оцінка якості роботи автоматичного морфемного аналізу. Якість роботи автоматичного морфемного аналізу в КУМ можна оцінити, якщо визначити скільки початкових форм (лем) текстової вибірки мають визначену морфемну структуру. Ця інформація систематизована у полі "morfem" таблиці "freq" (додаток 12), у якому за даними МБД АСМСА (розділ 2) через зіставлення початкової форми (леми) таблиці "freq" із початковою формою МБД записується модель морфемної структури слова, наприклад для леми *ревти* RF (табл.3.4). Як зазначалось, у полі "morfem" модель морфемної структури слова визначена не для всіх слів реєстру. Відсутність моделі морфемної структури слова зумовлена тим, що реєстр лем кожної текстової вибірки автоматично згенерований за слововживаннями тексту, а автоматичний морфемний аналіз проводиться лише для тих слів, які є у морфемній базі даних АСМСА, укладеної за нормативними словниками української мови. У лексичному реєстрі текстових вибірок є слова, які не були зафіксовані нормативними словниками і не ввійшли до МБД АСМСА, тому частина лексики реєстру залишається необробленою. Після доповнення МБД АСМСА новою лексикою було проведено оновлення даних у БД морфемних частотних словників. За останніми даними (листопад 2019 р.), представленими у табл. 3.9, кількість обробленої/необробленої лексики, за чотирма текстовими вибірками, може бути різною, що визначає різну ефективність роботи системи автоматичного морфемного аналізу в КУМ: за лексичними реєстрами Л. Костенко та Лесі Українки відсоток обробленої лексики становить ≈ 94 – 95%, а для лексичних реєстрів Т. Шевченка та В. Стуса ≈ 82% (попередній показник (за даними 2017 р.) обробленої лексики Т. Шевченка (опис таблиці "freq" становив 77,4%).

Таблиця 3.9. Кількісна характеристика обробленої/необробленої лексики у чотирьох текстових вибірках

Текстова вибірка	**Кількість одиниць реєстру початкових форм (лем) текстової вибірки**	**Лексика із визначеною морфемною структурою**		**Лексика із невизначеною морфемною структурою**	
		кількість лем	**%**	**кількість лем**	**%**
Т.Шевченко	7526	6191	82,26	1335	17,74
Л.Костенко	6347	5956	93,84	391	6,16
Леся Українка	5146	4908	95,38	238	4,62
В.Стус	9733	8029	82,49	1704	17,51

Як правило, необроблені слова із значенням абсолютної частоти (f) ≥ 10) є програмними помилками, що вимагає виправлення і внесення до бази даних АСМСА, але більшу частину цієї вибірки складають низькочастотні стилістично марковані слова. Наприклад, для текстової вибірки Т. Шевченка серед необробленої лексики: f 1 – 1024 слова; f 2 – 182 слова; f 3 – 79 слів. Реєстр таких слів (додаток 13) є надзвичайно важливим матеріалом для

редагування МБД АСМС, а також для вивчення особливостей стилю або ідіостилю і потребує обов'язкового та глибокого лінгвістичного дослідження. Таким чином, методика укладання ЧМС за текстами Корпусу української мови дозволяє автоматично генерувати лексичні реєстри, які формують гіпотетичну вибірку стилістично маркованих слів.

Даталогічна модель БД морфемних частотних словників апробована на 20 текстових вибірках і може використовуватися як еталонна в задачах комп'ютерної статистичної лексикографії різних мов, що визначає пріоритетним методологічний підхід обчислення статистичних характеристик морфемних одиниць у тексті на базі морфемної будови початкових форм текстових слововживань. Цей підхід, визначений інфологічною концепцією, базується на онтологічному принципі організації морфемної структури слів у флективних мовах. Морфемна структура основи слова, що відображає експліковану семантику слова, при словозміні залишається відносно стабільною: кількість морфем не змінюється, може виникати тільки словозмінна аломорфія, але вона враховується при лематизації і, за потреби, її можна автоматично приписати морфемам як потенційну ознаку.

3.2.3. Електронні частотні морфемні словники: параметри пошуку та класифікаційні можливості

Українська традиційна морфемна лексикографія має у своєму доробку лише два паперові морфемні словники, які систематизують статистичні морфемні дані: «Словник афіксальних морфем української мови» [САМУК 1998], «Кореневий гніздовий словник української мови» [Карпіловська 2002], укладені за даними Морфемно-словотвірного фонду української мови. Ці словники подають частотні характеристики (абсолютну та / або відносну частоту) структурних морфемних одиниць (різних морфемосполук), слів із визначеними морфемосполуками, кореневих морфем за статистичними даними генерального реєстру морфемно-словотвірного фонду (кількісна реалізація морфемних одиниць у системі мови) та «Частотного словника сучасної української художньої прози» [ЧССУХП 1981], з якого в 1981 р. розпочався розвиток української статистичної лексикографії.

Електронні частотні морфемні словники в Корпусі української мови – це єдина в сучасній українській комп'ютерній морфемній лексикографії текстоорієнтована науково-дослідна інформаційна система лінгвістичної морфемної галузі знань, оснащена пошуково-класифікаційними програмними аналізаторами, що забезпечують: ефективне та оперативне проведення морфемного аналізу на великих лексичних масивах текстів; отримання точних статистичних даних про морфемні одиниці в українськомовному тексті.

У Корпусі української мови в рубриці «Частотні словники» [ЧСКУМ 2018] автоматично укладені частотні морфемні словники за такими текстовими вибірками:

1. Частотні словники за стилістичними підкорпусами:

- поетичних текстів (станом на 2018 рік);
- наукових текстів (станом на 2018 рік);
- законодавчих текстів (станом на 2018 рік);
- художньої прози (станом на 2018 рік);
- художньої прози (станом на 2012 рік)
- публіцистики (станом на 2018 рік);
- публіцистики (станом на 2012 рік)
- медичних текстів (ендокринологія);
- фольклорних текстів (станом на 2014 рік);

2. Частотні словники за текстовими вибірками окремих авторів:

- збірки Лесі Українки ("На крилах пісень");
- книжки "Вибране" Ліни Костенко;
- збірки Василя Стуса "Палімпсести";
- збірки Василя Стуса "Круговерть";
- збірки Василя Стуса "Веселий цвинтар";
- поезії Тараса Шевченка (за виданням Твори : у 5 т. – Київ, 1970.)
- роману Василя Шкляра "Чорний ворон";
- романів Марії Матіос "Солодка Даруся", "Апокаліпсис";
- прозових творів Сергія Жадана;
- поетичних творів Сергія Жадана.

Кожен частотний словник – інтегральна лексикографічна система, що об’єднує 6 типів словників (див. § 3.1), серед яких два типи частотних морфемних словників: Частотний словник кореневих та афіксальних морфем; Частотний словник морфемних структур слів. Макроструктура кожного ЧМС побудована на основі інтегральної лексикографічної моделі, побудованої за зональним принципом структури. Кожен словник структурується на три функціональні зони-таблиці.

ЗОНА 1. Реєстр морфемних одиниць (морфструктур, коренів, афіксів) із такими статистичними характеристиками;

в) абсолютна частота вживання морфеми, морфструктури у вибірці текстів;

г) середня частота морфеми, морфструктури у вибірці текстів.

Кожен рядок таблиці першої зони представляє словникову статтю морфемної одиниці реєстру.

ЗОНА 2. Реєстр слів (лексем), у яких реалізовується, обрана користувачем у зоні 1, морфструктура / морфема, з такими характеристиками до кожного слова:

а) частина мови;

б) кількість текстів (творів), у яких вживається слово;
в) абсолютна частота;
г) середня частота;
ґ) середньоквадратичне відхилення;
д) коефіцієнт стабільності слова.

Кожен рядок таблиці другої зони представляє словникову статтю реєстрового слова.

ЗОНА 3. Конкорданс до кожного слова реєстру другої зони, у якому подано речення та твір-джерело, із якого взято речення. Словникова стаття третьої зони: реєстр речень, у яких вживається слово.

Даталогічна зовнішня модель інтегральної лексикографічної системи морфемних словників побудована з урахуванням інтерактивного характеру електронних словників: користувач за обраними опціями в інтерактивному режимі автоматично будує потрібні словники. В інтерфейсі частотних словників, створеному у вигляді веб-додатка ASP.Net, навігація між двома словниками (морфем та морфструктур) та 3-ма зонами у межах кожного словника здійснюється через гіперпокликання між одиницями електронних словників, що відображається у 3-ох інтерфейс-екранах (рис. 3.15.):

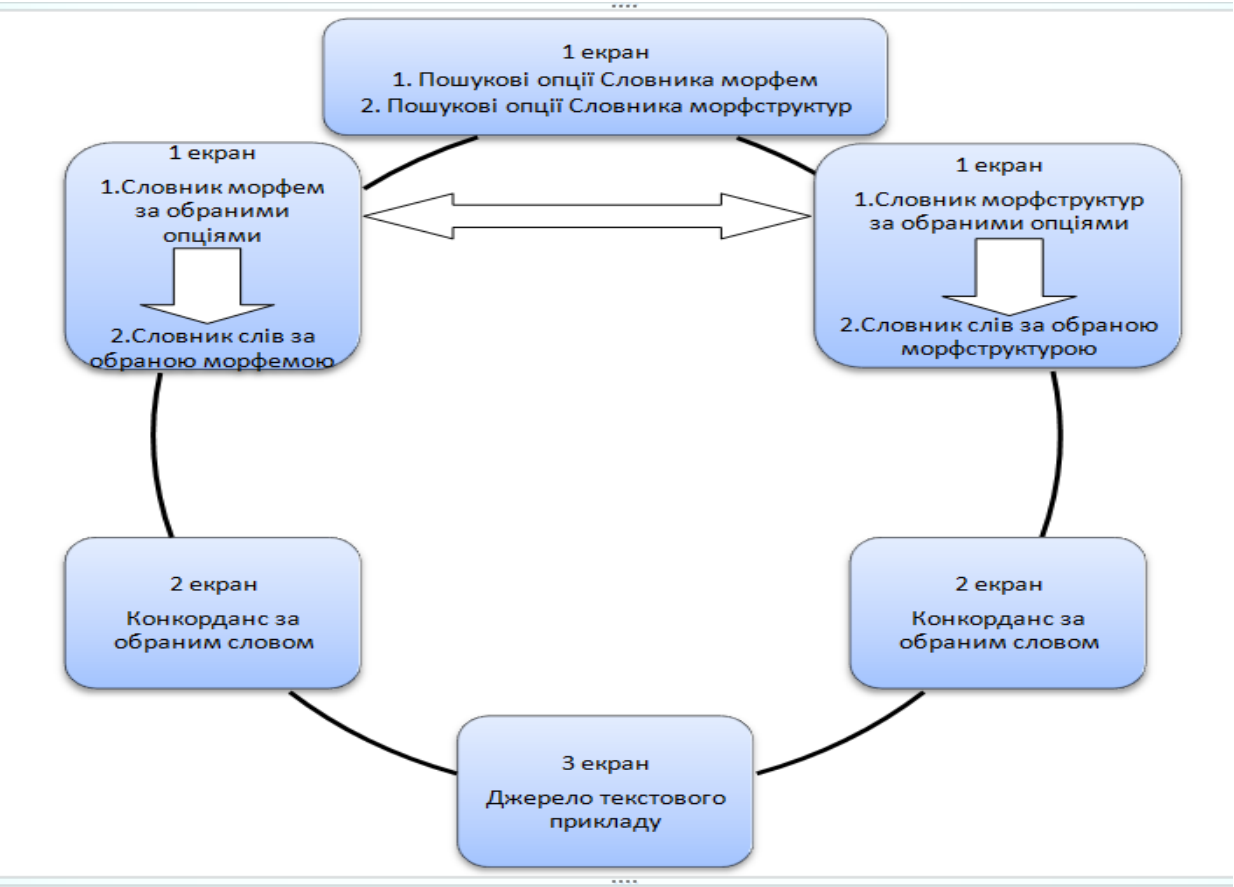


Рис. 3.15. Структура інтерфейсу лексикографічної системи ЧМС

Електронний ЧС морфем, укладений за текстами Т.Шевченка

ЗОНА 1 = 1-ий екран

Інтерфейс першого екрана електронного словника мови Тараса Шевченка, який подано в додатку 18, демонструє інтерактивний характер словника: користувач обирає потрібні йому класифікаційні опції і автоматично конструює морфемний частотний словник. Частотний список морфем укладається автоматично на лексичній вибірці, обмеженій двома перехресними запитами за вибором користувача: 1) функціональний тип морфеми, для якої користувач хоче побудувати частотний словник (рис.3.16): префікс, корінь, суфікс, інтерфікс (для флексій частотні словники не будуються); 2) частиною мови, яка визначає морфологічне поле лексичної вибірки тексту: морфемний словник укладається за всіма словами тексту або за словами окремої частини мови, визначеної користувачем у випадному списку опції "Частина мови" (рис.3.16).

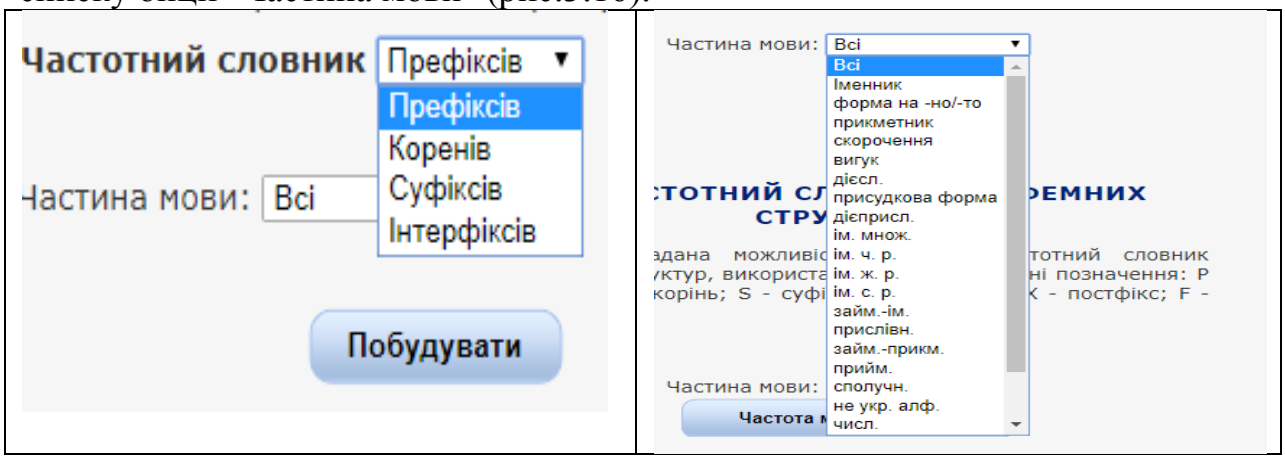


Рис. 3.16. Фрагменти інтерфейсу ЧМС Т.Шевченка: вибір функціонального типу морфеми та граматичної характеристики слів лексичної вибірки

За вибраними пошуковими параметрами, при активації опції "Побудувати", автоматично конструюється частотний словник морфем (зона 1). Демонстраційна версія (рис.3.17.) представляє частотний словник коренів за пошуковими опціями: 1) коренева морфема; 2) морфологічне поле лексичної вибірки – іменник.

Морфемно-частотний словник Коренів
всього записів: 1509

Морфема	Структури	Слова	Абсолютна частота	Середня частота
бу	Структури	Слова	531	8,70
пі	Структури	Слова	380	6,23
ста	Структури	Слова	340	5,57
він	Структури	Слова	335	5,49
бог	Структури	Слова	306	5,02
люд	Структури	Слова	306	5,02
світ	Структури	Слова	287	4,70
да	Структури	Слова	280	4,59
каз	Структури	Слова	280	4,59
див	Структури	Слова	273	4,48

1 2 3 4 5 6 7 8 9 10 ... >>

Рис. 3.17. Фрагмент частотного словника коренів, укладеного за іменниковою лексичною вибіркою

Словник коренів, які реалізовані в іменниках поетичних текстів Т. Шевченка, складає 1509 одиниць. Словник конструюється у вигляді таблиці, яка систематизує таку інформацію:

- поле 1 – Морфема: реєстр кореневих морфем, який при активації опції "Морфема" будується в алфавітному прямому або зворотному порядку;
- поле 2 – Структури: зв'язок частотного словника морфем із частотним словником морфемних структур слів, наприклад, при активації опції "Структури" у рядку кореня -*люд*- на окремому екрані будується список моделей морфемних структур слів-іменників, у яких вживається корінь -*люд*-, із інформацією про абсолютну частоту вживання цих структур у поетичних текстах Т. Шевченка (рис. 3.18.).

Всього записів: 5

Структура	Абсолютна частота
PRF	3
PRSF	2
RF	282
RIRSF	4
RSF	1

Рис. 3.18. Список моделей морфемних структур іменників, у яких вживається корінь -*люд*-

- поле 3 – Слова: зв'язок 1-ї зони із 2-ою зоною частотного словника морфем, яка показує реалізацію вибраної у першому полі морфеми в словах поетичних текстів Т. Шевченка (ця зона буде описана нижче);
- поле 4 – Абсолютна частота: абсолютна частота вживання кожного кореня в поетичних текстах Т. Шевченка без обмеження морфологічного поля частини мови лексичної вибірки;
- поле 5 – Середня частота: середня частота вживання кожного кореня у поетичних текстах Т.Шевченка без обмеження морфологічного поля частини мови лексичної вибірки. Рангові списки за абсолютною та середньою частотою будуються за спадом та ростом частот при активації кнопок "Абсолютна частота" та "Середня частота".

Морфемно-частотний словник коренів у межах іменникової лексики формується на основі програмних запитів до полів таблиць БД частотних словників (див. §.3.2.2):

1) поля "mname" таблиці "morfemstat", у якому за фільтром R формується вибірка кореневих морфем із абсолютною частотою вживання у текстах Т. Шевченка (додаток 19);

2) поля "cls" таблиці "freq", за фільтрами якого (коди іменника Й, К, Л) формується лексична вибірка іменників за текстами Т. Шевченка;

3) поля "mname" таблиці "morfems", у якому за фільтром R формується вибірка всіх кореневих морфем за текстами Т. Шевченка, до яких у полі "wrdFK" подано порядковий номер слова в БД частотних словників (додаток 20);

5) поля "Код" лексичної вибірки іменників, сформованої за частиномовними фільтрами іменника поля "cls" таблиці "freq". Через зіставлення кодів початкових форм слів у полях "wrdFK" таблиці "morfems" та "Код" таблиці "freq" автоматично формується реєстр кореневих морфем (1-ше поле інтерфейсу частотного морфемного словника: рис.3.17) за лексичною вибіркою іменників із збереженням інформації, яка записується у 4-ох наступних полях інтерфейсу частотного словника.

ЗОНА 2 = 1-ий екран

За умови активації опції "Слова", у рядку обраної морфеми, користувач здійснює зв'язок 1-ї зони частотного словника з 2-ою зоною – реалізація морфеми у словах текстової вибірки, що відкривається в окремій таблиці 1-го екрана інтерфейсу словника (рис. 3.19).

Морфемно-частотний словник Коренів
всього записів: 1509

Морфема	Структури	Слово	Абсолютна частота	Середня частота
бу	Структури	Слово	531	8,70
пі	Структури	Слово	380	6,23
ста	Структури	Слово	340	5,57
він	Структури	Слово	335	5,49
бог	Структури	Слово	306	5,02
люд	Структури	Слово	306	5,02
світ	Структури	Слово	287	4,70
да	Структури	Слово	280	4,59
каз	Структури	Слово	280	4,59
див	Структури	Слово	273	4,48

1 2 3 4 5 6 7 8 9 10 ... >>

Частотний словник по морфемі:R:люд, частина мови:Іменник Всього записів: 7

Слово	Частина мови	Абсолютна частота	Джерело	Середня частота	Середньоквадратичне відхилення	Коефіцієнт стабільності
люд	ім. ч. р.	12	8	0,20	0,57	2,88
люди	ім. множ.	270	125	4,43	2,68	0,61
людина	ім. ж. р.	1	1	0,02	0,13	7,75
людоїд	ім. ч. р.	3	3	0,05	0,22	4,40
людомор	ім. ч. р.	1	1	0,02	0,13	7,75
недолюд	ім. ч. р.	3	3	0,05	0,22	4,40
недолюдок	ім. ч. р.	2	2	0,03	0,18	5,43

Рис. 3.19. Лексична реалізація обраного кореня

У другій зоні словника подано лексичну реалізацію конкретного кореня, вибраного в таблиці першої зони: фрагмент демонструє реалізацію високочастотного кореня *–люд-* (з абсолютною частотою вживання 306 в усіх словах текстової вибірки, 1-ша таблиця) в іменниковій лексиці. Цей корінь реалізується в 7-ми словах-іменниках (2-га таблиця) із різною продуктивністю. Найчастотнішим є слово *люди*, яке реалізується у поетичному мовленні Т. Шевченка у 270 текстових слововживаннях.

Таблиця лексичної реалізації морфеми систематизує інформацію за такими полями:

- поле 1 – Слово: список початкових форм слів, у яких реалізовується вибрана морфема: корінь *-люд-*, реалізований у словах *люди, люд, людоїд, недолюд, недолюдок, людина, людомор*; активація слова (вибір мишею) зв'язує зону 2 із зоною 3 словника, що подає конкорданс до вибраного слова (додаток 21);

- поле 2 – Частина мови: морфологічна характеристика слова: *люди* – іменник, множина;
- поле 3 – Абсолютна частота: абсолютна частота іменника *люди* в поетичних текстах Т. Шевченка – 270; активація опції "Абсолютна частота" дозволяє автоматично будувати ранговий список за спадом та ростом значення абсолютної частоти;
- поле 4 – Джерело: активація цифри у рядку вибраного слова зв'язує зону 2 першого екрана із зоною 3(2) третього екрана, у якому подається список творів-джерел (додаток 22), із яких взято текстові фрагменти (речення) конкорданса до обраного слова;
- поле 5 – Середня частота: середня частота іменника *люди* – 4,42622950819672;
- поле 6 – Середнє квадратичне відхилення: середнє квадратичне відхилення іменника *люди* – 2,68242009884176;
- поле 7 – Коефіцієнт стабільності: коефіцієнт стабільності іменника *люди* – 0,606028244553139.

Таблиця лексичної реалізації вибраної морфеми автоматично будується і працює в інтерактивному режимі на основі програмних запитів до полів таблиці "freq" (додаток 12) БД частотних словників (див. §.3.2.2).

ЗОНА 3 = 2-ий екран + 3-ій екран

У третій зоні подаються контексти вживання вибраного у реєстрі 2-гої зони слова. Зона 3 відображається в інтерфейсі частотного словника 2-ма екранами.

Екран 2: конкорданс до вибраного слова (додаток 21), який відкривається при активації слова в першій колонці другої таблиці (рис. 3.19), наприклад:

Конкорданс до слова: **люди** ім. множ.

Морфемна структура: **RF /люд/и/**

Контекст	Джерело
Пошли ж ти їй долю — вона молоденька , Бо люде чужії її засміють .	>>
На чужині не ті люде — Тяжко з ними жити !	>>

Екран 3(1): твір-джерело, з якого взято речення-цитату у конкордансі. Навігація до цієї інформації здійснюється через активізацію позначки – >> у таблиці 2-го екрана. 3-ій екран – це сторінка твору-джерела в рубриці "Статистика текстів" (рис. 3.20), наприклад, при активації джерела речення *Пошли ж ти їй долю — вона молоденька , Бо люде чужії її засміють.* користувач потрапляє на екран 3(1).

Екран 3(2): список усіх творів-джерел (додаток 22), із яких взято речення конкорданса до обраного в другій таблиці слова (поле 4 на рис.3.19). Екрани 3.(1) та 3.(2) працюють в інтерактивному режимі на основі програмних запитів до полів таблиці "temp_freq" (додаток 11) БД частотних словників (див. §.3.2.2) та БД модуля "Текст" Корпусу української мови.

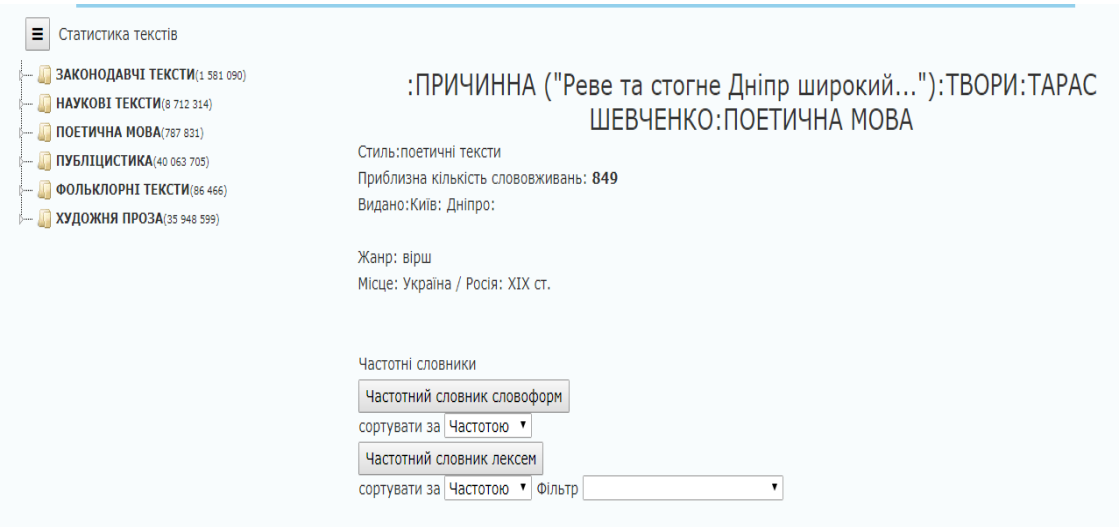


Рис. 3.20. Екран твору-джерала текстового фрагмента конкордансу

Електронний ЧС морфемних структур слів, укладений за текстами Т. Шевченка

Як і попередній словник, Частотний словник морфемних структур[28] слів поділено на 3 зони, а інтерактивна навігація по інтерфейсу на 3 екрани (рис.3.15.).

ЗОНА 1

У зоні 1 ЧС морфструктур користувач вибирає лише морфологічне поле лексичної вибірки слів за випадним списком опції "Частина мови" (додаток 23), у демонстраційному фрагменті – іменник. При активації опції "Частота морфструктур" автоматично укладається таблиця із трьох колонок (рис. 3.21.):

Всього записів: 39

Структура	Абсолютна частота	Середня частота
RF	16323	267,59
RSF	7482	122,66
R	6605	108,28
PRSF	3804	62,36
PRF	1631	26,74
RSSF	898	14,72
RS	604	9,90
PRS	499	8,18
RSS	359	5,89

Частотний словник по морфструктурі:**PRSF**, частина мови:Іменник Всього записів: 95 Всього записів: 95 Всього записів: 95

Слово	Частина мови	Абсолютна частота	Джерело	Середня частота	Середньоквадратичне відхилення	Коефіцієнт стабільності
пожар	ім. ч. р.	20	11	0,33	0,74	2,26
невольник	ім. ч. р.	15	11	0,25	0,64	2,62
пророк	ім. ч. р.	12	7	0,20	0,70	3,54
порада	ім. ж. р.	11	10	0,18	0,42	2,36

Рис. 3.21. Зони 1 та 2 ЧС морфемних структур іменників

Перша таблиця систематизує таку лінгвістичну інформацію:

- поле 1 – Структура: список 39 морфемних структур іменникової лексики;

[28] морфструктур

- поле 2 – Абсолютна частота: абсолютна частота морфструктури в усіх слововживаннях тексту без обмеження частиномовної характеристики слова;
- поле 3 – Середня частота: середня частота вживання морфструктури в усіх слововживаннях тексту без обмеження частиномовної характеристики слова.

Формування списку морфструктур можливе за спадом або за зростанням абсолютних частот. Наприклад, морфструктура PRSF є продуктивною і реалізується в 3804 слововживаннях тексту.

Частотний словник морфструктур за лексичною вибіркою іменникової лексики формується на основі запитів SQL до полів таблиці "freq" (додаток 12) БД частотних словників (див. §.3.2.2).

ЗОНА 2

За вибором морфструктури у 1-ій таблиці автоматично будується 2-га таблиця (рис. 3.21), у якій подано лексичну реалізацію морфструктури, вибраної в першій зоні, наприклад, морфструктура PRSF реалізована в 95 іменниках текстів Т. Шевченка. До кожного іменника першого поля таблиці у 5-ти наступних полях подано інформацію за схемою, аналогічною до таблиці 2 Частотного словника морфем (рис. 3.19). Навігація по інтерфейсу словника до зони 3 здійснюється за полем "Слово" та полем "Джерело".

ЗОНА 3

У третій зоні (екрани 2 і 3) ЧС морфструктур подаються: конкорданс до вибраного слова; джерело, із якого взято речення конкорданса; повний список творів-джерел, із яких взято контексти.

Частотний словник морфемних структур слів пов’язаний через пошукові опції із Частотним словником морфем. У заголовку до другої таблиці голубим кольором підсвічується інтерактивна опція PRSF (рис. 3.21.). При активації кожного символа моделі морфемної структури слова в окремому екрані інтерфейсу будується окрема таблиця (рис. 3.22.), яка систематизує список морфем, що можуть реалізовуватися в цій морфемній позиції, та подає статистичну характеристику абсолютної частоти морфеми в лексичній вибірці слів із морфемною структурою PRSF. Таким чином, користувач може системно аналізувати всі морфеми, що реалізовуються в моделі PRSF.

Морфемно-частотний
словник P позиція 0
всього записів: 23

Морфема	Абсолютна частота
по	33
за	9
о	9
при	8
пере	5
не	4
на	3
недо	3
під	3
без	2

1 2 3

Морфемно-частотний
словник R позиція 0
всього записів: 82

Морфема	Абсолютна частота
жар	3
каз	3
від	2
гиб	2
да	2
двір	2
пас	2
рад	2
сід	2
тик	2

1 2 3 4 5 6 7 8 9

Морфемно-частотний
словник S позиція 0
всього записів: 24

Морфема	Абсолютна частота
0	17
ок	12
к	4
н	3
ј	2
ищ	2
ник	2
оньк	2
в	1
ель	1

1 2 3

Морфемно-частотний
словник F позиція 0
всього записів: 5

Морфема	Абсолютна частота
а	24
я	5
е	3
и	3
о	1

Рис. 3.22. Списки морфем, що маніфестують символи моделі PRSF

Комп'ютерна лексикографічна система частотних морфемних словників у Корпусі української мови розкриває нові можливості в дослідженні морфемної організації українського слова. Автоматична систематизація різних морфемних одиниць за інтерактивними класифікаційними опціями частотних словників дозволяє аналізувати залучення морфем та морфемних моделей слів у творення нових слів, досліджувати морфемну довжину і глибину слів в українських текстах різних стилів, а також морфотактику різних типів морфем. Зв'язок частотних морфемних словників із конкордансом дозволяє проаналізувати різноаспектне функціонування морфем і морфструктур у текстах.

Автоматичне обчислення та систематизація статистичних характеристик морфемних та лексичних одиниць у частотних морфемних словниках, укладених за текстовими вибірками Корпусу української мови, відкривають велик перспективи для глибоких стилеметричних розвідок у дослідженні різних функціональних стилів та ідіостилів. Зокрема дослідження статистичних параметрів морфемного рівня організації поетичних текстів [Zuban 2019a], [Зубань 2014], [Зубань 2014a], [Зубань 2016] показують, що кількісні та статистичні характеристики морфемних структур слів, які формують відносно невеликий інвентар одиниць, виявляють закономірності будови тексту ідіостилю поета.

Досвід комп'ютерного лексикографічного моделювання морфемної системи української мови свідчить, що з метою вилучення з тексту реляційно-функціональних характеристик морфемних одиниць не обов'язково проводити морфемну або словотвірну анотацію текстів. Використана в комп'ютерному лексикографічному моделюванні методика автоматичного морфемного аналізу початкових форм не знижує ефективності та оперативності класифікаційно-пошукових опцій текстоорієнтованої лексикографічної системи, а також не применшує значущості результатів лінгвістичного дослідження на отриманому матеріалі, а навпаки, за рахунок систематизації та різноаспектної класифікації морфемної інформації підвищує експланаторність лінгвістичного дослідження.

Вивчення організації тексту на морфемному рівні за допомогою морфемного аналізу початкових форм слів, а не слововживань, виправданий практикою створення частотних морфемних словників у Корпусі української мови. Використання методики автоматичного морфемного аналізу лексики тексту за їх початковими формами в Корпусі української мови демонструє ефективність і оптимальність цієї методики: ≈ 200 тис. одиниць морфемної бази даних АСМСА дозволяють отримати інформацію про морфемну організацію мільйонів текстових слововживань з ілюстрацією контекстів їх вживання. Недолік цієї методики – неможливість автоматично зняти омонімію слів однієї частини мови за умови різної морфемної сегментації омографів: *вида-ти* (надрукувати), *ви-да-ти* (дати), *вид-а-ти* (бачити). У таких випадках морфемна сегментація редагується вручну.

Розбудова АСМСА (див. Розділ 2) на нинішньому етапі відкриває нові перспективи морфемного аналізу на матеріалі лексики текстів Корпусу української мови, зокрема поглиблення морфемного аналізу з урахуванням семантики морфем: групування спільнокореневої та спільноафіксальної лексики з урахуванням омонімії морфем та аломорфії. Методика комп'ютерного моделювання частотних словників за допомогою АСМСА узагальнює теоретичні та прикладні ідеї сучасного мовознавства, що робить інтерактивну лексикографічну систему ефективним і раціональним інструментом лінгвістичних досліджень. Зовнішня даталогічна лексикографічна модель електронних частотних морфемних словників, описана в цьому параграфі, може бути реалізована на всіх текстових вибірках Корпусу української мови на замовлення користувачів.

Р О З Д І Л 4

ЕЛЕКТРОНІ ЧАСТОТНІ МОРФЕМНІ СЛОВНИКИ – КОМП'ЮТЕРНИЙ ІНСТРУМЕНТ СТИЛЕМЕТРИЧНИХ ДОСЛІДЖЕНЬ

4.1. Методичні принципи організації статистичного аналізу в морфемній стилеметрії

Використання статистичних методів у мовознавчих дослідженнях набуває все більшого визнання: якщо раніше кількісні показники використовувалися лише для підтвердження висновків дослідження, проведеного традиційними методами якісного лінгвістичного аналізу, то сьогодні можна говорити про усталену тенденцію використання статистичних характеристик у встановленні закономірностей організації різних підсистем мови та мовлення (тексту), тому що: «... статистичні методи і характеристики допомагають глибше проникнути в закони мови і мовлення і можуть бути основою або відправною точкою для встановлення таких закономірностей, яких без цих методів не вдалося б побачити» [ВСС 1974: 3].

Особливого значення статистичні методи набувають у стилістиці, тому що відсутність кількісних показників у дослідженні функціональних та авторських стилів позбавляє наукове дослідження об'єктивності і доказовості зроблених висновків. Розвиток статистичної стилістики ознаменований в українському мовознавстві виходом у світ монографії «Статистичні параметри стилів» [СПС 1967], написаної колективом відділу структурно-математичної лінгвістики Інституту мовознавства ім. О.О.Потебні НАН України. У цій монографії В. Перебийніс обґрунтовує можливість використання статистичних методів у стилістичному дослідженні: «Можливість використання статистичних методів у стилістиці ґрунтується на тому, що всякий матеріал мовлення (тобто текст) є результатом добору певних одиниць із загальнонародної мови. Добір цей залежить не лише від теми висловлювання, але й від форми її викладу (поетичний чи драматичний твір, художня чи наукова проза, науково-популярний чи науковий виклад і т.д.), від законів і канонів стилю чи жанру, від особистих уподобань автора і, нарешті, від тих законів, за якими будується мовлення, а також від законів мови. Мова диктує свої закони кожному, хто нею користується, і ступінь підсвідомого чи свідомого засвоєння цих законів мовцем позначається на якості добору і розташуванні мовного матеріалу в мовленні[29]» [Перебийніс 1967: 23].

Фундаментальні статистичні дослідження колективу відділу структурно-математичної лінгвістики представлено низкою відомих праць: збірник «Вопросы статистической стилистики» [ВСС 1974], «Частотний словник сучасної української художньої прози» [ЧССУХП 1981], «Частотні

[29] Підкреслено автором монографії.

словники та їх використання» [Перебийніс 1985] та ін.. Лінгвістичні розвідки науковців цього відділу: «Система афіксального словотворення сучасної української мови» [Клименко 1973], «Морфемна структура слова» [МСС 1979], «Словник афіксальних морфем української мови» [САМУК 1998], «Основи морфеміки сучасної української мови» [Клименко 1998], «Суфіксальна підсистема сучасної української літературної мови: будова та реалізація» [Карпіловська 1999] – демонструють застосування статистичного аналізу в дослідженні морфемної системи української мови.

Перші стилістичні дослідження морфемної будови українських слів були проведені відомими українськими лінгвістами І. Білодідом [Білодід 1954] та Д. Баранником [Баранник 1958], [Баранник 1961] в аспекті стилістичного значення способів словотвору української мови. Дослідження статистичних параметрів основних стилів української мови на морфемному рівні організації тексту було започатковано ще у монографії «Статистичні параметри стилів» [СПС 1967], де аналізується частота кінцевих афіксів як статистичний параметр стилю, частота префіксів (статистичне вивчення 19 префіксів) і префіксальних словоформ, проте системний аналіз морфемної структури тексту не був предметом спеціального вивчення в аспекті параметризації стилю чи ідіостилю окремого автора. Причиною цього є "ручний" морфемний та статистичний аналіз слів у тексті – така робота навіть на текстах одного автора, на базі найменшої репрезентативної вибірки обсягом 1000 словоформ, є дуже трудомісткою. Системне вивчення морфемної структури слів в аспекті параметризації ідіостилю стало можливим з розвитком корпусної лінгвістики й комп'ютерної статистичної морфемної лексикографії, зокрема такі дослідження проводяться на матеріалі частотних морфемних словників, укладених на матеріалі текстів Корпусу української мови [Зубань 2014], [Зубань 2014], [Zuban 2019a].

Об'єктом статистичного аналізу в пропонованому стилеметричному дослідженні є модель морфемної структури (ММС) слова. Яку лінгвістичну цінність у стилеметричному дослідженні мають моделі морфемних структур слів та їхні статистичні характеристики? В електронних ЧС, які виступають комп'ютерним інструментом проведення статистичного дослідження ідіостилів, в описі морфемної структури слова використано метод моделювання, принципи якого описано у 1-му розділі монографії. Модель морфемної структури слова (PPPRSF – *не-с-по-ви-т-ий*) розглядається як інваріантна символьна модель морфемної будови слова, що представляє експліковану морфемами (найменшими знаковими одиницями мови) семантику слова на найвищому рівні абстрагування. Звичайно, морфемна структура слова не відображає повністю структури семеми, яка, залежно від критеріїв компонентного аналізу та кількості встановлених сем у слові, може визначатися по-різному: семантика слова залежить від екстралінгвістичних чинників, які формують денотативний, десигнативний, конотативний компоненти лінгвістичного знака, що є індивідуальними в різних мовців, у різних мовних ситуаціях і в

різних авторських стилях. Проте незаперечним є той факт, що ускладнення морфемної структури слова (збільшення кількості морфів) зумовлює ускладнення його семантичної структури. Морфемні структури слів є типовими семантико-структурними одиницями мови, які відображають будову експлікованої морфемами семної структури слова, що розуміється всіма мовцями однаково і є відтворюваною в мовній діяльності та повторюваною в мовленні. Досліджувати семантичну структуру слова як знакової одиниці тексту, особливо поетичного, статистичними та кількісними методами дуже важко, тому що слово об'єднує кількісно не чітку множину сем. Зважаючи на це, застосування методу моделювання в лінгвістиці вимагає спрощення об'єкта дослідження – слова. Тому моделювання морфемної будови слова як метод математичної експлікації мовного знака (спрощення) дає можливість досліднику виявляти складність морфемної структури слова, а через неї складність семантичної структури слова як інваріантну модель семеми.

Застосовуючи метод моделювання морфемних структур, можна аналізувати різні кількісні параметри морфемної будови слова:

1) морфемну довжину слова: кількість морфем у слові;

2) морфемну глибину слова: розгортання морфемної будови слова відносно кореня різною кількістю афіксів та різними функціональними типами морфем;

3) валентність функціональних типів морфем у морфемній структурі слова.

Уперше поняття "глибина слова" в морфемному аналізі застосував В. Москович: «... глибину слова можна визначити як кількість морфем у слові безвідносно до того, як ці морфеми розташовані» [Москович 1967: 18], а потім у цій самій статті приходить до висновку, що: «Величина глибини слова вказує не тільки на кількість морфем у слові, а й на кількість тактів породження цього слова» [Москович 1967: 33]. Таким чином лінгвіст робить висновок про розмежування понять "глибина слова" і "довжина слова" в проекції на різні структури слова: глибина слова визначається за кількістю морфем у слові, а довжина слова – за кількістю складів у слові. Крім того, В. Москович зауважує, що глибина слова залежить від функціонального стилю: «У різних стилях мови слова максимальної глибини й довжини зустрічаються з різною ймовірністю, тому в типологічному порівнянні мов варто спиратися на дані, вилучені із текстів одного стилю» [Москович 1967: 33]. Інтерпретуючи гіпотезу В. Інгве [Ингве 1965] про глибину речення В. Москович зазначає, що «Одна з підстав для висловлення припущення про 7 ± 2 морфем як верхню межу глибини слова, виникає за аналогією до гіпотези Інгве, друга, більш загальна, виходить із уявлення про те, що будь-яке перекодування одиниць нижчого рівня мови в одиниці вищого рівня пов'язано з обмеженим обсягом оперативної пам'яті людини» [Москович 1967: 19]).

У морфемології терміни "морфемна довжина слова" та "морфемна глибина слова" спочатку вживалися як взаємозамінні, оскільки глибина визначалася через кількість морфем у слові, а отже, становила морфемну довжину. Пізніше Н. Клименко чітко розмежовує ці поняття: «морфемна довжина слова – кількість морфем у слові, незалежно від їхнього розташування відносно кореня» [Клименко 1998: 150]; «... можливий розгляд морфемних структур слів з погляду їхнього породження з елементарних одиниць, тобто в перспективі розгортання в них ланцюжків морфем (префіксальних та суфіксальних) щодо ядерного (головного) компонента слова – кореня. У цьому плані встановлюють глибину слова, ...» [Клименко 1998: 153]. У стилеметричному дослідженні морфеміки ідіостилів використовується трактування термінів "морфемна довжина слова" та "морфемна глибина слова" у визначенні Н. Клименко.

Метод моделювання морфемних структур слів, який використано в електронних ЧМС, уможливлює дослідження взаємозалежності між кількісно-структурними ознаками ММС та статистичними характеристиками морфемних моделей і слів, у яких визначаються ці моделі, тому ці аспекти в стилістичних дослідженнях набувають статусу стилеметричних ознак.

Стилеметричне дослідження морфемних структур слів ґрунтується на вимозі системності опису ідіостилю, який передбачає аналіз не лише специфічних індивідуальних засобів мовлення, а всього масиву лінгвістичних одиниць якогось окремого рівня структури тексту. «Для лінгвостиліста становлять інтерес не лише ті аспекти мовлення, що формуються в результаті індивідуальних особливостей людини, а насамперед ті, які мають закономірний характер» [Кожина 1966: 16].

Підхід системності опису в статистичній стилістиці відрізняє її від стилістичних досліджень у літературознавстві. У монографії «Частотні словники та їх використання» автори, коментуючи завдання стилістики, визначені Ш. Баллі у відомій праці «Французька стилістика» [Балли 1961], зазначають із цього приводу: «Якщо для літературознавчої стилістики мова художнього твору є засобом розкриття ідейно-художнього змісту твору, то мовознавча стилістика повинна досліджувати загальні закономірності функціонування мови в художньому мовленні. Не можна погодитися із тими дослідниками, які вважають, що об’єктом стилістичного аналізу повинні бути експресивні, емоційно забарвлені відтінки мови, просторічні, діалектні елементи, словом стилістично марковані засоби. Адже, по-перше, ці засоби складають дуже малий процент від загального арсеналу зображувальних засобів. По-друге, можна знайти великі уривки тексту зовсім без стилістично маркованих елементів, а, між тим, текст є стилістично маркованим. Крім того, аналізуючи функціонування деяких мовних елементів, роблячи при цьому вибіркове дослідження, тобто "висмикуючи" окремі приклади, дослідники іноді просто переказують зміст твору, базуючись при цьому на власній інтуїції» [Перебийніс 1985: 127].

Стилеметричне дослідження морфемних структур слів керується тенденцією об'єднання двох аспектів дослідження стилістично маркованого тексту:

1) вивчення морфемної системи стилю з урахуванням усіх морфемних структур слів, які використані автором у текстах, що визначає індивідуальний добір цих одиниць за статистикою їх вживання, а отже робить морфемну структуру кожного слова стилістично маркованою одиницею;

2) вивчення тих морфемних структур слів, які формують групу індивідуальної функціонально-стилістичної та емоційно-експресивної лексики.

Отже, дослідження статистичних параметрів стилю на морфемному рівні його організації покликане:

1) встановити стилерозрізнювальні кількісно-статистичні закономірності функціонування морфемних структур слів у тексті стилю, які визначити іншими методами неможливо;

2) на базі статистичних характеристик ММС та дії статистичних законів, висувати гіпотези-оцінки матеріалу дослідження щодо наявності у статистичних лексичних вибірках стилістично маркованих одиниць, які можуть мати індивідуальні функціонально-стилістичні, емоційно-експресивні стилерозрізнювальні ознаки, що встановлюються й аналізуються якісними лінгвістичними методами: «...оскільки кількісний та якісний аспекти людського мовлення певним чином корельовані та взаємопов'язані, кількісні оцінки можуть бути сигналами, які спрямовують увагу дослідника на приховані від простого спостереження якісні особливості індивідуального або функціонального стилю» [Пиотровский 1968: 8],

Обидві цілі передбачають проведення зіставного стилістичного експерименту, який виявляє наявність або відсутність певного лінгвістичного явища в порівнюваних лексичних чи тестових вибірках. За влучним визначенням М. Пещак: «Наявність-відсутність досліджуваних елементів у певному місці структури цілого – один із основних стилістичних факторів твору, який виокремлює і об'єднує його з іншими у відібраній групі творів одного автора, а також у сукупності однотипних творів різних авторів, що належать до різних стилів однієї або декількох мов» [ВСС 1974: 217].

Методика стилеметрії передбачає дослідження статистичних параметрів стилів двома способами: 1) порівняння різних стилів; 2) порівняння з якимось еталоном, що є незалежним від цих стилів. Еталонним у стилеметрії при дослідженні одного стилю може бути або стилістично нейтральний штучно створений текст – нульовий стиль, який відображає закони мови, а не мовлення; або система мови: «…статистичні показники мови (мовні константи, встановлені на протилежність мовленню)» [СПС 1967: 30].

У стилеметричному дослідженні морфемних структур слів використовуються обидва підходи: у дослідженні ідіостилю одного автора,

визначальним є другий методичний підхід – зіставлення зі статистичними характеристиками морфемної будови слів у системі мови, а в дослідженні стилерозрізнювальних характеристик поетичних текстів двох і більше авторів – перший підхід.

Керуючись постулатом про те, що «... і будова мови, і її функціонування в мовленні, і співвідношення мови і мислення, мови і суспільства підкоряються дії статистичних законів, ...» [Перебийніс 2002: 7] висуваємо гіпотезу про дію в морфемній структурі тексту ідіостилю статистичного закону Дьюї, який у лінгвістиці названо законом переваги. «Він полягає в тому, що і мова, і мовлення віддають перевагу невеликій кількості одиниць, які часто використовуються й становлять ядро будь-якої мови чи мовленнєвої підсистеми, тоді як переважна кількість одиниць є низькочастотними» [Перебийніс 2002: 7].

Інтерпретація закону переваги в стилістиці традиційно досліджується через дію закону Ципфа – Мандельброта й визначення статистичної структури лексики тексту, що встановлюється за співвідношенням певної кількості слів лексичного реєстру, згрупованих за рейтингом однакової / подібної абсолютної частоти в тексті, із сумою абсолютних частот вживання цих слів у тексті. Цей закон визначає такі закономірності формальної організації тексту:

1) високочастотні слова тексту покривають найбільшу частину тексту, але становлять найменшу частину лексичного реєстру і є стилістично нейтральними, проте частота перших 10 слів рангового списку має стилерозрізнювальну потужність;

2) «середньочастотним словам властива найбільша "реактивність" відносно стилю автора» [Перебийніс 1985: 139];

3) низькочастотні слова тексту покривають найменшу частину тексту, але становлять найбільшу частину лексичного реєстру, і тому складають "багатство" лексикону автора, вони формують стилістичний критерій.

Відсоткове співвідношення цих статистичних характеристик у кожному стилі чи ідіостилі істотно відрізняється, тому «статистична структура реального тексту розуміється як його кількісна організація, як його модель. Статистична структура тексту, як певна рівнодіюча, дає можливість віднести той чи інший текст до певного функціонального стилю, визначити автора, період написання тощо» [Перебийніс 1985: 130].

Вважається, що термін "статистична структура лексики" був уведений Р. Фрумкіною, хоча сама дослідниця зазначає, що: «Ще в сорокових роках XX ст. було достовірно встановлено, що лексика кожного досить довгого тексту, будь то художня або наукова література, має певну статистичну структуру. У загальних рисах це означає, що для кожного автора існує визначене чітке співвідношення у вживанні більш частотних, менш частотних і рідкісних слів, і залежно від того, яке це співвідношення, ми суб'єктивно відчуваємо словник автора як багатий, різноманітний, або бідний, одноманітний» [Фрумкина 1960: 78].

У сучасній статистичній стилістиці поняття статистичної структури на рівні лексики стилю називають різними термінами: "статистична структура лексики", "статистична структура тексту", "статистична структура твору", "статистична структура стилю" – ці терміна не диференціюються. Статистичні парметри морфемних структур слів ще не були предметом вивчення у стилеметричних дослідженнях, тому необхідно обґрунтувати і термінологічно визначити поняття статистичної структури тексту на морфемному рівні його організації.

«Вважають, що статистична структура твору відома, якщо для будь-якої можливої частоти слова відома ймовірність, з якою можна навздогад вибрати з даного тексту слово із заданою частотою. Виявити статистичну структуру тексту – значить встановити залежність, що існує в ньому між частотами слів та ймовірністю появи в тексті слів із даною частотою» [Перебийніс 1985: 78]. Якщо модель морфемної структури слова, яка виступає об'єктом статистичного аналізу, – це структурна одиниця, що має реалізацію в конкретному слові, а опосередковано через цю одиницю реалізовується в тексті, то висуваємо гіпотезу про можливість застосування поняття статистичної структури тексту у вивченні морфемного рівня його організації.

Як методично правильно дослідити статистичну структуру тексту на морфемному рівні текстової системи стилю чи ідіостилю? Статичне дослідження морфемних структур слів враховує послідовність статистичної реалізації моделі морфемної структури слова в тексті, що закладена в систематизації лінгвістичної інформації в електронних ЧМС: одна ММС слова експлікується у певній кількості слів (початкових форм), які утворюють множину одиниць у лексичному реєстрі, а кожне слово цієї множини може реалізовуватися у тексті певною кількістю слововживань. Наприклад, ММС RSSSF у лексичному реєстрі представлена 31 словом (початковою формою), а в тексті 77 слововживаннями.

Базовими в стилеметричному дослідженні виступають два параметри: 1) кількісно-структурний: морфемна довжина слова, встановлена за кількістю морфем у ММС слова; 2) статистичний: абсолютна частота (f) та відносна частота (p – %) ММС у текстовій вибірці та лексичному реєстрі стилю. Оперуючи поняттям відносної частоти ММС у лексичному реєстрі, визначаємо це поняття як питома вага або лексична продуктивність ММС: відсоток лексики, який моделюється за визначеною ММС слова. Досліджуючи відносну частоту ММС у тексті, визначаємо цю статистичну характеристику як індекс покриття тексту: відсоток тексту, який "покривають" слововживання із визначеною морфемною структурою (ММС).

Відповідно, морфемна статистична структура стилю з урахуванням дихотомії мова / мовлення визначається у двох статистичних вибірках: лексичному реєстрі (система мови стилю) та тексті (система мовлення стилю). Проте модель морфемної статистичної структури стилю демонструє

іншу закономірність організації тексту, ніж традиційна модель лексичної статистичної структури тексту.

Статистична структура тексту на морфемному рівні встановлює співвідношення між кількістю слів однієї морфструктури в лексичному реєстрі та сумою абсолютних частот слововживань із цією морфструктурою в тексті і виявляє залежність, яка існує в мовній системі стилю між кількістю слів з однаковою морфемною будовою та ймовірністю появи в тексті таких слів. Наприклад, ММС RSSSF (*пт-аш-еч-к-а)* в лексичному реєстрі представлена 31 словом (сума абсолютних частот слів (лем) однієї ММС у лексичному реєстрі), а в тексті ця ММС представлена 77 слововживаннями (сума абсолютних частот слововживань однієї ММС у тексті).

У лексичній статистичній структурі тексту у встановленні співвідношення між кількістю слів у реєстрі і сумою абсолютних частот цих слів у тексті початковою точкою виступає формальна ознака – варіанта абсолютної частоти або інтервал варіаційного розподілу частот – за якою відбувається групування слів за однаковим показником абсолютної частоти (варіанта x_i). Наприклад: з абсолютною частотою 15 у тексті автора знайдено 150 слововживань, які в лексичному реєстрі формують групу з 10 слів. Тому статистична структура лексики є абсолютно формально-кількісною ознакою стилю: «... статистичні особливості тексту, що характеризують індивідуальний стиль автора, можна розглядати, відволікаючись від змісту, – у статистичній структурі тексту. Статистичні особливості тексту є формальними властивостями його організації, за допомогою яких можна розкрити якісні особливості стилю» [Перебийніс 1985: 130].

Морфемна статистична структура стилю також виявляє залежність розподілу абсолютних частот у лексичному реєстрі та тексті, але це співвідношення базується на зіставленні в лексичному реєстрі та тексті суми абсолютних частот слів, згрупованих за однаковою ММС слова, а не за рейтингом абсолютної частоти слів у тексті. Тому морфемна статистична структура виявляє не тільки формально-кількісні властивості організації стилю, а й лінгвістичну інформацію про його організацію: за якими статистичними закономірностями слова з різною кількістю морфем, з різними функціональними типами морфем, з різною кількістю кореневих морфем (і інші ознаки ММС) реалізуються в лексичному реєстрі та тексті стилю. Крім того, визначення статистичної структури тексту на рівні морфемних структур демонструє вищий рівень узагальнення кількісної моделі тексту, ніж на рівні лексикону, тому що кількість моделей морфструктур у тексті набагато менша ніж кількість слів тексту.

Центральним поняттям стилеметричного дослідження морфемної системи ідіостилю в нашому дослідженні виступає м о д е л ь м о р ф е м н о ї с т а т и с т и ч н о ї с т р у к т у р и і д і о с т и л ю, що визначається як розподіл ММС слів на три статистичні групи (високочастотні, середньочастотні, низькочастотні) через зіставлення відносної частоти слів однієї ММС у двох рейтингових списках ММС, укладених за: 1) лексичним реєстром

(лексемами), що формує поняття – морфемна статистична структура лексикону ідіостилю; 2) текстом (слововживаннями), що формує поняття – морфемна статистична структура тексту ідіостилю.

На базі визначеного поняття "модель морфемної статистичної структури ідіостилю" в стилеметричному дослідженні ставиться завдання проаналізувати статистичну "поведінку" ММС слів у лексичному реєстрі та тексті ідіостилю, а саме: 1) співвідношення різних статичних груп ММС слів; 2) співвідношення статистичних характеристик ММС з кількісними характеристиками морфемної довжини слів.

4.2. Статистичне моделювання морфемної системи поетичного ідіостилю Т. Шевченка

Мова ідіолекту Т. Шевченка була об'єктом досліджень численних лінгвістичних та літературознавчих студій як в україністиці, так і за її межами. Серед праць лінгвостилістичного характеру цього циклу необхідно відзначити дослідження П. Тимошенка [Тимошенко 2013], у яких системно використані кількісні характеристики мовних одиниць в аналізі лексичних та морфологічних особливостей ідіостилю Т. Шевченка. Статистичні дослідження поетичного мовлення Т. Шевченка в аспекті вивчення морфемної структури слів в українському мовознавстві не проводилися.

Стилеметричне дослідження морфемної будови слів мови Т. Шевченка проводиться за електронними частотними морфемними словниками, укладеними на базі ≈ 61 тис. слововживань (за даними 2017 р.) поетичних текстів Т. Шевченка. Досліджувану вибірку в ЧМС складає реєстр початкових форм слів (лем) обсягом 5279 одиниць, автоматично згенерований за слововживаннями тексту. Автоматичний морфемний аналіз проводиться лише для тих слів, які є у морфемній базі даних АСМСА, тому до реєстру ЧМС потрапляють не всі 6889 початкових форм загального реєстру лексикону Т.Шевченка[30].

Морфемна статистична структура ідіостилю. За визначеними в § 4.1. методичними засадами, у стилеметричному дослідженні ідіостилю Т. Шевченка, ставляться завдання:

- сформувати ранговий список лексичної продуктивності ММС;

[30] У процесі автоматичного укладання ЧМС за текстами Т. Шевченка, автоматично формується список слів з невизначеною морфемною структурою обсягом 1560 одиниць (додаток 13). Причини формування такого списку описані в § 3.2.2. Як правило, слова цієї вибірки з порогом абсолютної частоти нижче 4-ох (f < 4: f 1 – 1024 слова; f 2 – 182 слова; f 3 – 79 слів) є лексикою, що формує функціонально-стилістичні, емоційно-експресивні та ономастичні особливості ідіостилю. Для традиційних стилістичних досліджень такі слова є надзвичайно важливим лінгвістичним матеріалом, а для комп'ютерної стилеметрії – це методичний прийом формування в моделі лексичної статистичної структури тексту низькочастотної групи лексики, яка може вважатися гіпотетичною вибіркою стилістично маркованих слів.

- сформувати ранговий список індексу покриття тексту ММС;
- проаналізувати морфемну статистичну структуру ідіостилю й перевірити гіпотезу про дію закону переваги в морфемній системі ідіостилю Т. Шевченка.

Ідіолект Т. Шевченка нараховує 65 моделей морфемних структур слів (за даними Морфемно-словотвірного фонду, українська морфеміка нараховує 418 моделей [Клименко 1998]). Із 54 однокореневих морфемних структур, властивих українській мові, у поетичному мовленні Т. Шевченка встановлено 41 морфструктуру, реалізовану в 5127 словах, що становить ≈ 97,12% лексикону мови Т. Шевченка (табл. 4.1).

Таблиця 4. 1. Ранговий список ММС простих слів у лексичному реєстрі ідіостилю Т. Шевченка

№	ММС		Приклад реалізації ММС у слові	Питома вага у лексиконі		Абсолютна частота (f) у тексті
				Абсолютна частота (f)	Відносна частота (p %)	
1.	RF	75,68 %	*мій*	1216	23,03	16323
2.	RSF		*плакати*	1206	22,85	7482
3.	PRSF		*співати*	1012	19,17	3804
4.	PRF		*ніхто*	293	5,55	1631
5.	RSSF		*дівчина*	268	5,08	898
6.	PRSFX	92,93 %	*подивитися*	213	4,03	771
7.	RSFX		*дивитися*	116	2,20	683
8.	PRSS		*вранці*	105	1,99	254
9.	PRSSF		*наймичка*	100	1,89	212
10.	R		*де*	91	1,70	6605
11.	PRS		*знову*	79	1,49	499
12.	RSS		*тяжко*	77	1,46	359
13.	RS		*добре*	60	1,14	604
14.	PPRSF		*заспівати*	57	1,08	186
15.	PR		*нехай*	37	0,70	491
16.	RSSSF		*пташечка*	31	0,59	77
17.	PRFX		*довестися*	28	0,53	150
18.	RFX		*дітися*	19	0,36	110
19.	RISF		*гукати*	15	0,28	37
20.	PRSSS		*звичайне*	14	0,27	30
21.	PPRF		*сповити*	14	0,27	27
22.	RX		*десь*	13	0,25	38
23.	PRSSSF		*заквітчаний*	13	0,25	20
24.	PPRSSF		*приспівувати*	10	0,19	22
25.	PPRSFX		*простягатися*	10	0,19	22
26.	RSSS		*нищечком*	6	0,11	27
27.	PPRSS		*взаперті*	6	0,11	8
28.	PPRS		*незабаром*	5	0,09	17
29.	RSSSSF		*дівчаточко*	3	0,06	12

30.	PRISF	*незнаємий*	3	0,06	5
31.	RSSSS	*спатоньки*	2	0,04	4
32.	RISSF	*преторіанин*	2	0,04	5
33.	RISFX	*сміятися*	1	0,02	57
34.	PPR	*невлад*	1	0,02	1
35.	RSSFX	*квітчатися*	1	0,02	1
36.	PPRISS	*анікогісінько*	1	0,02	1
37.	PPPRSF	*несповитий*	1	0,02	1
38.	PRISSF	*нехристиянин*	1	0,02	1
39.	PRSSSS	*повінчано*	1	0,02	1
40.	RFXX	*хтось-то*	1	0,02	1
41.	RXX	*якось-то*	1	0,02	1

Із 314 кількакореневих морфструктур української мови в поетичних текстах Т. Шевченка було виявлено 23 двокореневі й 1 трикореневу морфструктуру, що реалізовані в 152 словах і становлять ≈ 2,88% лексикону мови Т. Шевченка (Таблиця 4.2).

Таблиця 4.2. Ранговий список ММС складних слів у лексичному реєстрі ідіостилю Т. Шевченка

№	ММС	Приклад реалізації ММС у слові	Питома вага у лексиконі		Абсолютна частота (f) у тексті
			Абсолютна частота (f)	Відносна частота (p %)	
1.	RIRSF	*чорнобривий*	44	0,82	125
2.	RRF	*чималий*	15	0,28	51
3.	RR	*бодай*	14	0,27	70
4.	RIRF	*обидва*	12	0,23	20
5.	RSIRSF	*животворящий*	7	0,13	8
6.	RRS	*чимало*	6	0,11	28
7.	RSRS	*довго-довго*	5	0,09	10
8.	RIRISF	*благодать*	3	0,06	9
9.	RIRSSF	*благовіститель*	3	0,06	3
10.	PRIRS	*достолиха*	3	0,06	3
11.	RFPR	*що-небудь*	2	0,04	8
12.	RIR	*добридень*	2	0,04	5
13.	PRIRF	*перекотиполе*	2	0,04	4
14.	RIRSSSF	*новобранець*	2	0,04	3
15.	RSRF	*Великдень*	2	0,04	3
16.	RSIRSSF	*великомученик*	2	0,04	2
17.	RSRSF	*малоліток*	2	0,04	2
18.	PRIRSF	*поблагословити*	2	0,04	2
19.	RIRS	*якомога*	2	0,04	2
20.	RSSRSS	*далеко-далеко*	1	0,02	1
21.	PRRS	*навіки-віки*	1	0,02	1
22.	RSPRSS	*рано-вранці*	1	0,02	1
23.	PRSRS	*спідтиха-тиха*	1	0,02	1
24.	RRRF	*чортзна-що*	1	0,02	1

Морфемна статистична структура ідіостилю Т. Шевченка представлена обмеженим реєстром одиниць – 65 ММС, що порівняно із лексичним реєстром зменшений майже в сотні разів (лексичний реєстр становить 5279 одиниць). Слова різної морфемної будови об'єднуються в різні, за кількістю, лексичні групи, які мають різну питому вагу в лексиконі. Таке групування формує ранговий список ММС з інтервалом питомої ваги від 23,03% до 0,02%.

Систематизована в таблицях інформація показує, що ≈75,68% лексикону поетичного мовлення Т. Шевченка моделюються за 5-ома моделями морфемних структур із найвищою питомою вагоюву лексичному реєстрі:

RF (*мій*[31]) ≈ 23,03%;
RSF (*плакати*) ≈ 22,85%;
PRSF (*співати*) ≈ 19,17%;
PRF(*ніхто*) ≈ 5,55%;
RSSF(*дівчина*) ≈ 5,08%.

Саме ці структури формують ядро морфеміки поетичного мовлення Т. Шевченка й визначаються як високопродуктивні одиниці морфемної статистичної структури лексикону ідіостилю. Ще ≈18% лексики описуються 10-ма, нижчими за питомою вагою, моделями морфемних структур, які разом з ядровими формують основу морфемної системи поетичних текстів, оскільки продукують слова, що входять до 90% (≈92,93%) лексики. Ці 18% морфемних структур належать до ближньої периферії морфеміки поетичного мовлення Т. Шевченка й формують групу середньопродуктивних одиниць морфемної статистичної структури лексикону ідіостилю:

PRSFX (*подивитися*) ≈ 4,03%;
RSFX(*дивитися*) ≈ 2,20%;
PRSS(*вранці*) ≈ 1,99%;
PRSSF(*наймичка*) ≈ 1,89%;
R (*де*) ≈ 1,7%;
PRS(*знову*)≈ 1,49%;
RSS(*тяжко*)≈ 1,46%;
RS (*добре*) ≈ 1,14%;
PPRSF(*заспівати*) ≈ 1,08%.

Усі інші морфемні структури: 27 однокореневих, 23 двокореневі, 1 трикоренева – характеризуються продуктивністю меншою одного відсотка й реалізуються в 7 % лексики. Такі структури належать до периферії морфеміки поетичного мовлення Т. Шевченка й належать до розряду низькопродуктивних одиниць морфемної статистичної структури лексикону ідіостилю.

[31]Для прикладу наводяться слова, які в межах вибірки слів однієї морфемної структури мають найвищий показник абсолютної та відносної частоти.

Отже, гіпотеза про дію закону переваги в розподілі морфемних структурних одиниць (ММС) у системі лексикону Т. Шевченка підтверджується: 5 високопродуктивних ММС формують ≈ 75 % лексикону; 9 середньопродуктивних ММС формують ≈ 18 % лексикону; 51 низькопродуктивна ММС формує ≈ 7 % лексики. Розподіл продуктивності моделей морфемних структур слів у морфемній статистичній структурі лексикону Т. Шевченка розглядаємо як стилеметричну ознаку ідіостилю, яка потребує перевірки в зіставному дослідженні з іншими ідіостилями[32].

Оскільки вважається, що стилістично маркованими є одиниці, які належать до низькочастотних, розглянемо насамперед вибірку слів із ММС, які мають питому вагу в лексиконі менше 1 %. Як свідчать дані табл. 4.1 та табл. 4.2, серед низькопродуктивних ММС лексикону є моделі, які характеризуються високим індексом покриття тексту (f > 70): PR; RSSSF; PRFX; RFX. Окремої уваги заслуговує морфструктура RISFX, яка реалізована лише в одному слові – *сміятися /сміjатися/* – з питомою вагою p 0,02 %, але має f 57 у тексті. Натомість, слова всіх інших 46 низькопродуктивних ММС (у статистичному інтервалі: p 0,02 % ≤ питома вага ≤ p 0,28%; f 1 ≤ абсолютна частота <f 70) характеризуються корелятивними відношеннями при зіставленні кількості утворюваних слів та їх активності в покритті тексту. Крім того, ММС слів цього низькочастотного діапазону характеризуються складною і нетиповою організацією морфемної будови слова в аспекті валентності різних функціональних типів морфем. Отже, низькопродуктивні ММС статистичного інтервалу питомої ваги в лексиконі – 0,02 % ≤ p ≤ 0,28% – формують гіпотетичну вибірку стилістично маркованої лексики. До цієї вибірки потрапляють слова, моделі морфемних структур яких у лексиконі і тексті мають корелятивні низькочастотні статистичні показники або характеризуються великим ростом статистичної активності у тексті, порівняно із лексиконом.

Зіставлення абсолютної частоти ММС у лексиконі ідіостилю (4-та колонка табл. 4.1 та 4.2) з їх абсолютною частотою вживання в текстах (6-та колонка табл. 4.1 та 4.2) показує деяку нерівномірність реалізації ММС у лексиконі та тексті ідіостилю). Наприклад, ММС R (*де*) реалізована в 91 слові лексикону, але в тексті має f 6605; ММС PRSF (*співати*) реалізована у 1012 словах лексикону, але в тексті має f 3804.

Використовуючи метод ранжування частот, представимо графічну модель морфемної статистичної структури ідіостилю. Припустимо, що система ординат *x* та *y* формує деяку двовимірну площину, яку можна визначити площиною ідіостилю. Ставиться завдання: відобразити на цій площині точки, які визначають місце ММС на площині ідіостилю. Оскільки за моделлю морфемної статистичної структури стилю, ми визначаємо статистичну реалізацію ММС у двох системах (лексиконі і тексті), то кожна

[32] Порівняльне стилеметричне дослідження описано в наступному параграфі.

ММС буде мати в площині ідіостилю дві точки, кожна з яких формує окрему лінію на графіку:

1) лінію продуктивності ММС у лексиконі;

2) лінію продуктивності ММС у тексті.

Оберемо для графічного моделювання 19 найпродуктивніших у лексиконі ММС з абсолютною частотою в статистичному інтервалі $1216 \leq f \leq 19$ та ті ж самі 19 ММС, з урахуванням абсолютної частоти в тексті (в статистичному інтервалі $16322 \leq f \leq 77$. У цій закритій системі ММС ранжування може мати тільки 19 рангів. Використовуючи ранговий список ММС двох попередніх таблиць будуємо нову таблицю (табл. 4.3), в якій записуємо 19 моделей морфемної структури слова за спадом абсолютної частоти у лексиконі (колонка "ММС").

Таблиця 4.3. Обернене ранжування ММС слів поетичного мовлення Т. Шевченка

№	ММС		Приклад лексичної реалізації ММС	Лексичний реєстр		Текст	
				f	Ранговий номер	f	Ранговий номер
1	RF	75,68 %	*мій*	1216	19	16323	19
2	RSF		*плакати*	1206	18	7482	18
3	PRSF		*співати*	1012	17	3804	16
4	PRF		*ніхто*	293	16	1631	15
5	RSSF		*дівчина*	268	15	898	14
6	PRSFX	92,93 %	*подивитися*	213	14	771	13
7	RSFX		*дивитися*	116	13	683	12
8	PRSS		*вранці*	105	12	254	7
9	PRSSF		*наймичка*	100	11	212	6
10	R		*де*	91	10	6605	17
11	PRS		*знову*	79	9	499	10
12	RSS		*тяжко*	77	8	359	8
13	RS		*добре*	60	7	604	11
14	PPRSF		*заспівати*	57	6	186	5
15	RIRSF		*чорнобривий*	44	5	125	3
16	PR		*нехай*	37	4	491	9
17	RSSSF		*пташечка*	31	3	77	1
18	PRFX		*довестися*	28	2	150	4
19	RFX		*дітися*	19	1	110	2

Цей ранговий список ММС відкладаємо на вісі *x* (рис.4.1), де перший ранг займає ММС із найвищим значенням абсолютної частоти. З метою кореляції чисел рангу із числами, що відображають статистичні характеристики ММС, здійснюється обернене ранжування ММС (колонка "ранговий номер") у зворотному напрямку рангів закритої системи одиниць: найвищий ранг (19) буде мати ММС RF із найвищою абсолютною частотою, а найнижчий ранг (1) буде мати ММС RFX із найнижчою абсолютною частотою. За системою оберненого ранжування приписуємо кожній ММС ранговий номер від 19 до 1 за спадом значення абсолютної частоти у лексичному реєстрі та тексті. Обернений ранговий список абсолютних частот лексичного реєстру (4-та колонка табл. 4.3) відповідає послідовності

розташування ММС в еталонному списку (2-га колонка), укладеному за прямим ранжуванням – вісь *x*. Обернений ранговий список абсолютних частот ММС у тексті порушує послідовність рангів, тому що ранг приписується кожній ММС за еталонним списком моделей 2-ої колонки. Обернене ранжування дозволить на графіку наочно показати зниження абсолютної частоти: обернені ранги відкладаються на вісі *y*.

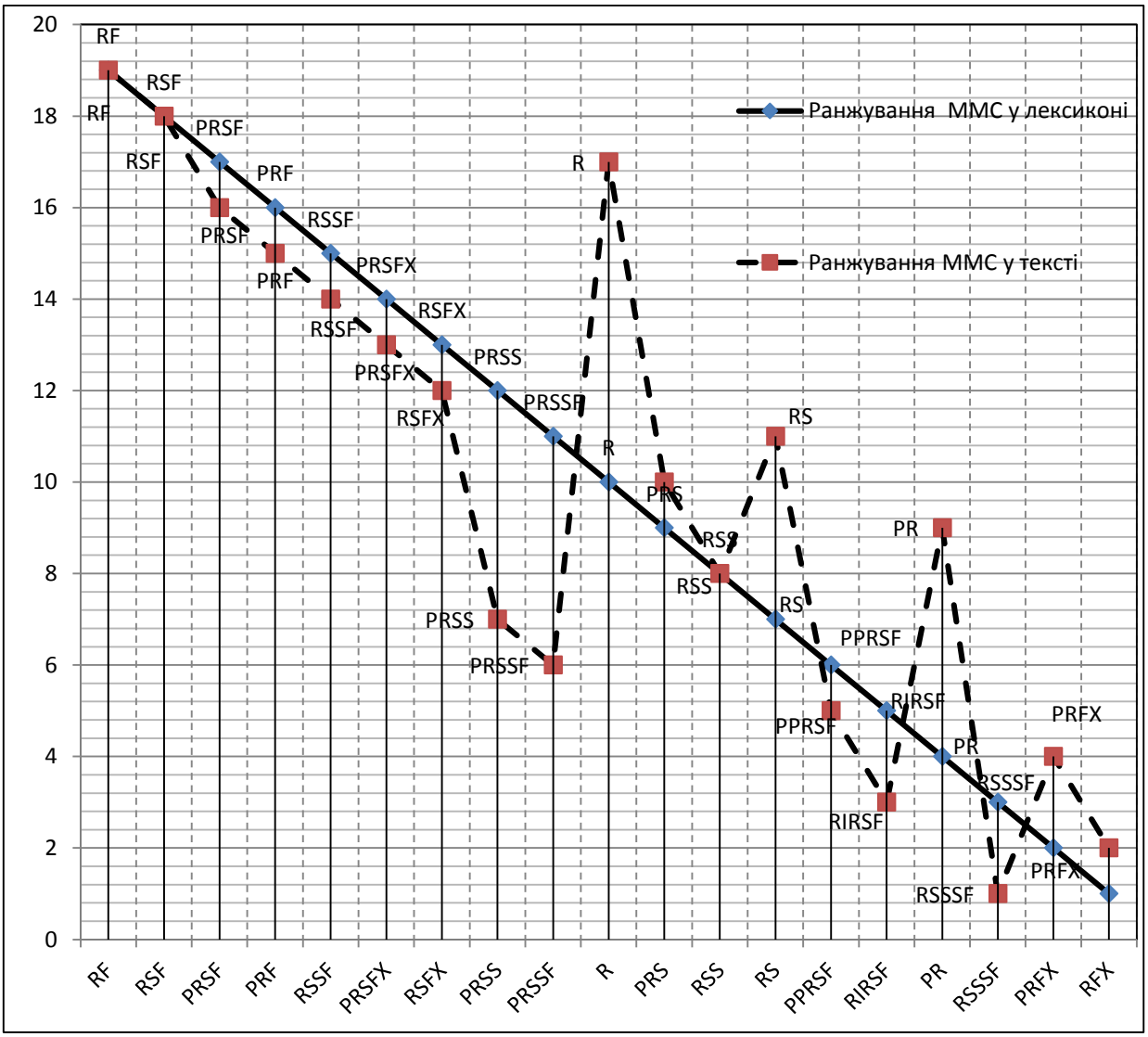


Рис. 4.1. Графічна модель морфемної статистичної структури ідіостилю Т. Шевченка за обереженим ранжуванням статистичних характеристик

Кожен обернений ранговий список моделюється на графіку окремою лінією. Суцільна спадна пряма лінія демонструє точки ММС у площині морфемної статистичної структури лексикону: ранги ММС рівномірно знижуються, тому що еталонний список ММС формувався за прямим ранговим списком абсолютних частот у лексичному реєстрі, відповідно числові показники рангів збігаються на обох осях ординат: $x(1)=y(19)$ ...

x(19)=*y*(1). Пунктирна ламана лінія моделює точки ММС у площині морфемної статистичної структури тексту й відображає підвищення або зниження рангу, порівняно з, рівномірно спадною, ранговою лінією лексикону. Зіставлення двох ліній графіка показує невідповідність ранжування абсолютних частот ММС у лексиконі та тексті й дозволяє описати деякі особливості моделі морфемної статистичної структури ідіостилю Т. Шевченка.

Три верхні піки ламаної ранжування ММС у тексті, які представляють точки R (*де*), RS (*добре*), PR (*нехай*), демонструють ріст рангу в тексті, порівняно з рангом у лексиконі. Ці морфемні структури не належать до ядрових морфемних структур лексикону Т.Шевченка, але, як свідчить графік ранжування, входять до множини високочастотних у тексті. Підвищується у тексті ранг периферійної морфструктури PRFX (*довестися*), що формує четвертий верхній пік ламаної ранжування ММС у тексті. Нижні піки ламаної морфемної статистичної структури тексту демонструють зниження рангу, порівняно з рангом лексичної продуктивності, у ММС PRSS (*вранці*); PRSSF (*наймичка*); RIRSF (*чорнобривий*); RSSSF (*пташечка*).

Отже, співвідношення слів однієї ММС у лексиконі та тексті не відповідає дії закону Дж. Ципфа (невелика кількість слів покриває велику частину тексту), тому що розподіл слів у лексиконі і тексті за моделлю морфемної статистичної структури ідіостилю здійснюється не за математичною моделлю – варіантою абсолютної частоти, а за лінгвістичною ознакою – морфемною будовою слова. Тому невідповідність співвідношення статистичної "поведінки" ММС у лексиконі і тексті не можна пояснити статистичними законами, але, очевидно, можна пояснити лінгвістичними закономірностями морфемної будови слова.

Для розуміння розподілу абсолютних частот слів у тексті різної морфемної будови, розглянемо, які абсолютні частоти можуть мати слова однієї ММС. Кожна ММС реалізовується в лексиконі у визначеній групі слів, кожне з яких може вживатися в тексті з різною абсолютною частотою: високою і низькою. Абсолютні частоти слів однієї ММС можуть варіювати. Різні значення абсолютної частоти слів з однаковою ММС формують варіативний ряд. Варіативний ряд – послідовність чисел (у нашому випадку значень f), розташовану в порядку збільшення їх величини. Кожне число в такій послідовності є варіантою – x_i. Проміжок між крайніми членами варіативного ряду (варіантами) називають інтервалом варіювання, а довжину цього інтервалу (кількість варіант) – розмахом. Розглянемо варіативні ряди абсолютних частот слів у множині одиниць однієї лексичної вибірки, сформованої за типовою лінгвістичної ознакою – слова вибірки мають однакову морфемну будову, тобто реалізують одну ММС. Таблиця 4.4 систематизує дані про варіанти абсолютних частот слів у лексичних вибірках пікових ММС.

Таблиця 4.4. Варіативні ряди абсолютних частот слів у лексичних вибірках пікових ММС ламаної тексту

ММС верхніх піків						
ММС	**Абсолютна частота у лексиконі (f)**	**Абсолютна частота у тексті (f)**	**Варіативний ряд (x_i) абсолютних частот у тексті**	**Кількість слів із варіантою x_i =1**	**Кількість слів інтервалу варіювання $1 \leq x_i \leq 10$**	
					у лексиконі	**у тексті**
R	91	6605	1, 2, 3, 4, 5, 6, 7, 9, 11, 12, 13, 19, 21, 23, 24, 25, 27, 28, 29, 31, 37, 38, 49, 52, 53, 56, 57, 62, 63, 71, 75, 82, 89, 91, 92, 102, 166, 190, 191, 212, 249, 312, 747, 1051, 1715	13	45	168
RS	60	604	1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 13, 15, 16, 18, 25, 26, 27, 33, 49, 123	18	44	128
PR	37	491	1, 2, 3, 4, 5, 6, 10, 11, 13, 14, 18, 19, 28, 31, 32, 59, 92, 98	14	25	58
ММС нижніх піків						
PRSS	105	254	1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 15, 17	61	101	199
PRSSF	100	212	1, 2, 3, 4, 5, 6, 8, 11, 22	64	98	179
RIRSF	44	125	1, 2, 3, 5, 8, 9, 34	26	43	91
RSSSF	31	77	1, 2, 3, 4, 5, 7, 9	16	31	77

ММС верхніх піків ламаної об’єднують слова як з низькими, так і з високими абсолютними частотами в таких інтервалах варіювання:

R: 1 – 1715; RS: 1 – 123; PR: 1 – 98.

Варіативні ряди цих ММС мають найбільший розмах (R: 46 варіант), причому в цих вибірках слова з абсолютними частотами від 1 до 10 становлять дуже малу частину, а слова з абсолютною частотою 1 утворюють лексичні вибірки обсягом менше 20 одиниць. Тому в цих випадках закон Дж. Ципфа спрацьовує.

ММС нижніх піків ламаної об’єднують слова з низькими абсолютними частотами в таких інтервалах варіювання:

PRSS: 1 – 17; PRSSF: 1 – 22; RIRSF: 1 – 34; RSSSF: 1 – 9.

Варіативні ряди ММС нижніх піків мають малий розмах (PRSS: 13 варіант), причому в цих вибірках переважають слова з абсолютними частотами від 1 до 10, а також велика частина слів з абсолютною частотою 1. Відповідно для таких моделей закон Дж. Ципфа не спрацьовує.

Закон Дж. Ципфа не діє у розподілі слів морфемної статистичної структури ідіостилю в таких випадках:

1) моделі з високою лексичною продуктивністю RF, RSF, PRSF, PRF, RSSF, PRSFX, RSFX мають і високу абсолютну частоту в тексті, яка за ранжуванням у 5-ти останніх моделях знижується тільки на один ранг;

2) моделі із середньою лексичною продуктивністю PRSS, PRSSF, R, PRS, RSS, RS, PPRSF, RIRSF, PR, RSSSF, PRFX, RFX можуть мати різну абсолютну частоту в тексті й, очевидно, це залежить від морфемної довжини слова.

Як відомо, Дж. Ципф [Zipf 1949] пояснював виведений ним математичний закон про співвідношення абсолютної частоти й рангу слів (абсолютної частоти й кількості слів) психофізіологічними властивостями мовної діяльності: «у мові взаємодіють дві протилежні тенденції – прагнення до спрощення артикуляції високочастотних слів і обмеження обсягу реєстру та пошуки максимальної точності висловлювання, звідки випливає збільшення реєстру» [цитовано за Перебийніс 1985: 132]. Відповідно, закон Дж. Ципфа пояснює піки спаду і росту абсолютної частоти слів однієї ММС у тексті (Рис. 4.1) кількісними характеристиками морфемної довжини слова: ММС верхніх піків ламаної тексту R, RS, PR – одно- та двоморфемні слова (тобто короткі слова), а ММС нижніх піків ламаної тексту PRSS, PRSSF, RIRSF, RSSSF – чотири- і п'ятиморфемні слова (тобто довгі слова).

Морфемна довжина слів. Морфемна довжина слів впливає на їхні статистичні характеристики, тому в стилеметричному аналізі ідіостилю Т. Шевченка ставиться завдання дослідити кількісні характеристики ММС слів і проаналізувати співвідношення морфемної довжини та морфемної глибини слів із їх лексичною продуктивністю.

Оскільки морфемні структури слів досліджуються у лексичному реєстрі, згенерованому за текстами Т. Шевченка, то зіставлення кількості слів, в яких реалізовується ММС різної морфемної довжини, можна проводити із лексичним реєстром сучасної української мови. Зіставним еталоном було обрано систему мови, а саме – дані морфемно-словотвірного фонду української мови (МСФ) про закономірності кількісної організації морфемних структур слів та їх статистичні характеристики, які проаналізовано Н. Клименко в праці «Основи морфеміки сучасної української мови» [Клименко 1998].

У Частотному словнику морфемних структур поетичного мовлення Т. Шевченка було використано ті ж символи моделювання морфемної структури слова, що й у МСФ, але з такими відмінностями у визначенні функціональних типів посткореневих афіксів: I – інтерфікс у нашому дослідженні визначається не тільки як міжкоренева сполучна морфема в складних словах, а й як афіксальна морфема в простих словах; X – постфікс у нашому дослідженні моделюється окремим символом, а в МСФ він визначається як суфікс – S; афікс інфінітива в МБД АСМСА визначається як флексія – F, а в МСФ – як суфікс S. Таким чином, варіативність моделей морфемних структур слів у двох системах моделювання відрізняється, але морфемно-кількісна характеристика цих моделей збігається, тому що сегментація у двох системах здійснювалася за морфемним словником І. Яценка [Яценко 1980]. Це дозволяє провести зіставний аналіз морфемних структур в аспекті морфемної довжини слова, яка може бути стилерозрізнювальною характеристикою.

Статистичний аналіз показує, що 65 моделей морфемних структур лексикону Т. Шевченка реалізуються в словах різної морфемної довжини, регульованої певними закономірностями кількісного обмеження. За даними

МСФ [Клименко 1998: 150 – 153], морфемна довжина українського слова обмежена інтервалом 1 – 13 морфем: простих слів інтервалом 1 – 11, складних слів 2 – 13. Морфемна довжина слова лексикону Т. Шевченка обмежена інтервалом 1 – 7 морфем: простих слів 1 – 6; складних 2 – 7. Крім того, морфемна довжина слова по-різному впливає на продуктивність морфструктури в словах системи української мови та в лексиконі Т. Шевченка. Закономірність такого впливу зручно описати за допомогою графічного моделювання (рис. 4.2).

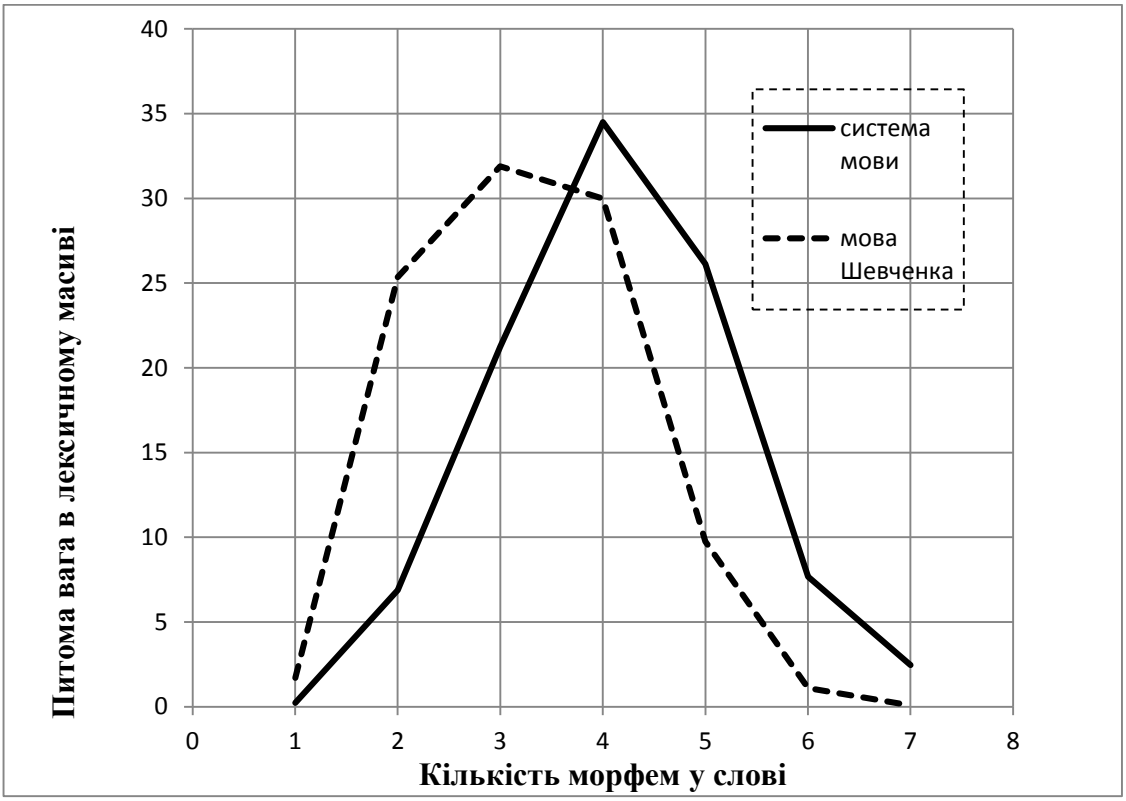


Рис. 4.2. Графік розподілу морфемної довжини слова у лексиці системи української мови та лексиконі поетичного мовлення Т. Шевченка

Найчисленнішими в українській мові є 3- (≈21,24%), 4- (≈ 34,5%), 5-морфемні слова (≈ 26,13%), що складають ядро морфемної системи мови й становлять ≈ 81,9% лексики. На графіку ці слова формують вершину ламаної «система мови».

У лексиконі поетичного мовлення Т. Шевченка найвищу продуктивність (вершина ламаної "мова Шевченка") мають 2- (≈ 25,35%), 3- (≈ 31, 85%), 4-морфемні слова – ≈ 29,99%, які становлять ≈ 87,29% лексики. 5-морфемні структури, на відміну від системи мови, становлять лише ≈ 9,75%, натомість 2-морфемні структури є продуктивними, хоча в системі мови вони становлять лише ≈ 7%. Одноморфемні структури в лексиконі Т. Шевченка мають питому вагу ≈ 1,7%, а в системі мови ≈ 0,22 %; 6-морфемні ≈ 1,1% (у мові ≈ 7,67%); 7-морфемні ≈ 0,08% (у мові ≈ 2,46%). Як показує ламана мови Т. Шевченка, зі збільшенням морфемної довжини слова до п’яти морфем їх питома вага в лексиконі Т. Шевченка різко зменшується,

а в 6-морфемних структурах падає до 1%. 7-морфемні структури представлені лише чотирма складними словами (*богобоязливий, новобранець, великомученик, великомучениця*) і не характерні для простих слів. Зменшення продуктивності морфструктур зі збільшенням морфемної довжини слова до 5 морфем проявляє дію закону простоти і переваги у морфемній будові слова лексикону Т. Шевченка. Натомість, у системі мови межею простоти морфемної будови слова є його довжина із 6 морфем [Клименко 1998: 157].

Кількісне зменшення морфемної довжини слова в лексиконі Т. Шевченка підтверджується й зіставленням показників середньої морфемної довжини слова. За даними МСФ середня морфемна довжина українського слова обмежується інтервалом 4 ± 1, адже пік зростання морфемної довжини припадає на 4-морфемні структури: «У морфемній системі перевага віддається словам з кількістю морфем, близькою до середньої глибини, що вимірюється 3,9 морфемами. Серед простих слів найбільше чотириморфемних слів, серед складних – п'ятиморфемних. Отже, в системі віддається перевага словам з глибиною 4 ± 1 морфема» [Клименко 2014а: 228]. У лексиконі Т. Шевченка найвищу продуктивність мають 3-морфемні (≈ 31, 85%) та 4-морфемні структури (≈ 29,99%), показники яких відрізняються одним відсотком. Вирахувавши середню морфемну довжину слова лексикону Т. Шевченка за формулою $\overline{x} = \frac{\sum x_i \cdot n_i}{\sum n_i}$ (x_i – кількість морфем морфструктури, n_i – кількість слів, у яких реалізована ця морфструктура), можна достовірно визначити, що $\overline{x}$ = 3,24. Таким чином, середня морфемна довжина слова поетичного мовлення Т. Шевченка входить до інтервалу 4 ± 1, але Т. Шевченко використовує морфемно-коротші слова із середньою довжиною 3,24.

Така статистика підтверджується стилістичними дослідженнями середньої фонемної довжини слова, довжини слова у складах та морфемної довжини слова [СПС 1967], [Москович 1967]: у поетичних текстах вживаються найкоротші слова, порівняно з іншими стилями. В. Москович це явище пояснює зміною функції тексту: «У більшості стилів мови основною є предметна функція, яка ставить у центрі уваги сам предмет повідомлення. У цих стилях спостерігається звичайна картина розподілу слів за глибиною і довжиною з оказіональним перевищенням максимумів глибини і довжини (наприклад, у стилі наукової прози, розглянутому вище). У тих же випадках, коли основною для певного стилю є не предметна, а інша функція – вокативна, виражальна або естетична – увага носіїв мови відволікається від предмета повідомлення й переходить або на адресат повідомлення (вокативна функція), або на його адресант (виражальна функція), або на саму структуру слова (естетична функція). Зміна уваги з предмета повідомлення на інші його аспекти викликає необхідність у такій зміні складу повідомлення, щоб у ньому були відсутні елементи, які ускладнюють сприйняття повідомлення, тобто, зокрема, слова з наближеною до

максимальної глибиною і довжиною. Цим пояснюється зменшення до мінімуму в поезії кількості слів з глибиною і довжиною близькою до максимальної» [Москович 1967: 29].

Отже, середня морфемна довжина слова 3,24 виступає як маркована ознака поетичного стилю, проте для визначення цієї ознаки маркером ідіостилю Т. Шевченка необхідно порівняти показник середньої морфемної довжини слова у поетичних текстах різних авторів.

Морфемна глибина слова. Визначення морфемної глибини слів як тактів розгортання морфемної структури слова відносно кореня здійснюється у мовознавстві двома способами: 1) описом морфемної будови слів у термінах аплікативної породжувальної моделі С. Шаумяна [Шаумян 1963]; 2) описом морфемної будови слів за моделлю сітки П. Менцерата [Menzerath 1954].

Використовуючи методику опису розгортання фонологічної структури слова – сітки П. Менцерата, яку Н. Клименко [Клименко 1998], [Клименко 1998б] та Є. Карпіловська [Карпіловська 1992] використали в описі морфемної будови слів, можна представити системний опис морфемних структур слів ідіолекту Т. Шевченка, що демонструє розгортання найпростішої морфемної будови слова до кількісно найскладнішої морфструктури як реально існуючої, так і потенційно можливої.

Таблиця 4.5. Морфемна сітка простих слів поетичного мовлення Т. Шевченка

3 ~~3PR~~	2 2PR 0.02 %	1 PR 0,70%	 **R 1,70%**
4 ~~3PRA~~	3 2PRF 0,27% 2PRS 0,09%	**2** **PRF 5,55%** **PRS 1,49%**	**1** **RF 23,03%** **RS 1,14%** RX 0,25
5 3PRSF 0,02 %	4 **2PRSF 1,08%** 2PR2S 0,11%	**3** **PRSF 19,17%** **PR2S 1,99** PRFX 0,53%	**2** **RSF 22,85%** **R2S 1,46%** RFX 0,36% R2X 0,02%
6 ~~3PR3A~~ 7 морфем	5 2PR2SF 0,19% 2PRSFX 0,19% 2PRI2S 0,02%	4 **PR2SF 1.89%** PR3S 0,27% **PRSFX 4,03%** PRISF 0,06%	3 **R2SF 5,08** R3S 0,11% **RSFX 2,20%** RF2X 0,02% RISF 0,28%
7 ~~3PR4A~~	6 ~~2PR4A~~ 7 морфем	5 PR3SF 0,25% PR4S 0,02% PRI2SF 0,02%	4 R3SF 0,59% R4S 0,04% RI2SF 0,04% RISFX 0,02% R2SFX 0,02%
8 ~~3PR5A~~	7 ~~2PR5A~~	6 ~~PR5A~~ 7 морфем	5 R4SF 0,06%

Морфемна сітка (табл. 4.5) відображає закономірності конструювання морфемних структур простих слів в аспектах морфемної довжини та морфемної глибини у такій кореляції кількісних параметрів: кожен породжувальний такт відображає збільшення морфемної структури слова на одну морфему, виходячи з нульового такту, який представляє слово-корінь. Тому кожен номер такту об'єднує ММС, кількість морфем у яких на одиницю більша за номер такту, наприклад: такт № 1 об'єднує двоморфемні слова, а такт № 5 – шестиморфемні слова. Відповідно, кількість тактів обмежена максимальною морфемною довжиною слова із зменшенням на одиницю: максимальна морфемна довжина простих слів у лексиконі Т. Шевченка становить 6, тому кількість породжувальних тактів морфемної структури слів у моделі сітки буде визначатися 5-ма тактами.

Таким чином, як показують статистичні дані у морфемній сітці, зі збільшенням породжувальних тактів від 1 до 4 продуктивність ММС становить більше 1% (ланки, до яких належать моделі з продуктивністю вище 1% виділені темним фоном), а зі збільшенням породжувальних тактів до п'яти питома вага ММС у лексиконі Т. Шевченка становить менше 1%. Ця статистична закономірність підтверджує висновок В. Московича про те, що «...більшість слів природних мов породжується на 1 – 4 тактах породжувального процесу; верхньою межею кількості тактів породження в конкретній мові є максимальна глибина слів у цій мові» [Москович 1967: 33], проте розподіл ММС за тактами свідчить про те, що в ідіостилі Т.Шевченка не всі морфемні моделі 1 – 4 тактів є продуктивними.

Морфемна сітка демонструє розгортання морфемної структури слова від найпростішої моделі R (*де*) у двох напрямках кількісного ускладнення:

1) лівобічного – додавання префіксальної частини слова по горизонталі справа – наліво;

2) правобічного – додавання постфіксальної частини слова по вертикалі зверху – вниз.

Морфструктура R може ускладнюватись за рахунок додавання афіксів у препозиції (2 морфструктури), у постпозиції (18 морфструктур), а також відцентрично (відносно кореня) у двох напрямках (20 морфструктур).

Препозитивна зона морфструктур може збільшуватися лише за рахунок префіксальних морфів. У поетичному мовленні Т. Шевченка максимальне препозитивне розгортання морфструктури реалізується у 3-афіксній префіксальній послідовності непродуктивної морфструктри PPP←RSF (*не-с-по-ви-т-ий*), що лексично представлена одним шестиморфемним словом (має абсолютну частоту вживання – 1). Чотирьохпрефіксальних морфструктур, які характерні для морфеміки української мови, у поетичному мовленні Т. Шевченка не виявлено. Одноафіксні та двоафіксні префіксальні послідовності реалізуються як в однобічному напрямку розгортання морфструктури, так і у двобічному.

Однобічне препозиційне ускладнення характерне для двох непродуктивних морфструктур: P←R (*не-хай*) ≈ 0,70% та PP←R (*не-в-лад*) ≈ 0,02%.

Модель двобічного афіксального ускладнення морфемної структури слова (рис. 4.3, рис. 4.4) представлена 19 морфструктурами.

3M	**P ←**	R	→	**A**	**PRF**[33]*ніхто* 5,55% PRS*знову* 1,49%
4M			→	**2A**	**PRSF***співати* 19,17% **PR2S***вранці* 1,99% PRFX*довестися* 0,53
5M			→	**3A**	**PR2SF***наймичка* 1,89% PR3S*звичайне* 0,27% **PRSFX***подивитися* 4,03% PRISF*незнаємий* 0,06%
6M			→	**4A**	PR3SF*заквітчаний* 0,25 PR4S*повінчано* 0,02% PRI2SF*нехристиянин* 0,02%

Рис. 4.3. Модель двобічного ускладнення морфемної структури слова з одним префіксом

Як показує модель на рис. 4.3, двобічне розгортання морфемної структури з додаванням одного префікса корелює з ускладненням постфіксальної одно-, двох-, трьох-, чотирьохафіксної послідовності. Такі морфемні структури реалізуються в 3-морфемних словах (≈ 7,04%), 4-морфемних (найвища продуктивність ≈ 21,69%), 5-морфемних (≈ 6,25%) та 6-морфемних (найнижча продуктивність ≈ 0,29%).

4M	**2P ←**	**R**	→	**A**	PPRF *сповити* 0,27% PPRS *незабаром* 0,09%
5M			→	**2A**	PPRSF *заспівати* 1,08% PPR2S *взаперті* 0,11%
6M			→	**3A**	PPR2SF *приспівувати* 0 ,19% PPRSFX *простягатися* 0,19% PPRI2S *анікогісінько* 0,02%

Рис. 4.4. Модель двобічного ускладнення морфемної структури слова з двома префіксами

Двобічне розгортання морфемної структури з додаванням двох префіксів (рис. 4.4) корелює з ускладненням постфіксальної одно-, дво-, трьохафіксної послідовності. Такі морфемні структури є малопродуктивними й реалізуються в 4-морфемних (≈ 0,36%), 5-морфемних (≈ 1,19%), 6-морфемних (≈ 0,40%) словах.

[33] Жирним шрифтом визначені продуктивні морфструктури, які формують основу морфемної системи поетичного мовлення Т.Шевченка.

Лівобічне ускладнення морфструктури префіксами обмежується кількісними параметрами морфемної довжини простих слів лексикону Т. Шевченка – 6 морфем, що регламентують кореляцію афіксів префіксальної зони з афіксальними послідовностями посткореневої зони: однопрефіксні морфемні структури корелюють із п'ятиафіксними постфіксальними послідовностями, а двопрефіксні морфемні структури не утворюють слів і чотирьохафіксними та п'ятиафіксними структурами посткореневої зони.

Постпозиційна зона морфструктур ускладнюється за рахунок додавання:

1) функціонально однотипних афіксів, наприклад:
   флексій (R→F (*мій-0*); PR→F (*ні-хт-о*));
   суфіксів (R→S (*добр-е*); R→SS (*тяж-к-о*));
   постфіксів (R→X (*де-сь*); R→XX (*як-ось-то*);
2) функціонально різних афіксів, наприклад:
   суфікс+флексія R→SF (*плак-а-ти*);
   суфікс+флексія+постфікс R→SFX (*див-и-ти-ся*);
   інтерфікс+суфікс+флексія R→ISF (*гу-к-а-ти*).

Функціонально різні, а також однотипні афікси, за винятком флексій, можуть поєднуватися за правилами комбінаторики в сталі посткореневі афіксальні послідовності, які виступають як лінгвістичні одиниці мовної структури, що утворюють парадигму типізованих морфемних структур посткореневої зони. Таким чином, у правобічному напрямку (за морфемною сіткою згори – вниз по вертикалі) морфемні структури слів розгортаються додаванням одиничних афіксів або цілісних афіксальних структур, що складаються з 2, 3, 4, 5 морфем. Афіксальні моделі цих структур подано у табличному записі (табл. 4.6).

Найвищу продуктивність у правобічному розгортанні морфемної структури мають флексії та двохафіксна посткоренева структура SF, які характеризуються високою продуктивністю як у безпрефіксних словах (R→**F**(*мій-0*) ≈ 23,03%; R→**SF** (*плак-а-ти*) ≈ 22,85%), так і в кореляції з одним префіксом (PR→**F**(*ні-хт-о*) ≈ 5,55%; PR→**SF** (*с-пі-ва-ти*) ≈ 19,17%).

Одноафіксна посткоренева зона репрезентована трьома функціональними типами морфем – флексією (висока продуктивність), суфіксом та постфіксом, які мають низьку продуктивність і реалізовані в 3% слів. Флексія та суфікс беруть участь у двобічному розгортанні морфемної структури й корелюють з одним або двома префіксами. Постфіксальне розгортання морфемної структури R→**X** (*де-сь*) 0,25% може бути тільки правобічним.

Двохафіксні посткореневі афіксальні послідовності описуються такими структурними моделями: SF (43,12%) із високим ступенем лексичної продуктивності; SS (3,56%) – середнім; FX (0,89%) XX (0,02%) – низьким. Суфіксально-флективна та двосуфіксальна морфемні послідовності забезпечують правобічне ускладнення морфемної структури як самостійно, так і в кореляції з одним та двома префіксами, крім того, структура SF

утворює слово з трьома префіксами. FX розгортає морфемну структуру самостійно та в кореляції з одним префіксом, а двопостфіксальна структура XX реалізується лише в одному безпрефіксальному слові (*якось-то*).

Таблиця 4.6. Афіксальні моделі посткореневої зони простих слів поетичного мовлення Т.Шевченка

Кількість афіксів посткоре-невої зони	**Афіксальні моделі посткореневої зони**	**Моделі морфемних структур слів**
1 (31,22%)	**F (28,25%)**	R**F** (*мій-0*) 23,03%
		PR**F***(ні-хт-о)* 5,55%
		PPR**F***(с-по-ви-ти)* 0,27%
	S (2,72%)	R**S***(добр-е)* 1,14%
		PR**S***(з-нов-у)* 1,49%
		PPR**S***(не-за-бар-ом)* 0,09%
	X (0,25%)	R**X***(де-сь)* 0,25%
2 (47,59%)	**SF (43,12%)**	R**SF***(плак-а-ти)* 22,85%
		PR**SF***(с-пі-ва-ти)* 19,17%
		PPR**SF***(за-с-пі-ва-ти)* 1,08%
		PPPR**SF***(не-с-по-ви-т-ий*) 0,02%
	SS (3,56%)	R**SS***(тяж-к-о)* 1,46%
		PR**SS***(в-ран-ц-і)* 1,99%
		PPR**SS***(в-за-пер-т-і)* 0,11%
	FX (0,89%)	R**FX***(ді-ти-ся)* 0,36%
		PR**FX***(до-вес-ти-ся)* 0,53%
	XX (0,02%)	R**XX***(як-ось-то)* 0,02%
3 (14,34%)	**SSF (7,16%)**	R**SSF***(дів-ч-ин-а)* 5,08%
		PR**SSF***(на-йм-ич-к-а*) 1,89%
		PPR**SSF***(при-с-пі-в-ува-ти)* 0,19%
	SSS (0,38%)	R**SSS***(ниш-ч-ечк-ом)* 0,11%
		PR**SSS***(з-вич-ай-н-е)* 0,27%
	SFX (6,42%)	R**SFX***(див-и-ти-ся)* 2,20%
		PR**SFX***(по-див-и-ти-ся*) 4,03%
		PPR**SFX***(про-стяг-а-ти-ся)* 0,19%
	FXX (0,02%)	RF**XX***(хт-о-сь-то)* 0,02%
	ISF (0,34%)	R**ISF***(гу-к-а-ти)* 0,28%
		PR**ISF**(*не-зна-й-ем-ий*) 0,06%
	ISS (0,02%)	PPR**ISS***(а-ні-к-ог-ісіньк-о*) 0,02%
4 (1%)	**SSSF (0,84%)**	R**SSSF***(пт-аш-еч-к-а*) 0,59%
		PR**SSSF***(за-квіт-ч-а-н-ий)* 0,25%
	SSSS (0,06%)	R**SSSS***(сп-а-т-оньк-и)* 0,04%
		PR**SSSS***(по-він-ч-а-н-о*) 0,02%
	ISSF (0,06%)	R**ISSF***(претор-і-ан-ин-0*) 0,04%
		PR**ISSF***(не-христ-ий-ан-ин-0*) 0,02%
	ISFX (0,02%)	R**ISFX***(смі-й-а-ти-ся*) 0,02%
	SSFX (0,02%)	R**SSFX***(квіт-ч-а-ти-ся)* 0,02%
5(0,06%)	**SSSSF(0,06%)**	R**SSSSF***(дів-ч-ат-оч-к-о*) 0,06%

Одноафіксні й двохафіксні моделі посткореневих зон ускладнюють морфемні структури слів, які мають найбільше лексичне наповнення: вони реалізуються у ≈ 78,81 % лексикону поетичного мовлення Т. Шевченка. Нижчу питому вагу лексичної реалізації мають морфструктури з

трьохафіксними посткореневими зонами ≈ 14,34 % і лише 1% складають чотирьохафіксні. П'ятиафіксна посткоренева послідовність представлена тільки в одній морфструктурі з питомою вагою 0,06 %.

Дослідження кількісної організації морфемної будови слів та статистичних параметрів морфемного рівня організації поетичного тексту Т. Шевченка показує, що кількісні та статистичні характеристики морфструктур, які формують відносно невеликий інвентар одиниць, виявляють закономірності будови тексту ідіостилю на морфемному рівні його організації. Статистичний аналіз морфемних структур демонструє лише деякі аспекти стилеметричних досліджень, які можуть проводитися на базі електронних морфемних частотних словників. Використання словників такого типу відкриває широкі можливості й перспективи для застосування інших статистичних методів у дослідженні різних функціональних стилів та ідіостилів.

4.3. Морфемна статистична структура стилю – стилеметрична модель ідіолектів українських поетів

Стилеметричне дослідження морфемного рівня організації поетичного мовлення Лесі Українки, Л. Костенко, В. Стуса та Т. Шевченка проводилося на матеріалі електронних частотних словників морфемних структур слів чотирьох поетів:

- ЧС збірки «На крилах пісень» Лесі Українки [ЧСУкраїнка 2019];
- ЧС збірки «Вибране» Л. Костенко [ЧСКостенко 2019];
- ЧС збірки «Палімпсести» В. Стуса [ЧССтус 2019];
- ЧС мови Т. Шевченка – Твори в п'яти томах [ЧСШевченко 2019].

Систематизовані в частотних словниках статистичні дані про морфемні структури слів у текстах Т. Шевченка, Лесі Українки, Л. Костенко та В. Стуса формують чотири частотні реєстри ММС, на базі вибірок різного обсягу слововживань:

- Т. Шевченка (ТШ) – 68295[34];
- Лесі Українки (ЛУ) – 36058;
- В. Стуса – 36 640 (ВС);
- Л. Костенко (ЛК) – 24238.

Обсяг текстової вибірки формується тільки за слововживаннями, початкові форми яких просегментовані на морфеми. Частина слововживань текстової вибірки не ввійшла до вибірки статистичних обчислень з причини

[34] Кількісні дані обсягу текстової вибірки поетичних творів Т. Шевченка, подані в § 4.2, не збігаються з кількісними даними, які використовуються в § 4.3, тому що статистичне дослідження в попередньому параграфі проводилося за даними ЧМС, укладеними в 2017 р. Після редагування та доповнення МБД АМСА у 2019 р. за результатами списку необробленої лексики (див. § 3.2.2) БД ЧМС були оновлені, текстова вибірка збільшена, а якість автоматичного морфемного аналізу підвищена. Тому обсяг текстових слововживань та обсяг лексичного реєстру слів значно збільшився.

відсутності визначення морфемної будови в початкових формах цих слововживань[35].

З метою вірогідності статистичних даних В. Перебийніс рекомендує в статистичному зіставленні авторських стилів використовувати вибірки обсягом 150 – 200 тис. слововживань для кожного автора [Перебийніс 1985: 17]. Такий обсяг текстового матеріалу необхідний для аналізу лексичних особливостей художньої прози, тому що кількісні масиви лексичної системи текстів художньої прози надзвичайно великі. У стилеметричному дослідження ММС слів поетичного мовлення можна зменшити обсяг вибірок на підставі того, що поетичні тексти набагато менші ніж тексти художньої прози.

Другою вимогою стилеметричних досліджень є кількісне збалансування текстових вибірок, якщо обсяг вибірок різний, у зіставних дослідженнях не можна використовувати показник абсолютної частоти. Тому в дослідженні використовуються статистичні дані відносної частоти у відсотках:

1) питома вага продуктивності ММС у лексиконі (відсоток кількості лексем, у яких реалізована ММС, по відношенню до загальної кількості лексем у реєстрі лексикону автора);

2) індекс покриття тексту ММС (відсоток кількості слововживань, у яких реалізована ММС, по відношенню до загальної кількості слововживань у тексті).

Статистичні дані, отримані на матеріалі текстів одного автора, порівнюються з даними інших поетів. У порівнянні морфемної довжини слова також проводиться зіставлення із закономірностями організації морфемної підсистеми української мови за даними МСФ [Клименко 1998].

За визначеними методичними засадами в § 4.1, у статистичному дослідженні чотирьох ідіостилів ставляться завдання:

- сформувати рангові списки продуктивності ММС за лексичними реєстрами чотирьох поетів і порівняти їх;
- сформувати рангові списки індексу покриття тексту ММС чотирьох поетів і порівняти їх;
- проаналізувати співвідношення відносної частоти ММС у лексиконі та тексті в чотирьох поетичних вибірках;
- визначити середню морфемну довжину слова в кожному ідіостилі й порівняти продуктивність ММС слів різної морфемної довжини в чотирьох лексичних реєстрах поетів.
- провести порівняльний аналіз моделей морфемної статистичної структури чотирьох ідіостилів і визначити, чи можна вважати цю модель стилеметричною ознакою ідіостилю.

Морфемна статистична структура ідіостилю. Реєстр моделей морфемних структур у ЧМС кожного поета формує окрему

[35] Пояснення такого обмеження подано в § 3.2.2.

морфемну систему, яка об'єднує різну кількість одиниць з різною продуктивністю в лексиконах поетів:

- Т. Шевченка (ТШ): 82 ММС – 6191 лексема [36];
- В. Стуса (ВС): 87 ММС – 8029 лексем;
- Л. Костенко (ЛК): 73 ММС – 5956 лексем;
- Лесі Українки (ЛУ): 61 ММС – 4907 лексем.

Обсяг лексичних реєстрів також формується тільки за початковими формами, які просегментовані на морфеми. Частина слів автоматично згенерованих реєстрів не ввійшла до лексичних вибірок статистичних обчислень із причини відсутності визначення морфемної будови в цих початкових формах[37].

Таблиця 4.7. Лексична продуктивність ММС у реєстрі чотирьох поетів

Питома вага (p%)у лексичному реєстрі					
MMS		**ТШ**	**ВС**	**ЛУ**	**ЛК**
RSF	*плак-а-ти*	22,98	22,42	23,74	24,58
RF	*мій-Ø*	21,80	19,22	19,65	26,46
PRSF	*с-пі-ва-ти*	18,35	19,97	21,38	15,51
RSSF	*дів-ч-ин-а*	5,22	5,57	5,73	5,51
PRF	*ні-хт-о*	5,15	4,70	4,75	4,45
Ядро		**73,50**		**75,25**	**76,51**
PRSFX	*по-див-и-ти-ся*	3,83	3,89	2,55	2,67
			Ядро 75,77		
R	*де*	2,91	2,32	2,89	2,84
RSFX	*див-и-ти-ся*	2,08	2,04	~~1,39~~	1,48
PRSSF	*на-йм-ич-к-а*	1,91	3,89	2,85	2,13
PRSS	*в-ран-ц-і*	1,83	~~1,05~~	~~0,84~~	1,21
PRS	*з-нов-у*	1,44	1,42	1,63	1,38
RS	*добр-е*	1,32	1,21	1,81	1,12
RSS	*тяж-к-о*	1,28	1,23	2,22	1,21
		Основа 90,10			**Основа 90,55**
PPRSF	*без-не-вин-н-ий*	1,03	1,23	1,43	0,86
				Основа 90,63	
RIRSF	*біл-о-рук-Ø-ий*	0,92	1,13	0,63	0,89
			Основа 90,24		
PR	*в-день*	0,84	1,00	0,92	0,82

У табл. 4.7, згідно з прийнятою в лінгвостатистиці традицією [Перебийніс 1970], визначено ядро (75 % масиву), основу (90 % масиву) системи ММС слів кожного поета за питомою вагою моделі в лексичному реєстрі. У порівняльному зіставленні еталонним ранговим списком вважається список, побудований за питомою вагою ММС у лексичному

[36] Якщо порівняти кількісні дані обсягу лексичного реєстру, кількість ММС та статистичні значення лексичної продуктивності ММС у лексичних реєстрах Т. Шевченка за табл. 4.1 (дані 2017 р.) та табл. 4.7 (дані 2019 р.), то можна зробити висновок, що зміна вихідних статистичних даних не вплинула на тенденцію формування ядра, основи й периферії системи ММС: їх формують ті самі моделі, хоча ранги ММС частково змінилися.

[37] Пояснення такого обмеження подано в § 3.2.2.

реєстрі Т.Шевченка (цей вибір зроблено умовно, тому що система поетичного мовлення Т. Шевченка була описана в попередньому параграфі).

Наведена у таблиці інформація показує, що ≈ 75 % лексикону поетичного мовлення чотирьох поетів моделюються за 5-ома спільними ММС: RSF (*плак-а-ти*), RF (*мій-Ø*), PRSF (*с-пі-ва-ти*), RSSF (*дів-ч-ин-а*), PRF (*ні-хт-о*), що складають ядро морфемної системи. Ядро морфемної системи В. Стуса формують 6 ММС: PRSFX також входить до ядра. Наступні ≈ 15% лексики (ТШ: 8 ММС; ВС: 8 ММС; ЛУ: 7 ММС; ЛК: 8 ММС) моделюється за різними ММС, серед яких визначаються:

а) спільні для 4 поетів моделі: R (*де*); PRSSF (*на-йм-ич-к-а*); PRS (*з-нов-у*); RS (*добр-е*); RSS (*тяж-к-о*);

б) моделі окремих ідіостилів:

- Т. Шевченка, Л. Костенко – PRSS (*в-ран-ц-і*);
- В. Стуса, Л. Костенко – PPRSF (*без-не-вин-н-ий*);
- Т. Шевченка, В. Стуса, Л. Костенко – RSFX (*див-и-ти-ся*).

Ці ММС разом з ядровими формують основу морфемної системи (моделюють ≈ 90% лексики), яка в кожного поета індивідуальна (табл. 4.7): Т. Шевченко – 13 ММС; В. Стуса – 15 ММС, Лесі Українки – 14 ММС; Л. Костенко – 13 ММС. Загалом, за 4-ма лексичними реєстрами 16 ММС мають ≈ p ≥ 1%. Всі інші ММС (≈ 10% лексики) є периферійними й характеризуються продуктивністю ≈ p < 1%. Такі ММС реалізуються в індивідуальній лексиці ідіостилю кожного поета. Автоматичне визначення низькопродуктивних ММС у ЧМС формує статистичний прийом вилучення з тексту індивідуальної лексики, що може з вірогідністю використовуватися в стилістичних дослідженнях.

ММС, які формують основу морфеміки чотирьох поетів (табл. 4.7), мають різну лексичну продуктивність. Розподіл питомої ваги продуктивних ММС у лексичному масиві кожного поета найкраще представити у вигляді графічного моделювання (рис. 4.5).

Графік демонструє, з якою продуктивністю (вісь *у* – p %)16 ММС (вісь *х*) морфемної системи кожного поета, покривають лексичний реєстр. Цю залежність відображає ламана кожного ідіостилю. Форма ліній визначає загальні та індивідуальні ознаки відсоткового розподілу ММС у чотирьох лексичних реєстрах. Загальні ознаки відсоткового розподілу визначають три зони лексичної продуктивності:

1) висока лексична продуктивність (24,58 – 15,51 %) – найвищі піки ламаних ліній – точки RSF, RF, PRSF;

2) середня лексична продуктивність (5,73 – 1,91 %) – різкий спад ламаних ліній до ≈ 5% (модель RSSF) – точки RSSF, PRF, PRSFX, R, RSFX, PRSSF;

3) низька лексична продуктивність (1,83 – 0,82 %) – низькі піки ламаних, які майже накладаються – точки PRSS, PRS, RS, RSS, PPRSF, RIRSF, PR.

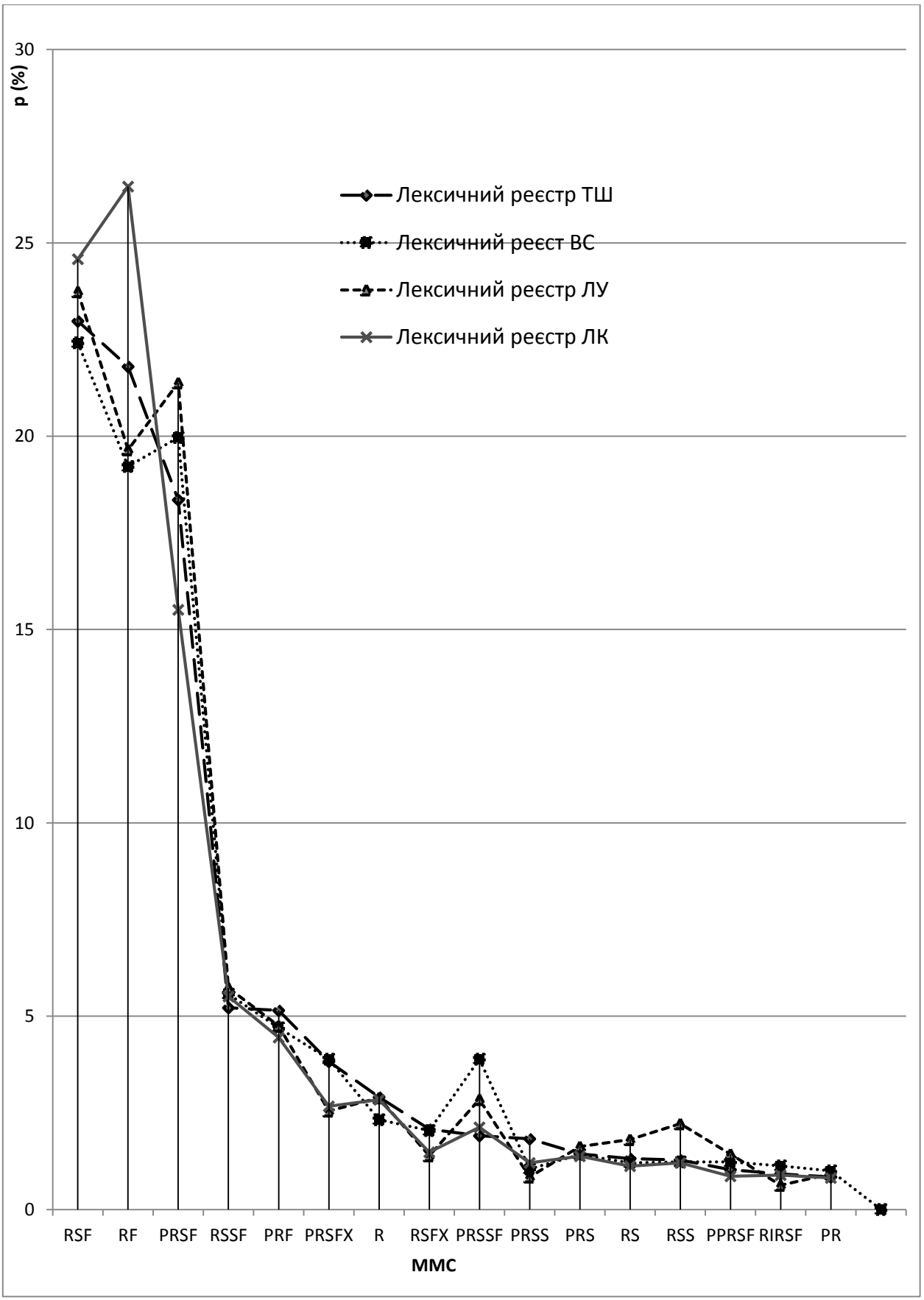


Рис. 4.5. Графік розподілу питомої ваги ММС у лексичних реєстрах чотирьох поетів

Індивідуальні ознаки відсоткового розподілу ММС характеризують насамперед морфемну статистичну структуру лексикону Л. Костенко: найвищий пік має друга модель RF (26, 46 %), а найнижчу продуктивність у 1-ій зоні має модель PRSF (15,51 %).

Ламана лексичної продуктивності Т. Шевченка (ТШ) характеризується спадом у точках RF (21,80 %), PRSF (18,35 %), а ламані ВС та ЛУ – зростанням у точці PRSF. Це свідчить, що розподіл ММС у 1-ій зоні визначає стилістичні особливості кожної вибірки. У 2-ій зоні потужною стилеметричною ознакою є модель PRSSF, яка має найвищий пік у ламаній ВС (3,89 %). У 3-ій зоні за піками ламаних вирізняється лінія ЛУ в точці RSS (2,22 %). У лінгвостатистиці розходження процентних показників досліджуваних одиниць не завжди є критерієм для висновку про значущість цих розходжень. «Дуже часто в лінгвістичних дослідженнях застосовуються проценти: дослідник визначає, що таке-то явище складає стільки-то процентів в одній сукупності і стільки-то у другій. Потім або цим використання процентних показників і завершується, або дослідник "на глазок" визначає, чи можна вважати значимими таке співвідношення процентних показників у різних вибірках. І, на жаль, у більшості випадків помиляється, приймаючи навіть незначні розходження за свідчення чогось значимого. Для запобігання таких помилок можна застосувати [...] обчислення показника Стьюдента, ...» [Перебийніс 2002].

Перевіримо істотність розходження відсоткових показників продуктивності в точках RF, PRSF, PRSSF, RSS за критерієм Стьюдента (t) для відсоткових показників (табл. 4.8) у зіставленні вибірок з вагомими відсотковими розходженнями.

Таблиця 4.8. Показник критерію Стьюдента

ММС	**Вибірки, що зіставляються**		**t**	**Критичний показник t для довірчого рівня 99%**
		кількість ступенів свободи		
RF	ЛК:ВС	13983	10,20	2,58
	ТШ:ВС	14218	3,91	
PRSF	ЛК:ЛУ	10861	7,83	
	ЛК:ТШ	12145	4,71	
PRSSF	ЛК:ВС	13983	5,87	
	ЛУ:ВС	13834	3,15	
RSS	ЛУ:ТШ	11096	3,48	

Дані показують, що в усіх 4-ох моделях показник критерію Стьюдента більший за критичний показник t > 2,58. Отже, із достовірністю 99% можна стверджувати, що розходження у відсотках для моделей RF, PRSF, PRSSF, RSS є істотним, а відсотковий показник цих моделей є статистичним параметром розмежування ідіостилів.

За визначеними методичними основами стилеметричного дослідження, морфемна статистична структура ідіостилю визначається на основі зіставлення статистичних характеристик ММС у лексиконі та тексті, тому що статистична "поведінка" ММС у лексиконі і тексті, за законом простоти і переваги, може відрізнятися, що підтверджує і модель морфемної статистичної структури ідіостилю Т. Шевченка, описана в попередньому параграфі (§ 4.2). Для визначення моделі морфемної статистичної структури чотирьох ідіостилів скористаємося зіставленням відносної частоти високопродуктивних та середньопродуктивних ММС у лексичних реєстрах та в текстах чотирьох поетів (табл. 4.9).

Таблиця 4.9. Відносна частота ММС у лексичних реєстрах та текстах 4-ох поетів

	Т. Шевченко		**В.Стус**		**Леся Українка**		**Л. Костенко**	
ММС	**р% у лексичному реєстрі**	**р% у текстах**	**р% у лексичному реєстрі**	**р% у текстах**	**р% у лексичному реєстрі**	**р% у тексті**	**р% у лексичному реєстрі**	**р% у текстах**
RSF	22,98	13,71	22,42	15,17	23,74	18,3	24,58	15,25
RF	21,8	34,43	19,22	31,67	19,65	34,83	26,46	35,08
PRSF	18,35	6,58	19,97	8,00	21,38	7,33	15,51	5,98
RSSF	5,22	1,65	5,57	2,47	5,73	2,19	5,51	1,99
PRF	5,15	2,71	4,7	2,84	4,75	2,63	4,45	2,25
PRSFX	3,83	1,21	3,89	1,44	2,55	0,67	2,67	0,89
R	2,91	30,04	2,32	26,89	2,89	24,92	2,84	28,81
RSFX	2,08	1,45	2,04	1,01	1,39	0,48	1,48	0,97
PRSSF	1,91	0,4	3,89	1,3	2,85	0,80	2,13	0,65
PRSS	1,83	0,42	1,05	0,31	0,84	0,19	1,21	0,45
PRS	1,44	0,95	1,42	1,03	1,63	0,78	1,38	0,78
RS	1,32	1,65	1,21	1,46	1,81	1,71	1,12	1,29
RSS	1,28	0,56	1,23	0,61	2,22	1,09	1,21	0,55
PPRSF	1,03	0,3	1,23	0,56	1,43	0,38	0,86	0,28
RIRSF	0,92	0,24	1,13	0,37	0,63	0,14	0,89	0,27
PR	0,84	1,12	1,00	1,21	0,92	1,08	0,82	0,66
Всього	**92,89**	**97,42**	**92,29**	**96,34**	**94,41**	**97,52**	**93,12**	**96,15**

16 ММС кожної вибірки покривають різний масив лексичного реєстру і тексту:

вибірка	% лексичного реєстру	% тексту
ТШ	92,89	97,42
ВС	92,29	96,35
ЛУ	94,44	97,52
ЛК	93,12	96,15

Зіставлення розподілу питомої ваги кожної моделі в лексичних реєстрах з індексом покриття тексту в кожній вибірці моделюємо за допомогою двох ламаних на графіках, де вісь x – р %; вісь y – 16 ММС (рис. 4.6; рис. 4.7; рис. 4.8; рис. 4.9).

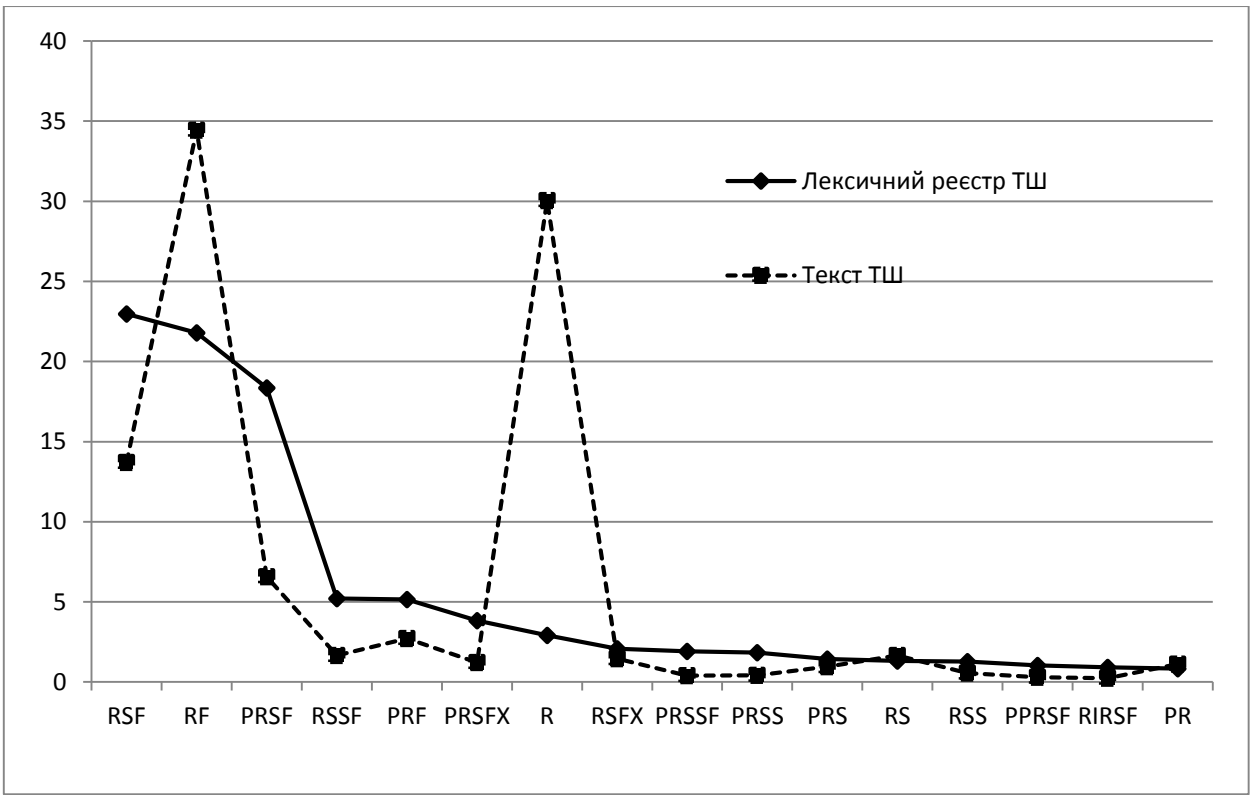


Рис. 4.6. Графік розподілу відносної частоти ММС у лексичному реєстрі та тексті Т. Шевченка

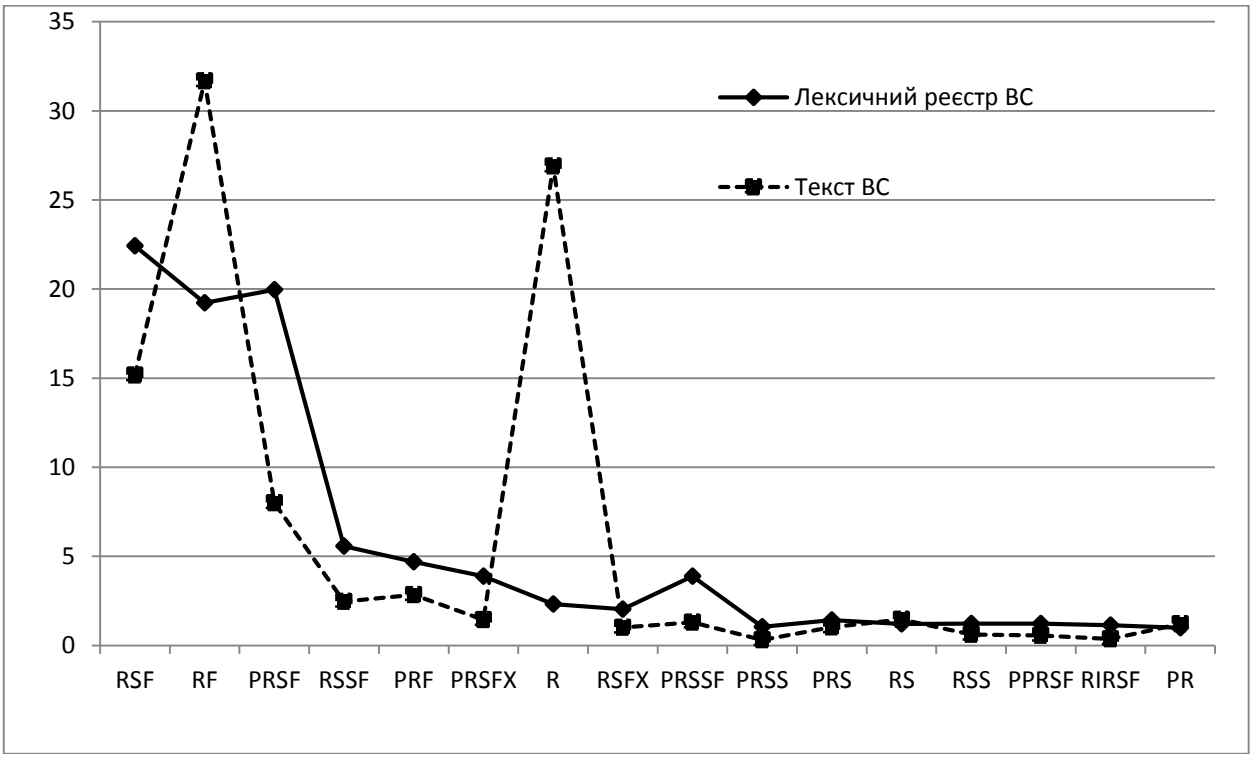


Рис. 4.7. Графік розподілу відносної частоти ММС у лексичному реєстрі та тексті В. Стуса

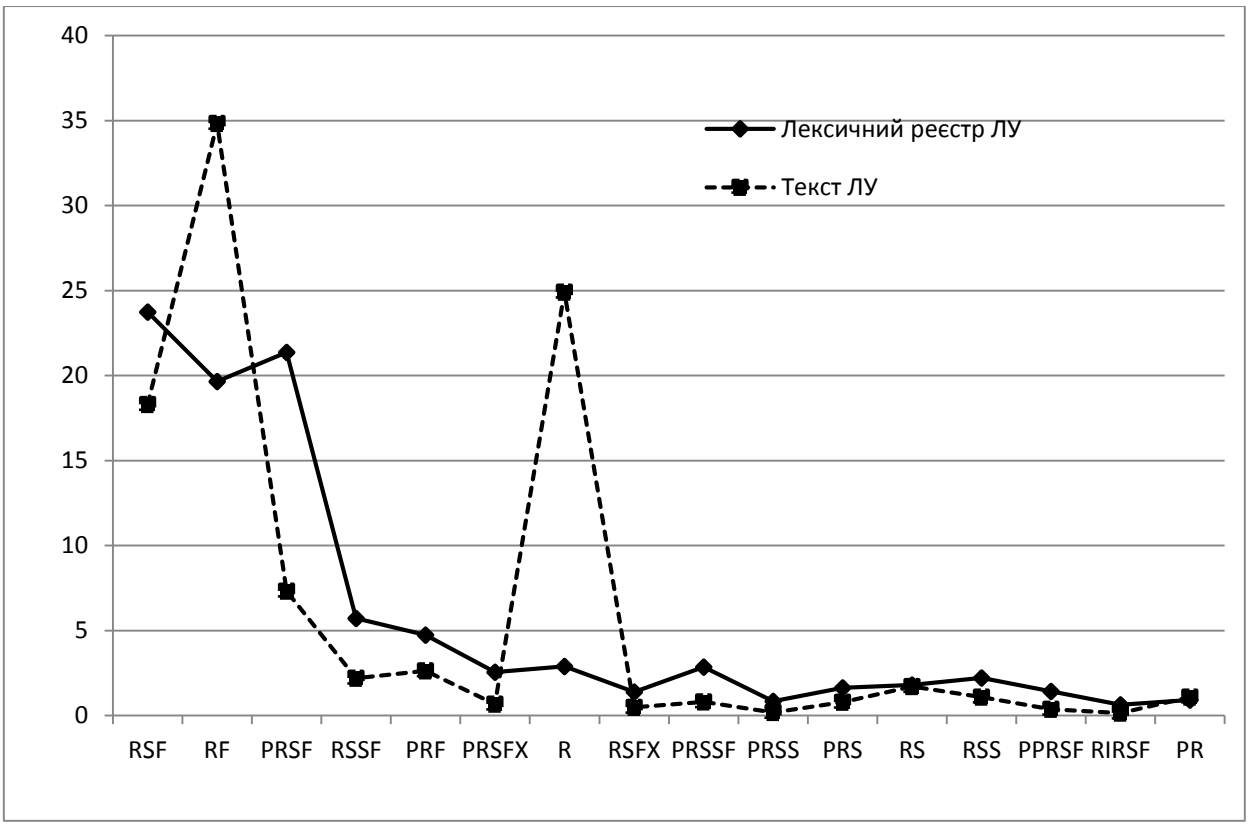


Рис. 4.8. Графік розподілу відносної частоти ММС у лексичному реєстрі та тексті Лесі Українки

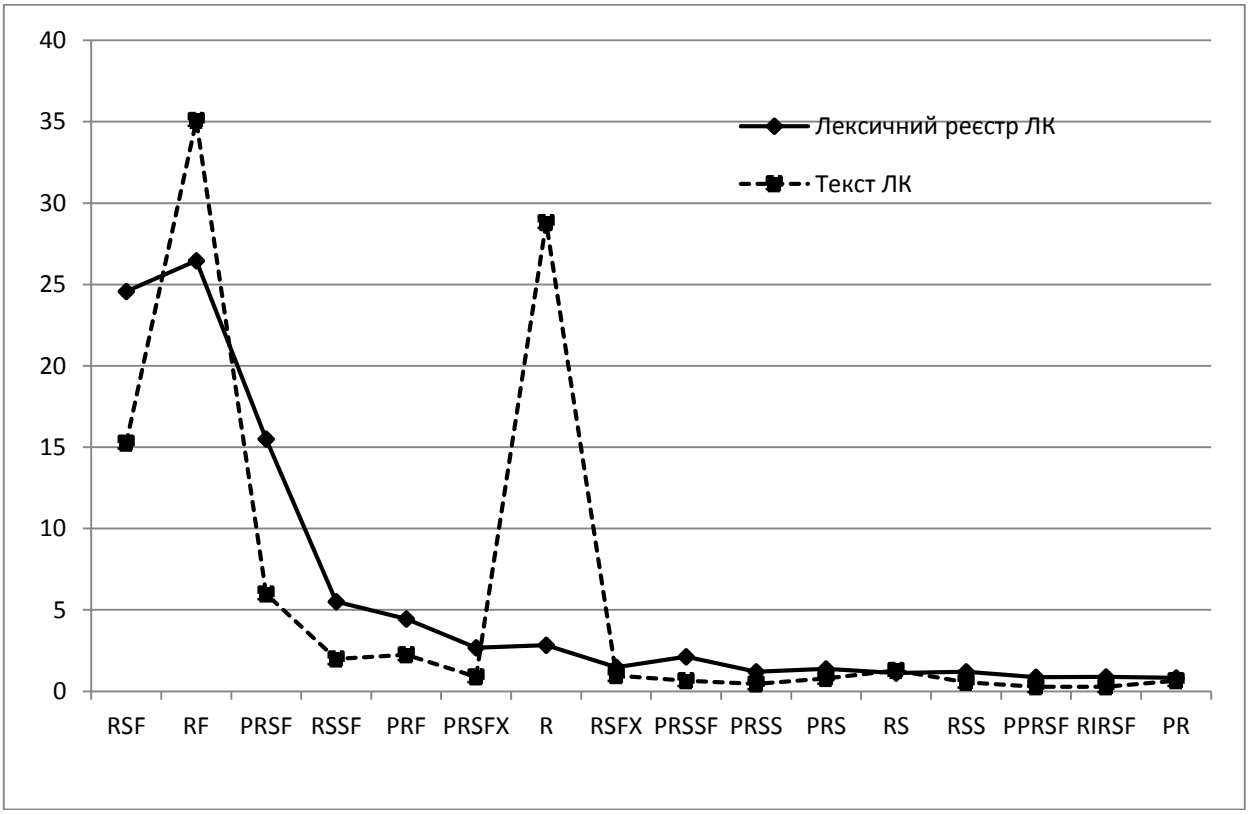


Рис. 4.9. Графік розподілу відносної частоти ММС у лексичному реєстрі та тексті Л. Костенко

Зіставлення двох ліній кожного графіка показує невідповідність продуктивності морфемних структур у мові (лексичному реєстрі) та в мовленні (тексті) й дозволяє описати особливості моделі морфемної статистичної структури ідіостилю. Ламані 4-ох графіків демонструють загальну закономірність: кожна ламана індексу покриття тексту повторює за формою ламану лексичної продуктивності, але продуктивність у тексті переважної кількості ММС нижча ніж у лексичному реєстрі, за винятком двох ММС – RF і R, що порушує загальну тенденцію до зниження індексу покриття тексту. Високий індекс покриття тексту моделі R пояснюється високою частотою в текстах сполучників, часток та прийменників, які переважно представляють цю модель у тексті, проте найвищий індекс покриття в тексті має модель RF. Піки ламаних тексту RF і R показують, що ці моделі покривають ≈ 58 – 65 % тексту кожного автора (для порівняння, відсоткова зона лексичного реєстру RF: 19,22 – 26,46 %; R: 2,32 – 2,99 %), і, очевидно, слова з такою морфемною будовою, за законами лінгвістичної статистики, визначаються як стилістично нейтральні, проте ці слова мають найменшу морфемну довжину, що є стилеметричною ознакою поетичного мовлення загалом. Це явище підтверджує необхідність уведення до моделі морфемної статистичної структури ідіостилю параметра морфемної довжини слова.

Морфемна довжина слова. Лексичні реєстри [38] поетичних вибірок включають слова з різною морфемною довжиною (табл. 4.10): ТШ, ЛУ (1 – 7 морфем); ВС, ЛК (1 – 8 морфем).

Таблиця 4.10. Співвідношення морфемної довжини слів з кількістю ММС та питомою вагою

Кількість морфем у слові	ТШ			ВС			ЛК			ЛУ			% відсоток у системі української мови
	Кількість ММS		% у лексичному реєстрі	КількістьММS		% у лексичному реєстрі	Кількість ММS		% у лексичному реєстрі	Кількість ММS		% у лексичному реєстрі	
	прості слова	складні слова		прості слова	складні слова		прості слова	складні слова		прості слова	складні слова		
1М	1		2,63	1		2,32	1		2,94	1		2,19	0,22
2М	4	1	24,63	4	1	22,77	4	1	28,93	4	1	22,70	6,87
3М	12	5	31,80	7	7	31,48	7	4	31,48	10	4	33,02	21,24
4М	12	8	30,15	11	8	29,39	12	8	27,44	11	3	31,20	34,51
5М	13	9	9,18	13	11	12,04	13	5	7,79	10	4	8,64	26,13
6М	10	5	1,15	7	7	1,41	7	5	0,96	7	4	1,06	7,67
7М		2	0,065	2	6	0,16	2	2	0,084	1	1	0,041	2,46
8М					2	0,025		2	0,034				
	52	**30**	**99,61**	**45**	**42**	**99,59**	**46**	**27**	**99,66**	**44**	**17**	**99,45**	**99,1**
Всього	**82**			**87**			**73**			**61**			

[38] З метою порівняння морфемної довжини слова в поетичних вибірках із системою української мови статистичні обчислення проводяться на матеріалі лексичних реєстрів, а не текстів.

Однакові за кількістю морфем ММС варіюють за функціональною структурою, наприклад, 4-морфемна структура (4М) може мати такі функціональні структури: RSSF – *дів-ч-ин-а*, PRSS – *в-ран-ц-і*, тому кожна кількісна морфемна модель може бути представлена різною кількістю ММС. У табл. 4.10 систематизовано дані про морфемну довжину слова, кількість ММС з однаковою кількістю морфем та відносну частоту в лексичному реєстрі.

Слова у поетичних текстах моделюються за різною кількістю ММС: Т. Шевченко (82 ММС); В. Стус (87 ММС); Л. Костенко (73 ММС), а лексика Лесі Українки представлена найменшою кількістю ММС – 61. По-різному представлена кількість ММС простих і складних слів у лексиконах поетів: найбільшу функціональну різнотипність мають ММС простих слів у лексиці Т. Шевченка (52 ММС), а ММС складних слів у лексиконах Т. Шевченка (30 ММС) та В. Стуса (42 ММС). 1-морфемні та 2-морфемні слова представлені в лексиконах чотирьох поетів однаковою кількістю ММС, що моделюють усі можливі функціональні типи одно- та двоморфемних простих і складних слів української мови: R (*де*), RR (*бо-дай*), RF (*мій-Ø*), RS (*добр-е*), RX (*де-сь*), PR (*не-хай*). Кількісний розподіл ММС слів за кількісно-морфемною (колонка – кількість морфем у слові) та функціонально-морфемною (колонки – кількість ММС) складністю показує, що зі збільшенням кількості морфем у слові кількість ММС починає варіювати. Кількість 4-морфемних ММС простих слів в усіх лексиконах майже однакова (11 – 12 ММС), а ММС складних слів найчисленніші у лексиці Т. Шевченка, В. Стуса, Л. Костенко (8 ММС). 5-морфемні ММС складних слів мають більшу функціональну варіативність у лексиці Т. Шевченка (9 ММС) та В. Стуса (11 ММС), а в лексиконах Л. Костенко (5 ММС) та Лесі Українки (4 ММС) варіативність цих моделей істотно знижується. Найменшу функціональну варіативність мають 6-морфемні слова в лексиці Лесі Українки (7 ММС складних слів, 4 ММС простих слів). 7-морфемні ММС мають низьку функціональну варіативність:

а) ММС простих слів реалізовані в словах В. Стуса (2 ММС: PRSSSSF – *з-мер-т-в-і-л-ий*; PPRSSSF – *не-до-кон-а-н-ість-Ø*), Л. Костенко (1 ММС: PPRSSFX – *при-з-вич-ај-і-ти-ся*), Лесі Українки (1 ММС: PPPRSSF – *не-до-с-пі-ва-н-ий*);

б) ММС складних слів реалізовані в словах Т. Шевченка (2 ММС: RIRSSSF – *нов-о-бр-а-н-ець-Ø*; RSIRSSF – *вел-ик-о-муч-ен-ик-Ø*), В. Стуса (6 ММС), Л. Костенко (2 ММС), Лесі Українки (1 ММС: RIPRSSF – *брат-о-в-би-в-ств-о*).

Кількісні дані табл. 4.10 показують, що ММС реалізуються в словах різної морфемної довжини, регульованої певними закономірностями кількісного обмеження. Морфемна довжина слова в лексиконах чотирьох поетів обмежена інтервалом 1 – 8 морфем:

1) прості слова в лексиці Т. Шевченка (1 – 6), В. Стуса, Л. Костенко та Лесі Українки (1 – 7);

2) складні слова в лексиці Т. Шевченка та Лесі Українки (2 – 7), а в лексиці В. Стуса та Л. Костенко (2 – 8).

Крім того, як свідчать статистичні дані табл.4.10, морфемна довжина слова по-різному впливає на лексичну продуктивність ММС у лексиконах української мови та чотирьох поетів. Закономірність впливу морфемної довжини слова на продуктивність ММС у лексичних системах чотирьох поетичних вибірок та в системі української мови демонструє графік на рис.4.10.

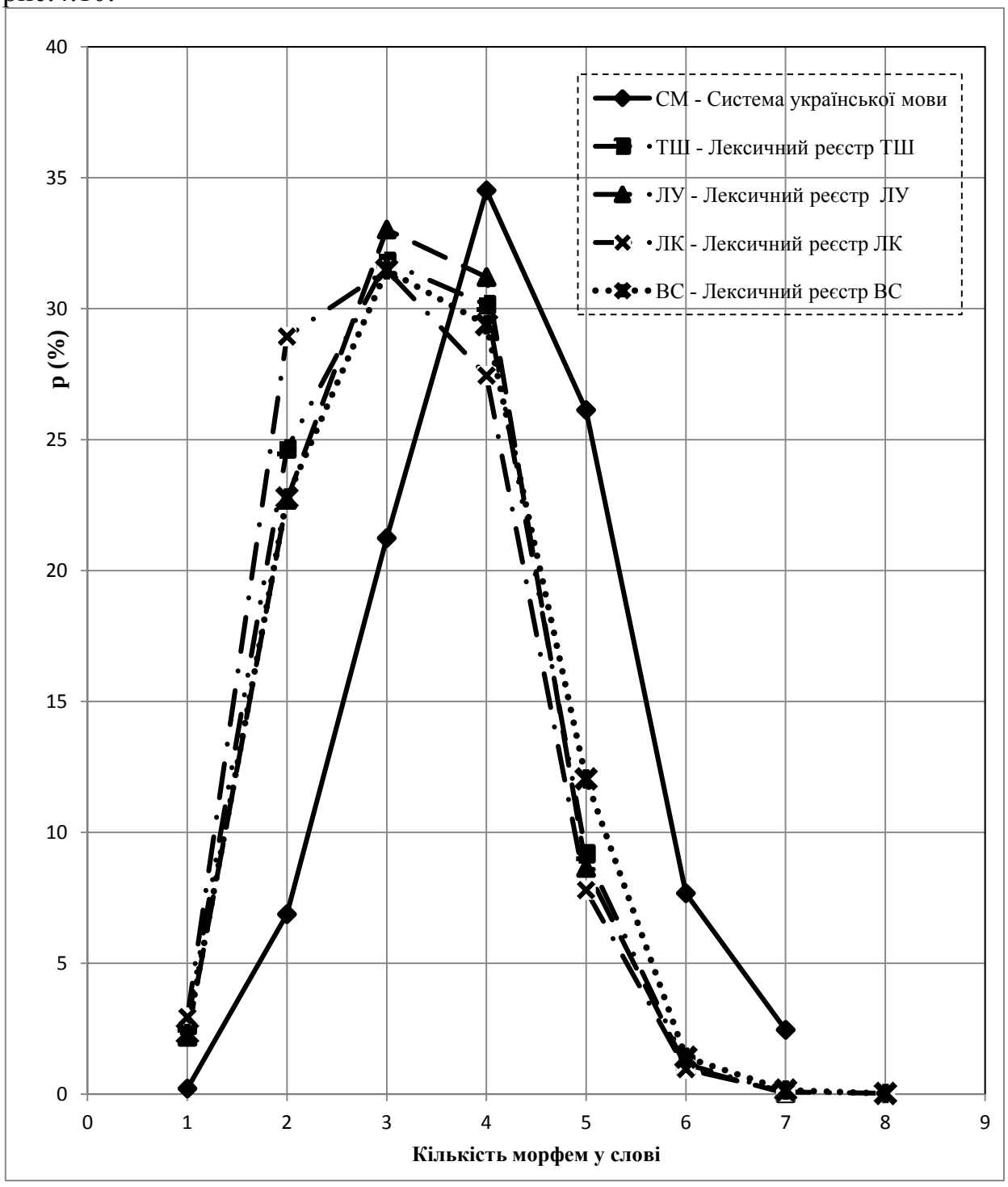


Рис. 4.10. Графік розподілу морфемної довжини слова в лексиконах чотирьох поетів та системі сучасної української мови

Вершину ламаної СМ формують 4-морфемні слова. Вершини ламаних чотирьох поетичних вибірок формують 3-морфемні слова. 5-морфемні структури, на відміну від системи мови, представляють на ламаних поетичних вибірок точку різкого зниження продуктивності. Розподіл морфемної довжини слова в лексиконах поетів найбільш подібний у вибірках ТШ та ВС: ламані майже збігаються. Подібно до них, формується ламана ЛУ, хоча вершина цієї ламаної демонструє найвищу питому вагу 3-морфемних структур (33,02 %). Найбільше відрізняється ламана ЛК, що демонструє найвищу питому вагу 2-морфемних структур (28,93 %), стрімке зниження питомої ваги (найнижчу продуктивність – 27,44 %) 4-морфемних структур.

Ламані графіка демонструють кількісне зменшення морфемної довжини слів ядрового масиву лексики поетів, порівняно із системою мови. Як уже зазначалося у попередньому параграфі, за даними МСФ середня морфемна довжина українського слова обмежується інтервалом 4 ± 1 і становить 3,9 [Клименко 1998: 150-153]. На базі чотирьох ЧС можна вирахувати середню морфемну довжину слова в лексиконі кожного поета за формулою $\bar{x} = \frac{\sum x_i \cdot n_i}{\sum n_i}$ (x_i – кількість морфем ММС, n_i – кількість слів, в якій реалізована ця ММС):

- ідіолект Т. Шевченка – 3,22;
- ідіолект В. Стуса – 3,34;
- ідіолект Лесі Українки – 3,23;
- ідіолект Ліни Костенко – 3,11.

Таким чином, середня морфемна довжина слова поетичного мовлення входить до інтервалу 4 ± 1, але поети використовують морфемно-коротші слова, що становить загальну стилеметричну ознаку поезії: середня морфемна довжина слова поетичного стилю сучасної української мови < 3,3. Крім того, показник середньої морфемної довжини слова, як свідчать дані чотирьох ідіостилів, є стилеметричною ознакою ідіостилю: поетичне мовлення кожного із досліджених авторів має свій статистичний показник середньою морфемної довжини слова.

Зіставлення статистичних даних електронних ЧМС чотирьох українських поетів в аспектах морфемної довжини слова, питомої ваги в лексичному реєстрі, індексу покриття тексту формує стилеметричну модель опису ідіостилю, яку було визначено як модель морфемної статистичної структури ідіостилю. Використання такої моделі в стилістичному дослідженні систематизує не лише специфічні морфемні явища, що формуються в результаті індивідуальних мовних особливостей поета, а всі структурні одиниці морфемного рівня системи тексту. Ця модель демонструє вищий рівень узагальнення кількісної моделі тексту, ніж модель статистичної структури лексики, тому що кількість ММС у тексті, порівняно з кількістю слів, зменшена в сотні разів. Невеликий реєстр ММС дозволяє систематизувати статистичні характеристики слів у великих за обсягом текстових вибірках.

Розподіл ММС у лексичних реєстрах та текстах здійснюється за законом Дж. Ципфа: 16 ММС 4-ох поетичних вибірок покривають ≈ 92 – 94% лексичного реєстру та ≈ 96 – 97 % тексту; 45 ММС (ЛУ), 71 ММС (ВС), 57 ММС (ЛК), 66 ММС (ТШ) покривають ≈ 6 – 8 % лексичного реєстру та ≈ 3 – 4 % тексту. Проте співвідношення кількості слів однієї ММС у лексичному реєстрі із сумою їх абсолютних частот у тексті не відповідає дії закону Дж. Ципфа, тому що розподіл слів у лексичному реєстрі і тексті здійснюється не за математичною моделлю – варіантою абсолютної частоти, а за лінгвістичною ознакою – морфемною будовою слова. І хоча закон Дж. Ципфа не діє в цьому розподілі, проте він пояснює піки росту відносної частоти одно- та двоморфемних ММС (R, RF) у тексті: високочастотні слова мають мінімальну морфемну довжину.

Зменшення середньої морфемної довжини слова в досліджуваних поетичних вибірках підтверджує думку В. Московича про вплив на це явище функції стилю: основою поетичного стилю є не предметна, а естетична функція, тому у поезії увага зосереджена на самій структурі слова: «Цим пояснюється зменшення до мінімуму в поезії кількості слів з глибиною і довжиною близькою до максимальної» [Москович 1967: 29].

Модель морфемної статичної структури ідіостилю може розглядатися як еталонна статистична модель у вивченні текстів різних стилів та ідіостилів на матеріалі однієї мови або в типологічних дослідженнях флективних мов.

В И С Н О В К И

У монографії розглянуто методи, стратегії та прийоми створення комп'ютерних інструментів для дослідження морфемної системи української мови. Результатом комп'ютерного конструювання стали електронні лексикографічні системи: автоматизована система морфемно-словотвірного аналізу (АСМСА) та інтерактивна лексикографічна система текстоорієнтованих частотних морфемних словників. Досвід створення цих систем засвідчує синтез знань і завдань трьох лінгвістичних галузей – морфемології, лексикографії та комп'ютерної лінгвістики, що дозволило визначити і обґрунтувати окрему прикладну лінгвістичну галузь – комп'ютерну морфемну лексикографію, об'єктом вивчення якої є морфемна система мови, а предметом конструювання – комп'ютерна модель лексикографічного опису цієї системи.

Концептуальна модель морфемної системи мови – базовий інфологічний конструкт у створенні електронної лексикографічної системи для користувача. На етапі софтвера створюється даталогічна лексикографічна модель, що виконує функцію систематизації лінгвістичної інформації у формі бази даних – автоматичного резидентного словника лінгвістичного процесора. Таким чином, комп'ютерна лексикографічна модель має дворівневий принцип будови й структурується на: 1) внутрішню модель бази даних, що забезпечує автоматичну роботу лексикографічного процесора; 2) зовнішню модель електронного словника – інтерфейс системи для користувача, який виступає результатом роботи лексикографічного процесора. У такому розумінні поняття "лексикографічна модель" та "лексикографічна система" використовуються ширше ніж у лексикографії, яка ставить завдання – розробити лексикографічні параметри мікро- та макроструктури паперового або електронний словника, що виступає зовнішньою моделлю лексикографічної системи. Тому, на наше переконання, комп'ютерна лексикографія в широкому розумінні тих задач, які вона виконує, є галуззю комп'ютерної лінгвістики, а не лексикографії.

Конструювання АСМСА базувалося на розробленні концептуальної та даталогічної моделей цієї системи, що формалізують опис об'єктів та процесів морфемної системи української мови, ґрунтуючись на методологічних узагальненнях теоретичних і прикладних ідей сучасної структурної лінгвістики. Вихідною й основною моделлю в АСМСА є даталогічна модель морфемної структури слова – функціонально-кількісно-графемна формула, що в поєднанні з програмними процедурами забезпечує проведення автоматичної морфемної сегментації слова. Ізоморфність цієї моделі з об'єктом-оригіналом була перевірена експериментально в процесі апробації системи автоматичного морфемного сегментування слів (початкових форм) у Корпусі української мови.

Функції АСМСА, обсяг лінгвістичного матеріалу, багатоаспектна систематизація морфемних та словотвірних одиниць, явищ і процесів визначають наукову та практичну значущість цієї системи:

(1) АСМСА репрезентує великий лексичний матеріал: обсяг реєстру слів із визначеною морфемною будовою становить ≈ 200 тис. одиниць, що дозволяє говорити про повноту лінгвістичного опису морфемної і словотвірної систем української мови.

(2) АСМСА має складну структуру даних, що репрезентує великий обсяг лінгвістичної інформації про організацію морфемної системи української мови:

- морфемну будову слів української мови: БД морфемних структур слів ≈ 200 тис. слів;
- аломорфію коренів української мови: БД аломорфічних коренів (≈ 2500 коренів);
- омонімію коренів української мови: БД омонімічних коренів (≈ 3100 коренів);
- аломорфію та омонімію афіксів української мови: БД афіксальних морфем (у процесі укладання);
- відношення словотвірної похідності між словами української мови: БД мотивувальних слів(у процесі укладання);
- морфонологічні процеси в мотивованому слові окремої словотвірної пари: БД морфонологічних процесів (у процесі укладання);
- лексичні значення слів української мови: компілятивна БД лексичних тлумачень;
- кількісно-морфемну характеристику спільнокореневої лексики української мови: БД вибірок спільнокореневих слів (у процесі верифікації);
- словотвірні гнізда української мови: БД словотвірних гнізд (у процесі укладання).

(3) У системі АСМСА можна проводити різноманітні автоматичні класифікаційні операції з урахуванням частиномовного обмеження лексичної вибірки, зокрема: групувати лексику української мови за функціональними типами морфем, за заданою конкретною морфемою, за моделлю морфемної структури слова.

(4) Класифікаційні можливості, обсяг і параметри систематизації різнопланової лінгвістичної інформації визначають АСМСА базою знань із морфемології та дериватології української мови. Ця система є ефективним і раціональним комп’ютерним інструментом лінгвістичних досліджень, що активно використовується в наукових розвідках та в навчальному процесі в Інституті філології Київського національного університету імені Тараса Шевченка.

(5) АСМСА є джерельною базою для автоматичного укладання різноманітних паперових та електронних словників із морфеміки і словотвору української мови.

(6) БД морфемних структур АСМСА використовується в автоматичному морфемному аналізу слів у текстах української мови.

Методика моделювання, використана в конструюванні АСМСА, була покладена в основу побудови концептуальної лексикографічної моделі морфемного електронного словника. Наукова новизна цієї лексикографічної моделі полягає в тому, що в ній закладено ідею комплексного лексикографічного опису морфеміки української мови на всіх рівнях її структури (синтагматичному, парадигматичному, ієрархічному) з урахуванням електронної форми представлення цього опису.

Такий підхід визначив інтегральний принцип побудови словникової статті морфеми як інваріантно-варіантного конструкта, що систематизує великий масив лінгвістичних даних і спричинює складність сприйняття великого обсягу лінгвістичної інформації користувачем. Із цієї причини в лексикографічній моделі аспект інтегральності опису було перенесено з мікроструктури на макроструктуру словника, що характеризується поєднанням і взаємодоповненням морфемних словників різних типів:

- словника морфів, диференційованого за функціональними типами морфів, із представленням повного реєстру слів, у яких реалізований кожен морф;
- словника морфемних структур слів, типізованих в окремі групи за символьними моделями морфемних структур;
- тлумачно-морфонологічного словника морфем, у якому морфема представлена як інваріантно-варіантний конструкт.

Інтегральна лексикографічна модель електронного морфемного словника української мови є першою спробою застосування інтегрального принципу побудови нелексичного словника в українській комп'ютерній лексикографії. Реалізація такої інтегральної лексикографічної моделі можлива тільки в інтерактивних електронних лексикографічних системах, у яких вирішується проблема протиріччя між великим обсягом словника та глибиною опису словникових одиниць, з одного боку, і простотою та зручністю користування цим словником, з іншого. Інтегральна лексикографічна модель була апробована при конструюванні лексикографічної системи електронних частотних морфемних словників у Корпусі української мові.

Морфемний автоматичний аналіз у Корпусі української мови розроблений не за традиційним принципом анотації морфемних одиниць у тексті, а за новаторським лексикографічним принципом представлення результатів аналізу в електронних частотних морфемних словниках: тексти КУМ виступають тільки матеріалом для автоматичного укладання частотних морфемних словників. Комп'ютерне лексикографічне моделювання морфемної системи українськомовних текстів свідчить, що з метою вилучення з тексту реляційно-функціональних характеристик морфемних одиниць не обов'язково проводити морфемну або словотвірну анотацію текстів. Використана в комп'ютерному лексикографічному моделюванні методика автоматичного морфемного аналізу початкових форм (лем) текстів не знижує ефективності та оперативності класифікаційно-пошукових опцій створеної текстоорієнтованої системи морфемного аналізу тексту, а також не

применшує значущості результатів лінгвістичного дослідження на отриманому матеріалі, а навпаки, за рахунок систематизації та різноаспектної класифікації морфемної інформації підвищує експланаторність лінгвістичного дослідження.

Даталогічна лексикографічна модель БД ЧМС – це перша і єдина в українській комп'ютерній лінгвістиці комп'ютерна динамічна експериментальна текстоорієнтована аналітична модель, на базі якої лінгвістичний процесор в автоматичному режимі здійснює:

- ефективний та релевантний автоматичний морфемний аналіз лем за згенерованими лексичними реєстрами великих текстових вибірок;
- обчислення статистичних характеристик морфем та моделей морфемних структур слів у текстових вибірках;
- укладання реєстрів позиційно-функціональних типів морфем;
- укладання лексичних реєстрів за типами морфемних одиниць, визначених у словах;
- укладання конкордансів до слів, у яких визначена певна морфемна одиниця.

Даталогічна модель БД частотних морфемних словників має велике практичне значення: вона може застосовуватися для конструювання електронних морфемних словників за будь-якою текстовою вибіркою КУМ, а також може використовуватися як еталонна в задачах комп'ютерної статистичної лексикографії різних мов, що визначає пріоритетним методологічний підхід обчислення статистичних характеристик морфемних одиниць у тексті на базі морфемної будови початкових форм (лем) текстових слововживань. Нині ця модель апробована на двадцяти текстових вибірках, у результаті чого здійснено автоматичне лексикографічне конструювання двадцяти електронних лексикографічних систем ЧМС [ЧСКУМ 2018], інтерфейс яких створено за визначеними принципами інтегральної лексикографічної моделі.

Електронні лексикографічні системи ЧМС реалізовують лише два типи інтерактивних словників: словник морфем та словник морфструктур. Інтерфейс ЧМС, побудований з урахуванням трьох функціональних зон, забезпечує інтерактивну навігацію та різноманітні автоматичні класифікації морфемних одиниць. На нинішньому етапі ставиться завдання розбудови інтерфейсу з метою покращення діалогового режиму з користувачем: збільшити кількість пошукових опцій лінгвістичної інформації, релевантної до прогностичних запитів лінгвіста-дослідника, зокрема: додати пошукову опцію за конкретно введеною морфемою, додати опції автоматичної суми виділених частот та інші операції, які підвищать мобільність та дослідницькі можливості лексикографічної системи.

Вивчення організації тексту на морфемному рівні за допомогою морфемного аналізу початкових форм слів, а не слововживань, виправданий практикою створення частотних морфемних словників у Корпусі української мови. Використання методики автоматичного морфемного аналізу лексики

тексту за їх початковими формами в Корпусі української мови демонструє ефективність й оптимальність цієї методики: ≈ 200 тис. одиниць морфемної бази даних АСМСА дозволяють отримати інформацію про морфемну організацію мільйонів текстових слововживань з ілюстрацією контекстів їх вживання. Проте застосована методика має свої недоліки: 1) неможливість автоматично зняти омонімію слів однієї частини мови за умови різної морфемної сегментації омографів; 2) неможливість визначити морфемну структуру в словах, які не були зафіксовані словниками української мови або є помилками. Вибірка таких слів, з одного боку, засвідчує недолік роботи автоматичного морфемного аналізу, що вимагає редагування МБД АСМСА, а з іншого, є надзвичайно важливим матеріалом для стилістичних досліджень. Уважне вивчення "необробленої" лексики дозволяє зробити висновок, що даталогічна модель, використана в конструюванні електронних ЧМС, продукує експериментальну статистичну процедуру формування вибірки низькочастотної лексики, яка може вважатися гіпотетичною вибіркою стилістично маркованих слів.

Лексикографічна система ЧМС – перша в українській комп'ютерній лексикографії система текстоорієнтованих електронних морфемних словників. Ця системи відкриває перед філологами нові можливості та перспективи в дослідженні морфемної будови слова в українськомовних текстах, зокрема вивчення морфемної будови неологізмів, морфемної довжини та глибини слів, валентності різних типів морфем, статистичної "поведінки" морфем та морфемних структур слів у різних текстових вибірках. Корпусноорієнтовані дослідження, проведені на основі електронних морфемних словників, матимуть велике теоретичне значення, тому що ці словники систематизують нові лінгвістичні дані про функціонування морфемних одиниць в українськомовних текстах різного функціонально-комунікативного спрямування.

Автоматичне обчислення та систематизація статистичних характеристик морфемних і лексичних одиниць у частотних морфемних словниках, укладених за текстовими вибірками Корпусу української мови, відкрили широкі можливості для глибокого стилеметричного дослідження, у якому було теоретично обґрунтовано й апробовано нову метричну модель – модель морфемної статистичної структури стилю. Ця модель може розглядатися як типологічна статистична модель у вивченні текстів різних стилів та ідіостилів і демонструє вищий рівень узагальнення кількісної моделі тексту, ніж модель статистичної структури лексики, тому що кількість моделей морфемних структур слів у тексті, порівняно з кількістю слів, зменшена у сотні разів.

Дослідження кількісної організації морфемної будови слів та статистичних параметрів морфемного рівня організації поетичного тексту доводить, що кількісно-структурні та статистичні характеристики морфструктур, які формують відносно невеликий інвентар одиниць, виявляють закономірності будови тексту ідіостилю на морфемному рівні

його організації, а статистична "поведінка" морфемних структур слів є параметром авторського стилю. Розроблені теоретичні принципи статистичного аналізу морфемної системи стилю та проведені стилеметричні дослідження дозволяють зробити висновок про формування в українському мовознавстві нового розділу статистичної стилістики – морфемної стилеметрії.

Інструментально-процедурна методика та формалізація моделі функціонування морфемної системи української мови й українськомовного тексту, закладені в комп'ютерному лексикографічному моделюванні, демонструють когнітивний синтез декларативних та процедурних лінгвістичних знань у галузі морфемології та дериватології української мови й представляють розвиток практики і теорії комп'ютерної лінгвістики в аспекті ідеології "машини знань" [Баранов 2001: 311]. Створені електронні лінгвістичні системи – АСМСА та ЧМС – є інтелектуальними експертними програмами, що імітують діяльність лінгвіста-дослідника і продукують експертні оцінки, автоматично визначених, морфемних одиниць.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Алексеенко 2004: Алексеенко Л. А., Дарчук Н.П., Зубань О.М., Сорокин В.М. Параметризированная база данных поэтической речи как источник и инструмент филологических студий. *Международная конференция "Прикладная лингвистика без границ". Материалы конференции*. Санкт- Петербург, 2004. С. 80 – 87.
2. Алексієнко 2001: Алексієнко Л. А., Дарчук Н.П., Зубань О.М. Методика створення автоматизованої системи морфемно-словотвірного аналізу (АСМСА) слів української мови. *Наукова спадщина професора С.В. Семчинського і сучасна філологія: Збірник наукових праць: У 2 ч.* / Упоряд. В. Ф. Чемес. Київ: Видавничо-поліграфічний центр "Київський університет", 2001. Ч. 1. С. 38–49.
3. Алексієнко 2004а: Алексієнко Л. А., Дарчук Н.П., Зубань О.М., Сорокін В.М Автоматизована система морфемно-словотвірного аналізу як інструмент лінгвістичних досліджень. *Мова. Науково-теоретичний часопис. Матеріали міжнародної науково-практичної конференції "Проблеми прикладної лінгвістики"*. Одеса, 2004. С. 19–26.
4. Андреев 1963: Андреев Н. Д. Алгоритмы статистико-комбинаторного моделирования морфологии, синтаксиса, словообразования и семантики. *Материалы по математической лингвистике и машинному переводу. Сб. 2*. Ленинград: Изд-во ЛГУ, 1963. С. 3–45.
5. Андреев 1995: Андреев Н. Д. Статистико-комбинаторные методы в теоретическом и прикладном языковедении. Ленинград: Наука,1967. 403 с.
6. Апресян 1990: Апресян Ю. Д. Формальная модель языка и представление лексикографических знаний. *Вопросы языкознания*. Москва, 1990. № 6. С. 123–139.
7. Апресян 1995: Апресян Ю.Д. Избранные труды: в 2 т. Том II. Интегральное описание языка и системная лексикография. Москва: Школа «Языки русской культуры», 1995. 767 с
8. АРСУН 2013: Активні ресурси сучасної української номінації: Ідеографічний словник нової лексики / Відп. ред. Є.А. Карпіловська. Київ: ТОВ «КММ», 2013. 416 с.
9. Б. де Куртене 1963: Бодуен де Куртене И. А. Введение в языкознание. Избранные труды: В 2-х т. Москва: Изд-во Акад. наук СССР, 1963. Т.2. С. 246–294.
10. Балли 1961: Балли Ш. Французская стилистика. Москва: Издательство "Иностранной литературы", 1961. 394 с.

11. Баранник 1958: Баранник Д.Х. Стилістична роль префіксальних утворень у діалозі української класичної п’єси. *Наукові записки Дніпропетровського університету*. Дніпропетровськ, 1958. Т. 68, Вип. 16. С. 11–17.

12. Баранник 1961: Баранник Д.Х. Стилістичне використання словотвору в п’єсах І.К.Карпенка-Карого, М.Л.Кропивницького і М.П.Старицького. *Наукові записки Дніпропетровського університету*. Дніпропетровськ, 1961. Т. 65, Вип. XII. С. 79–83.

13. Баранов 2001: Баранов А.Н. Введение в прикладную лингвистику: Учебное пособие. Москва: Эдиториал УРСС, 2001. 360 с.

14. Белоногов 1983: Белоногов Г.Г., Калинин Ю.П., Поздняк М.П.,Хорощилов А.А., Яфаева Г.М. Алгоритм многоступенчатого морфологического анализа русских слов. *Научно-техническая информация*. Серия 2., 1983. №1. С. 6–10.

15. Бирюкова 1975: Бирюкова Л. П. О функциональном подходе к морфемному анализу. *Актуальные проблемы русского словообразования. Учен. зап. Ташкент. ун-та.* Ташкент, 1975. Т. 143. С. 295.

16. Білодід 1954: Білодід І.К. Стилістичне використання засобів словотвору. *Українська мова в школі*. Київ, 1954. №. 3. С. 24–31.

17. Блумфилд 1968: Блумфилд Л. Язык. Москва: Прогрес, 1968. 607 с.

18. Бобкова 2014: Бобкова Т.В. Корпус текстів: основні аспекти визначення. *Науковий вісник кафедри ЮНЕСКО Київського національного університету. Філологія, педагогіка, психологія*. Київ, 2014. Вип. 29. С. 11–20.

19. Богданов 1980: Богданов С.И. Семантика морфемы и способы её определения в современной морфологии. *Вести ЛГУ*. Ленинград, 1980. №14. С. 82–86.

20. Богданов 1997: Богданов С. И. Форма слова и морфологическая форма. Санкт-Петербург: Изд. Санкт-Петербургского университета, 1997. 245 с.

21. Бондаренко 1974: Бондаренко Т.Ф. Статистико-комбінаторне формування словозмінних типів української мови. *Структура мови і статистика мовлення*. Київ: Наукова думка,1974. С.112-121.

22. БрУК 2018: Браунський корпус української мови. URL: https://r2u.org.ua/corpus (дата звернення 8.10.2018).

23. Вандриес 1964: Вандриес Ж. Язык (извлечения).История языкознания 19–20 веков в очерках и извлечениях: В 2-х ч. / Под ред. В.А.Звегинцева. Москва: Просвящение, 1964. Ч.1. С. 440–459.

24. Вахек 1964: Вахек И. Лингвистический словарь пражской школы. Москва: Прогресс, 1964. 350 с.

25. Виноградова 1984: Виноградова В.Н. Стилистический аспект русского словообразования. Москва: Наука, 1984. 184 с.

26. Винокур 1946: Винокур Г.О. Заметки по русскому словообразованию. Изв. АН СССР. Отделение литературы и языка. Москва, 1946. Т.5, Вып. 4. С. 315–331.

27. ВСНТ 2018: Відкритий словник новітніх термінів. URL: http://www.mova.info/wordlist.aspx?l1=179 (дата звернення 24.11.2018).

28. ВСС 1974: Вопросы статистической стилистики. Сб. ст. Киев, 1974. 329 с.

29. ВТССУМ 2005: Великий тлумачний словник сучасної української мови: 250000 / уклад. та голов. ред. В. Т. Бусел. Київ; Ірпінь: Перун, 2005. 1728 с.

30. Гайна 2005: Гайна Г.А. Основи проектування баз даних: Навчальний посібник. Київ: КНУБА, 2005. 204 с.

31. Гапьперин 1974: Гапьперин И. Р. Информативность единиц языка. Москва: Высшая школа, 1974. 174 с.

32. Герд 1973: Герд А.С. Структурные типы слов в современном русском языке. *Исследования по грамматике русского языка. V (Учен. Зап. Ленингр. Ун-та)*. Ленинград, 1973. № 375, Сер. Филол. Наук., Вып 77. С.56−65.

33. Герд 1983: Герд А. С. Семантика морфемы: значение или значимость (valeur)? *Структурная и прикладная лингвистика.* Ленинград, 1983. Вып. 1. С. 47−53.

34. Глисон 1961: Глисон Г. Введение в дескриптивную лингвистику. Москва:Изд-во иностр. лит., 1961. 486 с.

35. ГЛОСА 2003: Англо-український словник „ГЛОСА” / В. І. Перебийніс та ін. Київ, 2003. URL: http://www.mova.info/Page3.aspx?l1=63 (дата звернення 13.05.2018).

36. Головащук 1989: Головащук С.І. Словник-довідник з правопису та слововживань. Київ: Наукова думка, 1989. 832 с.

37. Головин 1977: Головин Б.Н. Замечания к теории словообразования. *Ученые записки Горьковского университета*. Горький, 1977. Вып. 76. С. 3–41.

38. Голубкова 2014: Голубкова Е.Е. Новый облик современной лексикографии. Будет ли существовать словарь через 20 лет? *Вестник Московского государственного лингвистического университета.*

Серия: Гуманитарные науки. Москва: Изд-во МГЛУ, 2014. № 20 (706). С.73–80.

39. Горпинич 1999: Горпинич В. О. Сучасна українська літературна мова. Морфеміка. Словотвір. Морфонологія:: навч. посіб. для студ. філол. спец. вищ. закл. освіти. Київ: Вища шк., 1999. 206 с.

40. Горский 1961: Горский Д.П. Вопросы абстракции и образования понятий. Москва: Изд-во АН СССР, 1961. 352 с

41. ГРАК 2018: Генеральний регіонально анотований корпус української мови. URL: http://uacorpus.org/(дата звернення 20.10.2018) .

42. Гринберг 1963: Гринберг Дж.. Квантитативный подход к морфологической типологии языков. *Новое в лингвистике*. Москва: Изд-во иностр. лит., 1963. Вып. 3. С. 60–94.

43. Гришина 2009: Гришина Е.А., Иткин И.Б., Ляшевская О.Н. и др. О задачах и методах словообразовательной разметки в корпусе текстов. *Полярный Вестник.* 2009. № 12. С. 5–25.

44. Ґрещук 1995: Ґрещук В. Український відприкметниковий словотвір. Івано-Франківськ: Плай, 1995. 206 с.

45. Д.-Кульчицька 2005: Демська-Кульчицька О. Основи національного корпусу української мови. Київ, 2005. 218 с.

46. Данилюк 2013: Данилюк І. Корпус текстів для вивчення граматичної службовості. *Лінгвістичні студії : зб. наук. праць*. Донецьк: ДонНУ, 2013. Вип. 26. С. 224–229.

47. Дарчук 2008: Дарчук Н. П. Комп’ютерна лінгвістика (автоматичне опрацювання тексту): підручник. Київ: Видавничо-поліграфічний центр “Київський університет”, 2008. 351 с.

48. Дарчук 2010: Дарчук Н. *Корпус украинского языка.* Prace Filologiczne. Warszawa, 2010. t. LXIII, S.99–108.

49. Дарчук 2010а: Дарчук Н. П. Дослідницький корпус української мови: основні засади і перспективи. *Вісник Київського національного університету імені Тараса Шевченка.* Серія: Літературознавство. Мовознавство. Фольклористика. Київ: Видавничо-поліграфічний центр “Київський університет”, 2010. № 21. С. 45–49.

50. Дарчук 2013: Дарчук Н. П. Комп’ютерне анотування тексту: результати і перспективи: монографія. К: Освіта України, 2013. 543 с.

51. Дарчук 2016: Дарчук Н., Зубань О., Лангенбах М., Ходаківська Я. АГАТ-семантика: семантична розмітка Корпусу української мови. *Українське мовознавство*. Київ: Видавничо-поліграфічний центр “Київський університет”, 2016. Вип. 1 (46). С. 3–10.

52. Демська 2005: Демська-Кульчицька О. Основи національного корпусу української мови. Київ: Інститут української мови НАНУ, 2005. 219 с.

53. Демська 2009. Демська О. М. Два аспекти лексикографії: місце у системі мовознавчих дисциплін і структура. Магістеріум. Київ: НаУКМА, 2009. Вип. 37 : Мовознавчі студії. С. 18-23.

54. Дубичинский 2000: Дубичинский В. В. Основные принципы неографии. *«Słowa, słowa, słowa»... w komunikacji językowej* / Pod red. M. Grabskiej. Gdańsk, 2000. S. 61–68.

55. ЕГСУЛМ 2018: Електронний граматичний словник української літературної мови (словозміна). URL: http://www.mova.info/Page.aspx?l1=222 (дата звернення 24. 10.2018)

56. Ельмслев 1960: Ельмслев Л. Пролегомены к теории языка. *Новое в лингвистике.* Москва:Изд-во иностр. лит.,1960. Вып.1. С. 264–390.

57. ЕСМТШ 2017: Електронний словник мови Тараса Шевченка. URL: http://www.mova.info/cfqsh_2.aspx (дата звернення 15.03.2017).

58. Ефремова 1968: Ефремова Т.Ф. Из наблюдений над структурой русского языка на уровне морфем. *Семантические и фонологические проблемы прикладной лингвистики*. Москва: Изд-во МГУ, 1968. С. 45–56.

59. Ефремова 1970: Ефремова Т. Ф. Опыт описания современного русского языка на уровне морфов: автореферат дис. канд. филол. наук. / Московский университет. Москва, 1970. – 18 с.

60. Ефремова 1996: Ефремова Т. Ф. Толковый словарь словообразовательных единиц русского языка. Москва : Рус. яз., 1996. 639 с

61. Ефремова 2001: Ефремова Т. Ф. Новый словарь русского языка: толково-образовательный. В 2-х т. Москва: Рус. яз., 2001. Т.1, 1210 с., Т.2, 1084 с.

62. Жуковська 2013: Жуковська В.В. Вступ до корпусної лінгвістики: навчальний посібник. Житомир: Вид-во ЖДУ ім. І. Франка, 2013. 142 с.

63. Зализняк 1977: Зализняк А.А. Грамматический словарь русского языка. Москва: Русский язык, 1977. 880 с.

64. Засорина 1971: Засорина Л. Н. Введение в структурную лингвистику. Москва: Высшая школа, 1971. 319 с.

65. Земская 1973: Земская Е.А. Современный русский язык. Словообразование. Москва: Просвящение, 1973. 325 с.

66. Зубань 1997: Зубань О. М. Методологічні питання суфіксальної парадигматики українського дієслова. *Українське мовознавство*. Київ:

Видавничо-поліграфічний центр “Київський університет”, 1997. Вип. 21. С.16 –25.

67. Зубань 1997а: Зубань О. М. Морфонологічний аспект суфіксальної парадигматики українського дієслова. *Українське мовознавство*. Київ: Видавничо-поліграфічний центр “Київський університет”, 1997. Вип. 21. С.73 –81.

68. Зубань 1997б: Зубань О. М. Типи морфемних структур суфіксальних послідовностей у дієслівних основах української мови. *Вісник Київського університету. Літературознавство. Мовознавство. Фольклористика.* Київ: Видавничо-поліграфічний центр “Київський університет”, 1997. Вип. 5. С. 37–40.

69. Зубань 1998: Зубань О.М. Морфеміка суфіксальної зони українського дієслова: автореферат дисертації на здобуття наукового ступеня кандидата філологічних наук / Київський університет імені Тараса Шевченка. Київ, 1998. 18 с.

70. Зубань 2000: Зубань О. М. Статус конфікса у системі української мови. *Слов’янські мови і сучасний світ*. Київ: Видавничо-поліграфічний центр “Київський університет”, 2000. С. 115–121.

71. Зубань 2000а: Зубань О. М. Морфеміка як рівень мовної стратифікації та як предмет вивчення у системі мовознавчих дисциплінє *Українське мовознавство*. Київ: Видавничо-поліграфічний центр “Київський університет”, 2000. Вип. 24. С. 54–60.

72. Зубань 2001: Зубань О. М. Створення морфемної бази даних: принципи опису морфеми як інваріантної одиниці. *Українське мовознавство*. Київ: Видавничо-поліграфічний центр “Київський університет”, 2001. Вип. 23. С. 52–64.

73. Зубань 2006: Зубань О. Параметризована база даних як інструмент дослідження корпусу текстів. *Лексикографічний бюлетень*: Зб. наук. пр. Київ: Ін-т української мови НАН України, 2006. Вип. 13. С. 37–43.

74. Зубань 2014: Зубань О. Особливості морфемної будови слів у поетичних текстах Т. Шевченка (на матеріалі Корпусу української мови). *Українське мовознавство*. Київ: Видавничо-поліграфічний центр “Київський університет”, 2014. № 44/1. С. 123–133.

75. Зубань 2014а: Зубань О. Стилеметричні ознаки морфемних структур слів у поетичному мовленні Т. Шевченка (на матеріалі Корпусу української мови). *Мовні і концептуальні картини світу*. Київ: Видавничо-поліграфічний центр “Київський університет”, 2014. Вип. 48. С. 165–179.

76. Зубань 2015: Зубань О. Електронні частотні морфемні словники в Корпусі української мови. *Науковий вісник Східноєвропейського*

національного університету імені Лесі Українки, Серія: Філологічні науки. Луцьк, 2015. № 3 (304). С. 315–320.

77. Зубань 2016: Зубань О. Частотні морфемні словники в Корпусі української мови – джерело стилеметричних досліджень. *Acta Universitatis Palackianae Olomucensis Philologica 104 – 2016: UCRAINICA VII: Současná ukrajinistika Problémy jazyka, literatury a kultury*. Olomouc, 2016. S. 224–231

78. Зубань 2016а: Зубань О. М. Електронні словники у Корпусі української мови: параметри пошуку та систематизації мовних одиниць. *Мовні і концептуальні картини світу*. Київ: Видавничо-поліграфічний центр "Київський університет", 2016. Вип. 54.С. 190–201.

79. Зубань 2017: Зубань О. М. Електронна лексикографічна система "Морфограф": теоретичні засади та методика конструювання (проект) // *Людина. Комп'ютер. Комунікація: збірник наукових праць*. Львів: Видавництво Львівської політехніки, 2017. С. 60 -68.

80. Зубань 2017а: Зубань О. Н. Задачи и методы автоматического морфемного анализа в Корпусе украинского языка. *Актуальные проблемы современной прикладной лингвистики: Сб. науч. ст., посвященных 80-летию доктора филол. наук, профессора академика Международной академии информатизации А. В. Зубова*. Минск, 2017. С. 265– 273.

81. Зубань 2018: Зубань О. Корпус української мови – комп'ютерна експертна система лінгвістичного аналізу українськомовного тексту. *Teka komisji polsko-ukraińskich związków kulturowych*. Lublin: Wydawnictwo KUL, 2018. vol 13. С. 191–206.

82. Зубань 2019: Зубань О. Автоматична конвертація паперового словника «Активні ресурси сучасної української номінації» в електронну лексикографічну систему. *Людина. Комп'ютер. Комунікація: збірник наукових праць*. Львів: Видавництво Львівської політехніки, 2019. 1 електрон. опт. диск (DVD-ROM). С. 61–72.

83. Зятковская 1980: Зятковская Р. Г. Суффиксальная система современного английского языка. Москва: Высшая школа, 1980. 147 с.

84. Ингве 1965: Ингве В. Гипотеза глубины. *Новое в лингвистике*. Москва:Изд-во иностр. лит., 1965. Вып. 4. С. 126–138.

85. Караулов 1976: Караулов Ю.Н. Общая и русская идеография. Москва: Наука. 1976. 355 с.

86. Караулов 1981: Караулов Ю.Н. Лингвистическое конструирование и тезаурус литературного языка. Москва: Наука, 1981. 366 с.

87. Карпіловська 1986: Карпіловська Є.А. Карпіловський В.С. Автоматизація побудови нових лінгвістичних об'єктів. *Мовознавство: науково-теоретичний журнал*. Київ, 1986. №6. С. 63 – 67.

88. Карпіловська 1990: Карпіловська Є.А. Конструювання складних словотворчих одиниць.К.: Наукова думка,1990. 156 с.

89. Карпіловська 1992: Карпіловська Є.А. Морфемна сітка як інструмент дослідження будови слова. *Українське мовознавство*. Київ, 1992. Вип.19. С.100 – 110.

90. Карпіловська 1999: Карпіловська Є. А. Суфіксальна підсистема сучасної української мови: будова та реалізація. Київ: УкрНДІПСК, 1999. – 297 с.

91. Карпіловська 2002: Карпіловська Є. А. Кореневий гніздовий словник української мови. Київ: Українська енциклопедія, 2002. 912 с..

92. Карпіловська 2006: Карпіловська Є.А. Вступ до прикладної лінгвістики: комп'ютерна лінгвістика: підручник. Донецьк: ТОВ «Юго-Восток, Лтд», 2006. 188 с.

93. Карпіловська 2007, 2008: Карпіловська Є.А. Тенденції розвитку сучасного українського лексикону: чинники стабілізації інновацій. *Українська мова: науково-теоретичний журнал*. Київ, 2007. № 4. С. 3–15; 2008. № 1. С. 24–35.

94. Карпіловська 2008: Карпіловська Є.А. Вплив інновацій на стабільність мовної системи: регулятори системної рівноваги. *Мовознавство: науково-теоретичний журнал*. Київ, 2008. № 2/3. С.148–158.

95. Карпіловська 2019: Карпіловська Є.А. Здобутки академічної структурної та математичної лінгвістики у моделюванні українського слова. *Українська мова: науково-теоретичний журнал*. Київ, 2019. № 1(69). С. 18–36.

96. Кибрик 1987: Кибрик А.Е. Лингвистические предпосылки моделирования языковой деятельности. *Моделирование языковой деятельности в интелектуальных системах* / Под ред. А.Е. Кибрика. Москва: Наука, 1987. С. 33–50.

97. Кислюк 2012: Кислюк Л.П. Сучасна словотвірна норма української мови: мовна практика та кодифікація // *Українська мова: науково-теоретичний журнал*. Київ, 2012. № 1. С. 52–66.

98. Кислюк 2017: Кислюк Л.П. Сучасна українська словотвірна номінація: ресурси та тенденції розвитку. Київ: Видавничий Дім Дмитра Бураго, 2017. 424 с.

99. ККТРГ 2018: Компьютерный корпус текстов русских газет конца XX-ого века. URL: http://www.philol.msu.ru/~lex/corpus/corp_descr.html (дата звернення 09.10.2018).

100. КЛ 2005: Корпусна лінгвістика / Широков В. А. та ін.; НАН України, Укр. мов.-інформ. фонд. Київ: Довіра, 2005. 472 с.

101. Клименко 1973: Клименко Н.Ф. Система афіксального словотворення сучасної української мови: монографія . К: Наукова дука, 1973. 186 с.

102. Клименко 1975: Клименко Н.Ф. Морфологічна будова композитів. *Морфологічна будова сучасної української мови*. Київ: Наукова думка, 1975. С. 5–34.

103. Клименко 1984: Клименко Н.Ф. Словотворча структура і семантика складних слів у сучасній українській мові. Київ: Наукова думка, 1984. 251 с.

104. Клименко 1996: Клименко Н. Морфеміка і словотворення як частини категорійної граматики української мови. *3 Міжнародний конгрес україністів. Харків, 26-29 серпня 1996: [Тези і повідомл.]*.Харків,1996. С.196-200.

105. Клименко 1998б: Клименко Н.Ф., Карпіловська Є.А. Морфеміка слов'янських мов як об'єкт типологічного вивчення. *Мовознавство: науково-теоретичний журнал*. Київ, 1998. № 2 – 3. С. 117 – 135.

106. Клименко 1998: Клименко Н. Ф. Основи морфеміки сучасної української мови: навч. посіб . Київ: ІЗМН,1998. – 182 с.

107. Клименко 1998а: Клименко Н. Ф., Карпіловська Є.А. Словотвірна морфеміка сучасної української мови. Київ: Інститут мовознавства НАН України, 1998. 270 с.

108. Клименко 2008: Клименко Н. Ф., Карпіловська Є. А., Кислюк Л. П. Динамічні процеси в сучасному українському лексиконі. Київ: Видавничий Дім Дмитра Бураго, 2008. 336 с.

109. Клименко 2014: Клименко Н.Ф. Составление словарей морфем с помощью ЭВМ. *Клименко Н.Ф. Вибрані праці* / Упор. Є.А. Карпіловська, О.Д.Пономарів, А.О. Савенко. Київ: Видавничий дім Дмитра Бураго, 2014. С. 535–544.

110. Клименко 2014а: Клименко Н.Ф., Карпіловська Є.А. Морфемні структури слів у сучасній українській літературній мові. *Клименко Н.Ф. Вибрані праці* / Упор. Є.А. Карпіловська, О.Д.Пономарів, А.О. Савенко. Київ: Видавничий дім Дмитра Бураго, 2014. С. 223–239.

111. Клименко 2014б: Клименко Н.Ф., Карпіловська Є.А., Комарова Л.І. та ін. Морфемно-словотвірний фонд української мови як дослідницька та інформаційно-довідкова система. *Клименко Н.Ф. Вибрані праці* / Упор. Є.А. Карпіловська, О.Д.Пономарів, А.О. Савенко. Київ: Видавничий дім Дмитра Бураго, 2014. С.545–558.

112. Клобуков 1976: Клобуков Е.В. Морфонология как парадигматическая морфемика. *Вопросы русского языкознания*. Москва: Изд.МГУ, 1976. Вып. 1. С. 82–92.

113. Ковалик 1958: Ковалик І. І. Питання слов'янського іменникового словотвору. Львів: Видавництво Львівського університету, 1958. 154 с.

114. Ковалик 1961: Ковалик І. І. Вчення про словотвір. Львів: Видавництво Львівського університету,1961. 83 с.

115. Ковалик 1971: Ковалик І. І. Дериватологія (словотвір) як самостійна лінгвістична дисципліна та її місце у системі науки про мову. *Словотвір сучасної української літературної мови*. Київ: Наукова думка, 1971. С.5–56.

116. Кожина 1966: Кожина М. Н. О специфике художественной и научной речи в аспекте функциональной стилистики. Пермь: [б. и.], 1966. 213 с.

117. Козленко 2014: Козленко І. Морфемологія. Морфологія сучасної української мови: навчальний посібник / Л. Алексієнко, О. Зубань, І. Козленко. Київ: ВПЦ "Київський університет", 2014. С. 28–215.

118. Копотев 2008: Копотев М., Мустайоки А. Современная корпусная русистика. *Инструментарий русистики: корпусные подходы.* Хельсинки, 2008. С. 7-24.

119. Котова 1978: Котова Н. В., Янакиев М. О. О многообразии морфем в славянских языках. *Славянская филология*. Москва, 1978. Вып. X. С. 4–8.

120. Кретов 1999: Кретов А. А Морфемно-морфонологический словарь языка А.С. Пушкина / А. А. Кретов, Л. Н. Матыцина. Воронеж: Центрально-Черноземное книжное издательство, 1999. 208 с.

121. КТУМ 2018: Корпуси текстів української мови. URL: http://corpora.donnu.edu.ua/ (дата звернення 07.11.2018).

122. Кубрякова 1974: Кубрякова Е.С. Основы морфологического анализа. Москва: Наука, 1974. 216 с.

123. Кубрякова 1983: Кубрякова Е.С., Панкрац Ю.Д. Морфонология в описании языков. Москва: Наука, 1983. 117 с.

124. Кубрякова 1991: Кубрякова Е.С. Понятие морфемы в современных грамматических исследованиях за рубежом. *Морфема и проблемы типологии*. Москва: Наука,1991. С.150–177ст.

125. Кузнецова 1977: Кузнецова А. И. Морфемный анализ и проблемы диахронии. *Словообразовательные и семантико-синтаксические процессы в языке*: Межвуз. сб. науч. трудов. Пермь, 1977. С. 55–63.

126. Кузнецова 1986: Кузнецова А.И., Ефремова Т.Ф. Словарь морфем русского языка. Москва: Русский язык, 1986. 1136 с.

127. Кукушкина 2005: Кукушкина О.В., Поликарпов А.А., Пирятинская Е.Ф. Полистилевой корпус текстов современного русского языка: Аннотация. *«HumLang – Язык Человека»:* Сайт филологического факультета МГУ, 2005 URL: http://www.philol.msu.ru/~humlang/articles/polystylcorp.html (дата звернення 16.12.2018).

128. Кукушкина 2006: Кукушкина О., Поликарпов А., Токтонов Г. Корпусная неография и неология: системный анализ характеристик лексических неологизмов (на материале газетного раздела «Полистилевого корпуса современного русского языка»). Лексикографічний бюлетень: Зб. наук. пр. Київ: Ін-т української мови НАН України, 2006. Вип. 13. С. 10–15.

129. Кукушкина 2006а: Кукушкина О.В., Поликарпов А.А., Токтонов А.Г. Анализ системных характеристик словообразовательного процесса (На основе анализа новых лексических единиц газетного материала «Полистилевого корпуса современного русского языка»). *«HumLang – Язык Человека»:* Сайт филологического факультета МГУ, 2006 URL: http://www.philol.msu.ru/~humlang/articles/polystylcorp.html (дата звернення 05.10.2018)

130. КУМ 2019: Корпус української мови. URL: http://www.mova.info/corpus.aspx (дата звернення 12. 11. 2019).

131. Купріянов 2018: Купріянов Є. В. Лексикографічна система іспанської мови: феноменологія інтегрального опису. Київ: УМІФ НАНУ, 2018. 254 с.

132. Курилович 1965: Курилович Е. Структура морфемы. *Очерки полингвистике*. Москва: Изд.-во ин. лит., 1965. С. 71–92.

133. КУСС 2005: Короткий українсько-сербський словник сполучуваності слів. Навчальний словник /Айданич Д., Білоног Ю. Київ: ВПЦ "Київський університет", 2005. 126 с. [онлайнова версія]. URL: http://www.mova.info/Page3.aspx?l1=65 (дата звернення 03.03.2019)

134. Лопатин 1977: Лопатин В. В. Русская словообразовательная морфемика. Проблемы и принципы описания. Москва: Наука, 1977. 316 с.

135. ЛЭС 1990: Лингвистический энциклопедический словарь / гл. ред. В. Н. Ярцева. Москва: Сов. энциклопедия, 1990. 683 с.

136. Ляшевская 2016: Ляшевская О. Н. Корпусные инструменты в грамматических исследованиях русского языка. Москва: Издательский Дом ЯСК: Рукописные памятники Древней Руси, 2016. 520 с.

137. Мартине 1963: Мартине А. Основы общей лингвистики. *Новое в лингвистике*. Москва: Изд.-во ин. лит., 1963. Вып.3. С. 366–558.

138. Маслов 1961: Маслов Ю.С. О некоторых расхождениях в понимании термина ”морфема”. *Проблемы языкознания. Сборник в честь академика И.И.Мещанинова.* Ленинград: Изд-во Ленинградского ун-та, 1961. С. 140–153.

139. Маслов 1978: Маслов Ю.С. К семантической типологии морфем. *Русский язык: вопросы его истории и современного состояния*. Москва: Наука, 1978. С. 5–18.

140. Мацюк 2018: Мацюк Р. Таблиця окремих термінів, вживаних у диференційній геометрії (багатомовний перекладний словник). URL: http://www.mova.info/Page2.aspx?l1=221 (дата звернення 24.04.2018).

141. Мельников 1978: Мельников Г. П. Системология и языковые аспекты кибернетики. Москва: Сов. радио, 1978. 368 с.

142. Мельчук 2016: Мельчук И. А., Жолковский А. К. Толково-комбинаторный словарь русского языка: Опыты семантико-синтаксического описания русской лексики. 2-е изд., испр. Москва: Глобал Ком: Языки славянской культуры, 2016. 544 с

143. Моисеев 1968: Моисеев А. И. Наименование лиц по профессии в современном руском литературном языке: автореф. докт. дис. Ленинград, 1968. 27 с.

144. Моисеев 1987: Моисеев А. И. Основные вопросы словообразования в современном русском литературном языке: Учеб. пособие. Ленинград: Изд-во Ленинград, ун-та, 1987. 207 с.

145. Москович 1967: Москович В. А. Глубина и длина слова в естественных языках. *Вопросы языкознания*. Москва, 1967. № 6. С. 17–33.

146. МСС 1979: Морфемна структура слова: монографія / Т. О. Грязнухіна та ін. Київ: Наукова думка, 1979. 327 с.

147. Нелюбин 1983: Нелюбин Л. П. Перевод и прикладная лингвистика. Москва: Высшаяшкола, 1983. 208 с.

148. НКРЯ 2018: Национальный корпус русского языка. URL: http://www.ruscorpora.ru (дата звернення 06.09.2018).

149. ОВСУМ 2018: Онлайнова версія академічного тлумачного «Словника української мови» в 11 томах (1970–1980). URL: http://sum.in.ua (дата звернення 09.08.2019).

150. Оливериус 1976: Оливериус З. Ф. Морфемы русского языка: частотный словарь. Praga: Univerzita Karlova, 1976. 175 с.

151. Откупщикова 1963: Откупщикова М. И. Об одном возможном способе построения формальной морфологии. *Материалы по математической лингвистике и машинному переводу*. Ленинград: Изд-во Ленинградского университета, 1963. Сб. 2. С. 61–66.

152. Пазельская 2009: Пазельская А. Г. Модели деривации и синтаксическая позиция отглагольных существительных по корпусным данным. *Компьютерная лингвистика и интелектуальные технологии: метриалы ежегод. Международной конференции "Диалог – 2009*". Москва: Изд-во РГГУ, 2009. Вып. 8 (15). С. 373–378.

153. Пазельская 2009а: Пазельская А. Г. Модели деривации отглагольных существительных: взгляд из корпуса. *Корпусные исследования по русской грамматике* / Ред.-сост. К.Л. Киселева, В.А. Плунгян, Е.В. Рахилина, С.Г. Татевосов. Москва: Пробел, 2009. С. 65–91.

154. Палкова 2015: Палкова А.В. Основные понятия электронной лексикографии. *Вестник ТвГУ. Серия «Филология»*. Тверь: Твер. гос. ун-т, 2015. № 4. С. 88-93

155. Пацкин 2002: Пацкин А. И. Гиперсловари на базе системы «Абриаль». *Компьютерная лингвистика и интелектуальные технологии: материалы ежегод. Международной конференции «Диалог – 2002»*. Москва: Изд-во РГГУ, 2002. URL: http://www.dialog-21.ru/digest/2002/articles/packin/ (дата звернення 06.07.2018).

156. Пацкин 2004: Пацкин А.И. Опыт построения полной морфемно-ориентированной семантической сети для русского языка *Компьютерная лингвистика и интелектуальные технологии: материалы ежегод. Международной конференции «Диалог – 2004»*. Москва: Изд-во РГГУ, 2004. URL: http://www.dialog-21.ru/media/2553/packin.pdf. (дата звернення 06.07.2018).

157. Перванов 2011: Перванов Я. А. Заметки по электронной лексикографии. *Ithaca NY. Cornell University Library-arxiv.org,* 2011. 8 с. URL: https://arxiv.org/pdf/1107.1753.pdf (дата звернення 12.07.2018)

158. Перебийніс 1969: Перебийніс В. С. Напрями і школи у сучасній структурній лінгвістиці. Проблеми та методи структурної лінгвістики. Київ: [б. в.],1969. Вип.2. С. 8–24.

159. Перебийніс 1970: Перебийніс В. С. Кількісні та якісні характеристики системи фонем сучасної української літературної мови. К: Наукова думка, 1970. 272 с.

160. Перебийніс 1985: Перебийніс В.С., Муравицька М.П., Дарчук Н.П. Частотні словники та їх використання: монографія / за ред. В.С. Перебийніс. Київ: Наукова думка, 1985. 203 с.

161. Перебийніс 2002: Перебийніс В.С. Статистичні методи для лінгвістів. Вінниця: Нова Книга, 2001. 168 с.

162. Перебийніс 2009: Перебийніс В. І., Сорокін В.М. Традиційна та комп'ютерна лексикографія: навч. посібник. Київ: Вид. центр КНЛУ, 2009. 218 с.

163. Пещак 1966: Пещак М. М. Суфіксальні поля та їх особливості в системі словотвору українських топонімів. *Статистичні та структурні лінгвістичні моделі*. Київ: Наукова думка, 1966. С.97–107.

164. Пещак 1974: Пещак М. М., Гриднєва Л. М. Комбінаторика графем української літературної мови. Структура мови і статистика мовлення. Київ: Наукова думка, 1974. С.51–73.

165. Пиотровский 1968: Пиотровский Р. Г. Информационные измерения языка. Ленинград: Наука, 1968. 116 с.

166. Пиотровский 1999: Пиотровский Р. Г. Лингвистический автомат и его речемыслительное обоснование: Учеб. пособие. Минск: Изд-во Мин. гос. лингвист. ун-та, 1999. 196 с.

167. Плунгян 2008: Плунгян В. А. Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики. *Русский язык в научном освещении.* Москва: Нестор – История. №2 (16), 2008. С.7–20.

168. Плунгян 2009: Плунгян В. А. Почему современная лингвистика должна быть лингвистикой корпусов?: (публичная лекция, прочитанная 01.10.2009) URL: http://www.polit.ru/lectures/2009/10/23/corpus.html (дата звернення 05.04.2018).

169. Поликарпов 1998: Поликарпов А. А., Богданов В. В., Крюкова О. С. Хронологический морфемно-словообразовательный словарь русского языка: создание базы данных и ее системно-квантитативный анализ. *Вопросы общего, сравнительно-исторического, сопоставительного языкознания*. Москва: Московский лицей, 1998. Т. 32. С. 172–184.

170. Поликарпов 2013: Поликарпов А.А. Модель жизненного цикла знака: к теоретическим основаниям исторической лексикологии и дериватологии. *Славянская лексикография. Международная коллективная монография* / Ред. М. И. Чернышева. Москва: Азбуковник, 2013. С. 679–702.

171. Полюга 1983: Полюга Л. М. Морфемний словник. Київ: Радянська школа, 1983. 462с.

172. Полюга 2009: Полюга Л. М. Словник українських морфем: 3-є вид. Київ: Довіра, 2009. 554 с.

173. Поляков 2008: Поляков А. Е. Словарь языка А. С. Грибоедова. Т. 1: А–З. Москва: Языки славянской культуры, 2008. 432 с.

174. Попко 2007: Попко Л. П. Неологизация в языке как трансляция культурно-лингвистической национальной ментальности. Київ: ГАРККиИ, 2007. 168 с.

175. Пролинг РУТА 5.0: Лингвистические модули программы РУТА. URL: http://prolingoffice.com/product/ruta (дата звернення 26.11.2018).

176. Різників 2015. Різників О. Українські словогрона. Словогрона духу. Одеса: Симєекс-принт, 2015. 440 с.

177. Савченко 1974: Савченко І.Ф. Розподіл довжини слова в словнику української мови. *Структура мови і статистика мовлення.* Київ: Наукова думка, 1974. С.30–38.

178. Савченко 1990: Савченко И.Ф. Типология фонемной структуры морфем (на материале префиксальных прилагательных украинского языка) : автореф. дис. на соиск. уч. степ. канд. филол. наук, Киевский ордена Ленина и ордена Октябрьской Революции гос. ун-т им. Т. Г. Шевченка. Киев, 1990. 17 с.

179. САМУК 1998: Словник афіксальних морфем української мови / за ред. Н. Ф. Клименко. Київ: ВАТ УкрНДІПСК, 1998. 434 с.

180. СГСУЛМ 1972: Структурна граматика сучасної української літературної мови (проспект) / АН УРСР, Ін-т мовознав. ім. О. О. Потебні; відп. ред. В. С. Перебийніс. Київ: Наукова думка, 1972. 99с.

181. Селегей 2008: Селегей В. Электронные словари и компьютерная лексикография, 2008. – URL: http://www.lingvoda.ru/transforum/ articles/selegey_a1.asp (дата звернення 05.06.2018)

182. Селігей 2014: Селігей П. О. Етимологічний словник запозичених суфіксів і суфіксоїдів в українській мові. Київ: Академперіодика, 2014. 324 с.

183. Сікорська 1995: Сікорська З. С. Українсько-російський словотворчий словник: 2-е вид. Київ: Освіта, 1995. 256 с.

184. Сірук 2018: Сірук О., Сорокін В. Тезаурус комп'ютерної лексикографії. URL: http://www.mova.info/Page3.aspx?l1=188&vocid=1/ (дата звернення 24.04.2018).

185. СІС 1985: Словник іншомовних слів / за ред. О. С. Мельничука: 2-е видання, випр. і доп. Київ:УРЕ, 1985. 966с.

186. Скаличка 1967: Скаличка В. О грамматике венгерского языка. Пражский лингвистический кружок. Москва: Прогресс, 1967. С. 128–196.

187. Скаличка 1967а: Скаличка В. О. Асимметрический дуализм языковых единиц, в кн.: Пражский лингвистический кружок. Москва: Прогресс, 1967. С. 114–128.

188. Скляревская 2013: Скляревская Г. Н. Современная русская лексикография: достижения и лакуны. *Славянская лексикография*. Москва: Азбуковник, 2013. С. 579–615.

189. Смирницкий 1948: Смирницкий А. И. Некоторые замечания о принципах морфологического анализа основ. *Доклады и сообщения филологического факультета МГУ*. Москва, 1948. Вып. 5. С. 21–26.

190. Сова 1970: Сова Л.З. Аналитическая лингвистика. Москва: Наука, 1970. 253 с.

191. Солнцев 1971: Солнцев В. М. Язык как системно-структурное образование. Москва:Наука, 1971. 292 с.

192. СПС 1967: Статистичні параметри стилів / за ред. В. С. Перебийніс. Київ: Наукова Думка, 1967. 260 с.

193. ССУМ 2005: Семантичний словник української мови. Київ, 2005. URL: http://www.mova.info/semvoc.aspx?l1=193 (дата звернення 21.05.2019).

194. Степанов 1975: Степанов Ю.С. Методы и принципы современной лингвистики. Москва: Наука, 1975. 310 с.

195. Степанов 1975а: Степанов Ю.С. Основы общего языкознания. Москва: Просвящение,1975. 271 с.

196. СтилеАнализатор-2 2014: Комплексная тексто-аналитическая система «СтилеАнализатор-2», основанная на Web-технологиях: разработка, наполнение данными и тестирование на прикладных задачах / А. А. Поликарпов и др. *Сайт лаборатории общей и компьютерной лексикологии и лексикографии*. Москва, 2014. URL: http://www.philol.msu.ru/~lex/khmelev/papers.html (дата звернення 17.05.2019).

197. Субботін 2008: Субботін С. О. Подання й обробка знань у системах штучного інтелекту та підтримки прийняття рішень: Навчальний посібник. Запоріжжя: ЗНТУ, 2008. 341 с.

198. СУЛМ 1973: Сучасна українська літературна мова: Лексика, фразеологія / Відп. ред. М. А. Жовтобрюх. Київ : Наук. думка, 1973. 438 с.

199. СУМ 1970–1980: Словник української мови. К: Наукова думка, 1970–1980. Т. 1–11.

200. Сухотин 1984: Сухотин Б.В. Выделение морфем в текстах без пробелов между словами. М: Наука, 1984. 97 с.

201. Таран 2011: Таран А. Конкурування номінацій у сучасній українській літературній мові: тенденції стабілізації нової лексики. Черкаси: Вид. Чабаненко Ю., 2011. 232 с.

202. ТАС 2018: Труднощі англійського слововживання для українців: словник-довідник. URL: http://www.mova.info/Page3.aspx?l1=207 (дата звернення 28.08.2018).

203. Татевосов 2009: Татевосов С. Г. Множественная префиксация и анатомия русского глагола. *Корпусные исследования по русской грамматике* / Под ред. К. Л. Киселева, В. А. Плунгян, Е. Рахилина, С. Г. Татевосова. Москва: Пробел-2000, 2009. С. 92–156.

204. Теньер 1988: Теньер Л. Основы структурного синтаксиса. Москва: Прогресс, 1988. 654 с.

205. Тимошенко 2013: Тимошенко П. Д. Студії над мовою Тараса Шевченка / Інститут української мови НАН України . Київ: КММ, 2013. – 224 с.

206. Тихонов 1971: Тихонов А. Н. Морфема как значимая часть слова. *Филологические науки*. Москва, 1971. № 6. С. 39–52.

207. Тихонов 1990: Тихонов А. Н. Словообразовательный словарь русского языка: В 2 т. Москва: Рус. яз., 2-е изд. 1990.

208. Токтонов 2006: Токтонов А. Г. Новая лексика в русских газетах 1990-х годов: системно-словообразовательный анализ: На материале «Компьютерного корпуса текстов русских газет конца XX века: диссертация кандидата филологических наук : 10.02.01. Москва, 2006. 207 с.

209. Трубецкой 1960: Трубецкой Н. С. Основы фонологии.Москва: Изд-во иностр. лит., 1960. 372 с.

210. Трубецкой 1967: Трубецкой Н. С. Некоторые соображения относительно морфонологии. *Пражский лингвистический кружок*. Москва: Прогресс, 1967. С. 115–119.

211. УГ 1986: Украинская грамматика / за ред. В. М. Русанівського. Київ:Наукова думка, 1986. 339с.

212. УІГС 2001: Українсько-італійський граматичний словник дієслів (інформаційно-довідкова система) / Л. Алексієнко та ін. Київ – Флоренція, 2001. URL: http://www.mova.info/italvoc.aspx?l1=94 (дата звернення 28.08.2018).

213. Улуханов 1996: Улуханов И.С. Единицы словообразовательной системы русского языка и их лексическая реализация. Москва: [б. и.], 1996. 221 с.

214. Уорт 1983: Уорт Д.С. Русский словообразовательный словарь. Введение: Пер. с англ. *Новое в зарубежной лингвистике*. Москва: Прогресс, 1983. Вып. 14. С. 227– 260.

215. УРАТ 2018: Українсько-російсько-англійський тезаурус з лінгвістичної термінології (інформаційно-пошукова система). URL: http://www.mova.info/mov_thes.aspx?l1=68 (дата звернення 28.08.2018).

216. Фриз 1962: Фриз Ч. Значение и лингвистический анализ. *Новое в лингвистике*. Москва: Изд-во иностр. лит., 1962. Вып 2. С. 98–117

217. Фрумкина 1960: Фрумкина Р. М. Статистическая структура лексики Пушкина. *Вопросы языкознания*. Москва,1960. № 3. С. 78–81.

218. Харрис 1962: Харрис З. Совместная встречаемость и трансформация в языковой структуре. *Новое в лингвистике*. Москва: Изд-во иностр. лит., 1962. Вып.2. С. 528–637.

219. Харрис 1965: Харрис З. Метод в структурной лингвистике (Методологические предпосылки). История языкознания XIX – XX веков в очерках и извлечениях: В 2-х ч. / Под ред. В. А. Звегинцева. 3-е изд. Москва: Просвящение, 1965. Ч.2. С. 209–228.

220. Хомский 1961: Хомский Н. Три модели описания языка. Кибернетический сборник. Москва, 1961. Вып.2. С. 237–266.

221. Чепик 2006: Чепик Е. Ю. Компьютерная лексикография как одно из направлений современной прикладной лингвистики. *Ученые записки ТНУ*. Симферополь, 2006. Т.19 (58). No2: Филология. С. 274–280.

222. ЧСКостенко 2019: Частотний словник збірки «Вибране» Л. Костенко. URL: http://www.mova.info/cfq.aspx?fdid=lkzb (дата звернення 15.03.2019).

223. ЧСКУМ 2018: Частотні словники Корпусу української мови. URL: http://www.mova.info/article.aspx?l1=210&DID=5215 (дата звернення 15.03.2018).

224. ЧСНС 2017: Частотний словник наукового стилю. URL: http://www.mova.info/Page2.aspx?l1=176 (дата звернення 15.03.2017).

225. ЧССтус 2019: Частотний словник збірки «Палімпсести» В. Стуса. URL: http://www.mova.info/cfq.aspx?fdid=stuspol1(дата звернення 15.03.2019) .

226. ЧССУП 2017: Частотний словник сучасної української публіцистики. URL: http://www.mova.info/Page2.aspx?l1=91 (дата звернення 15.03.2017).

227. ЧССУПМ 2017: Частотний словник сучасної української поетичної мови. URL: http://www.mova.info/Page2.aspx?l1=89 (дата звернення 15.03.2017).

228. ЧССУХП 1981: Частотний словник сучасної української художньої прози. К: Наукова думка, 1981. Т. 1–2.

229. ЧСТХФ 2017: Чотиримовний словник термінів з хімії та фізики / коорд. проекту Ульрік Грубе. URL: http://www.mova.info/Page3.aspx?l1=60 (дата звернення 15.03.2017).

230. ЧСУкраїнка 2019: Частотний словник збірки "На крилах пісень" Лесі Українки. URL: http://www.mova.info/cfq.aspx?fdid=lunp (дата звернення 15.03.2019).

231. ЧСХП 2017: Частотний словник художньої прози. URL: http://www.mova.info/Page2.aspx?l1=90 (дата звернення 15.03.2017).

232. ЧСШевченко 2017: Частотний словник мови Т. Шевченка: "Твори в п’яти томах. URL: http://www.mova.info/cfqsh_2.aspx(дата звернення 15.03.2017).

233. ЧСШевченко 2019: Частотний словник мови Т. Шевченка: "Твори в п’яти томах. URL: http://www.mova.info/cfqsh_2.aspx(дата звернення 23.03.2019)

234. ЧЮЖ 1965: Чжао-Юань-Жень. Модели в лингвистике и модели вообще. *Математическая логика и её применения*. Москва: Мир,1965. С. 281–293.

235. Шанский 1959: Шанский Н.М. Очерки по русскому словообразованию. Москва: Учпедгиз, 1959. 245 с.

236. Шаумян 1963: Шаумян С. К., Соболева П. А. Аппликативная порождающая модель и исчисление трансформаций в русском языке. Москва: Изд-во Акад. наук СССР, 1963. 125 с.

237. Шаумян 1965: Шаумян С. К. Структурная лингвистика. Москва: Наука, 1965. 395с.

238. Шевелёва 1973: Шевелёва П. А. Алгоритм вычленения морфов внутри беспобельного текста. НТИ, серия 2. Москва,1973. №6. С. 20–23.

239. Широков 1998: Широков В. А. Інформаційна теорія лексикографічних систем. Київ: Довіра, 1998. 331 с.

240. Широков 2004: Широков В. А. Феноменологія лексикографічних систем. Київ: Наукова думка, 2004. 327 с.

241. Широков 2005: Широков В. А. Елементи лексикографії. Київ: Довіра, 2005. 304 с.

242. Широков 2011: Широков В. А. Комп’ютерна лексикографія. Київ: Наук. думка, 2011. 351 с.

243. ШСССУМ 2005: Шкільний словотвірний словник сучасної української мови / Н. Ф. Клименко та ін. Київ: Наукова думка, 2005. 264 с.

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

245. ЯКМИСЛ 2006: Языковая картина мира и системная лексикография / Отв. ред Ю. Д. Апресян. Москва: Языки славянских культур, 2006. 912 с.

246. Яценко 1980: Яценко І. Т. Морфемний аналіз: Словник-довідник: У 2 т. / За ред. Н. Ф. Клименко. Київ: Вища школа, 1980–1981. Т.1: А–Н. 355с. Т.2: О–Я. 352 с.

247. Anderson 1982: Anderson St.R. Where's Morphology. *Linquistic Inquiri*, 1982. vol.13, №4. C. 610–678.

248. Fodor 1980: Fodor J. D. Semantics: Theories of Meaning in Generative Grammar. Harvard University Press, 1980. 236 p.

249. Fries 1952: Fries Ch. C. The Structure of English: An Introduction to the Construction of English Sentences. New York: Harcourt, Brace, 1952. 304 p.

250. GWJH 1984: Gramatyka współczesnego jezyka polskiego. Morfologia. Warszawa, 1984. 545 s.

251. Hockett 1947: Hockett Ch. Problems of morphemic analysis. *Lanquage*, 1947. vol. 23, №4. P. 321–343.

252. Horecki 1964: Horecki I. Morfematická struktura slovenciny. Bratislava, 1964.– 423 s.

253. ISO/IEC 2015: ISO/IEC 2382:2015, Information technology: Vocabulary. Part 1: Fundamental terms. URL: https://www.iso.org/standard/63598.html. (дата звернення: 17.12.2018).

254. eLex 2019: eLex 2019 – electronic lexicography in the 21 st century. URL: https://elex.link/elex2019/ (дата звернення: 10.09.2019).

255. Leech 1997: Leech G. Introducing corpus annotation. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Addison Wesley Longman, 1997. P. 1–9.

256. Menzerath 1954: Menzerath P. Die Architektonik des deutschen Wortschatzes. Phonetische Studien. Bonn, Hannover, Stuttgart: Ferdinand Dummlers Verlag, 1954. 131 s.

257. Slavíčková 1975: Slavíčková E.: Retrográdní morfematický slovník češtiny. Praha: EAV, 1975. 645 s.

258. Teubert 207: Teubert W. Corpus linguistics and lexicography. *Text Corpora and Multilingual Lexicography*. Amsterdam/ Philadelphia: John Benjamins Publishing Company, 2007. P. 109–134.

259. Zipf 1949: Zipf G. Human Behavior and the Principle of Least Effort. Cambridge (Maas): Addison-Wesley, 1949. – 573 p.

260. Zuban 2015: Zuban O. M. Morphemic and derivational analysis in the corpus of the Ukrainian language. *Українське мовознавство*. Київ: Видавничо-поліграфічний центр "Київський університет", 2015. Вип. 1 (45). С. 3–10.

261. Zuban 2017: Zuban O. Automatic Morphemic Analysis in the Corpus of the Ukrainian Language: Results and Prospects. *Jazykovedný časopis / Journal of Linguistics*. Bratislava, 2017. ROČNÍK 68, № 2. P. 415–426.

262. Zuban 2019: Zuban O. Lexicographical Database of Frequency Dictionaries of Morphemes Developed on the Basis of the Corpus of Ukrainian Language. *Advances in Intelligent Systems and Computing IV. CCSIT 2019. Advances in Intelligent Systems and Computing*. Springer, Cham, 2019 vol 1080. P. 549–566.

263. Zuban 2019a: Zuban O. The Morphemic System Stylometric Analysis of the Ukrainian Poets' Idiostyles: Corpus Based Approach. *Лінгвістичні студії Linguistic Studies: зб. наук. праць*. Вінниця, 2019. vol. 38. С. 96–104.

Д О Д А Т К И

ДОДАТОК 1. Табличний запис дистрибутивної характеристики афіксальної морфеми -АК- [АК; АЧ; К; Ч; АЦ; АЦ’; Ц; Ц’] (Подано за статтею Зубань О. М. Створення морфемної бази даних: принципи опису морфеми як інваріантної одиниці. [Зубань 2001: 57 – 58].)

R#F	кр′іп> К -Ø^{2}~ вой> -а^{1}~(-і-)	шв> АЧ –Ø^{2}~ хиж> –у^{1д}~ йун> –ий~		
R#S	піш> -`а			
R#SF	бурл> -уйу-т′^{ді} свой> -ин′-а^{1} хиж> -ів-Ø піш> -ов-ий прост> -уват-ий	шв> -к-а закр′іп> -уйу^{ді}-т′ розкр′іп> -ен′н′-а вой> -ен′к-о бід> -ис′к-о риб> -их-а сл′іп> -ишч-е прост> -ок-Ø коз> -ин-а шв> -ів-Ø		
RS#F	мерз>л′- -Ø^{2}~ сі>в- -Ø^{2}~ пис′>м- -Ø^{2}~ паруб>ч- -Ø^{2}~ глит>ай- -а^{1}~(-і-) завод′>ій- -а^{1}~(-і-) бід>ол- -а^{1}~(-і-)	с′і>в- -Ø^{2}~	спожи>ва- Ч -Ø^{2}~ оповід>а- -Ø^{2}~ завід>ува- -Ø^{2}~ жебр>а- -а-т′^{2д}	жебр>а- К -Ø^{2}~ з′ів>а- -а^{1}~(і) с′пі>в- -Ø^{2}~ зада>ва- -а^{1}~(і)
R#SS	бурл> -ува-ти	хиж> -и-ти закр′іп> -ува-ти		
R#SSF		шв> -чин-а коз> -чин-а свой> -енниц′-а перв> -ечн-к-а глум> -нник-Ø	спону>к-а- -Ø^{2}~ посвід>ч-ува- -Ø^{2}~	
RS#SF	хлоп>ч- -уват-ий мерз>л′- -ів-Ø сі>в- -ів-Ø пис′>м- -ів-Ø паруб>ч- -ів-Ø	мерз>л′- -к-а с′і>в- -к-а глит>ай- -ин-Ø завод′>ій- -ин-Ø бід>ол- -ин-Ø	оповід>а- -к-а завід>ува- -к-а спі>ва- -ок-Ø спі>ва- -к-а тк>а- -енк-о тк>а- -их-а тк>а- -ук-Ø тк>а- -івн-а	жебр>а- -уй-ут′^{1д} жебр>а- -ів-Ø зада>ва- -уват-ий

Продовження таблиці

RS#SS			жебр>а- -и[2]-ти	жебр>а- -ува-ти
RSS#F			дос′пі>в-ува- -Ø[2]~ с′і>й-а- -Ø[2]~ влашт>ов-ува- -Ø[2]~	
RS#SSF		мерз>л′- -чин-Ø с′і>в- -чин-Ø	пк>а- -ишнин-Ø оповід>а- -чин-Ø завід>ува- -чин-Ø с′пі>ва- -чин-Ø	жебр>а-К - уват′-іс′т′-Ø
RSS#SF			пос′від>ч-ува- -к-а зас′пі>в-ува- -к-а с′і>й-а- -к-а пос′від>ч-ува- -ів-Ø зас′пі>в-ува- -ів-Ø с′і>й-а- -ів-Ø	
RSS#SSF			зас′пі>в-ува-Ч-ч-ин-Ø сі>й-а- -чин-Ø	
R#F	вой> **АЦ′** -і			
R#SF	кр′іп> -к-ий	хиж> **АЦ**- тв-о		
RS#F	слит>ай- -і завод′>ій -і бід>ол- -і			з′ів>а– **Ц′** –і
R#SS	покр′іп> -к-и покр′іп> -к-ому			
RS#SF	с′і>в- -к-ий пис′>м- -к-ий чуд>ерн- -к-ий бід>н′- -к-ий	чуд>ерн- -тв-о пис′>м- -тв-о	жебр>а – **Ц** – тв-о	жебр>а- -к-ий с′пі>ва- -к-ий завід>ува- -к-ий
RS#SS	пос′і>в- -к-и, ому попис′>м- -к-и, ому почуд>ерн- -к-и ому побідн>н′- -к-и, ому			пожебр>а-**Ц′**-к-и, ому пос′пі>ва-**Ц′**-к-и, ому завід>ува-**Ц′**-к-и, ому

ДОДАТОК 1.1. Список програмних процедур МС посткороневої зони українського дієслова

№	процедура	приклад квазіфлексії, яку моделює процедура	приклад словоформи
1	*>	пови	*упови>ти*
2	>*	ба	*руб>а-ти*
3	>**	ава	*да>ва-ти*
4	>***	авува	*прав>ува-ти*
5	>****	чугур	*куч>угур-ю*
6	>*****	оловинь	*попол>овинь*
7	>*-*	ражда	*враж>д-а-ти*
8	>*-**	рава	*нагр>а-ва-ти*
9	>*-***	ьбува	*гань>б-ува-ти*
10	>*-****	жествл	*обож>е-ствл-ю*
11	>**-*	атіша	*багат>іш-а-ти*
12	>**-**	бесну	*хлеб>ес-ну-ти*
13	>**-***	авдовува	*виправд>ов-ува-ти*
14	>***-*	рмани	*дур>ман-и-ти*
15	>***-**	йчикуй	*лакей>чик-уй*
16	>***-***	атствува	*меценат>ств-ува-ти*
17	>****-*	инувачи	*звин>увач-и-ти*
18	>****-**	стикулюю	*жест>икул-юю*
19	>****-***	нувачува	*звин>увач-ува-ти*
20	>*-*-*	бали	*недб>а-л-и-ти*
21	>*-*-**	одаткую	*опода>т-к-ую*
22	>*-*-***	япакува	*дряп>а-к-ува-ти*
23	>*-**-*	ергота	*джер>г-от-а-ти*
24	>*-**-**	ивішаю	*жи>в-іш-аю*
25	>*-**-***	адковува	*успад>к-ов-ува-ти*
26	>*-***-*	айомля	*озна>й-омл-я-ть*
27	>*-***-**	чителюю	*вч>и-тел-юю*
28	>*-***-***	азифікува	*газ>и-фік-ува-ти*
29	>*-****-*	жествля	*обож>е-ствл-я-ти*
30	>*-****-**	ествлюй	*обож>е-ствл-юй*
31	>*-****-***	ествлюва	*обож>е-ствл-ювa-ти*
32	>**-*-*	накша	*переін>ак-а-ти*
33	>**-*-**	ломкну	*чол>ом-к-ну-ти*
34	>**-*-***	твертува	*четв>ер-т-ува-ти*
35	>**-**-*	кавіша	*ласк>ав-іш-а-ти*
36	>**-**-**	ливішає	*сміл>ив-іш-ає*
37	>**-**-***	ріалізува	*індустрі>ал-із-ува-ти*
38	>**-***-*	ітнича	*співроб>ім-нич-а-ти*
39	>**-***-**	гатствую	*ренег>ат-ств-ую*
40	>**-***-***	отствува	*нім>от-ств-ува-ти*
41	>***-*-**	шиствую	*фаш>ист-в-ую*
42	>***-*-***	шиствува	*фаш>ист-в-ува-ти*
43	>***-**-*	течніша	*стат>ечн-іш-а-ти*

44	>***-**-**	фіковую	*перекваліфік-ов-ую*
45	>***-**-***	овува	*перекваліфік-ов-ува-ти*
46	>***-***-*	лостивля	*умилост-ивл-я-ти*
47	>***-***-**	лостивляю	*умилост-ивл-яю*
48	>***-***-***	ентствува	*президент-ств-ува-ти*
49	>****-**-**	оналізую	*націонал-із-ую*
50	>****-**-***	оналізува	*професіонал-із-ува-ти*
51	>*-*-*-**	анкую	*дра-н-к-ую*
52	>*-*-*-***	анкува	*дра-н-к-ува-ти*
53	>*-*-**-*	айкота	*галай-к-от-а-ти*
54	>*-*-**-**	атковуй	*опота-к-ов-уй*
55	>*-*-**-***	атковува	*опота-к-ов-ува-ти*
56	>*-**-*-*	ркотні	*помар-к-от-н-і-ти*
57	>*-**-*-**	ркотнію	*мар-к-от-н-ію*
58	>*-**-**-*	глявіша	*смуг-л-яв-іш-а-ти*
59	>*-**-**-**	хнявішаю	*порох-н-яв-іш-аю*
60	>*-**-**-***	уалізува	*індивід-у-ал-із-ува-ти*
61	>*-**-***-**	иянствую	*христ-и-ян-ств-ую*
62	>*-**-***-***	иянствува	*христ-и-ян-ств-ува-ти*
63	>**-*-**-*	чайніша	*позвич-ай-н-іш-а-ти*
64	>**-*-**-**	арничаю	*куст-ар-н-ич-аю*
65	>**-*-**-***	урнішува	*окульт-ур-н-іш-ува-ти*
66	>**-**-**-**	іалізовую	*матері-ал-із-ов-ую*
67	>**-**-**-***	іалізовува	*матері-ал-із-ов-ува-ти*
68	>**-**-***-**	тиканствую	*політ-ик-ан-ств-ую*
69	>**-**-***-***	тиканствува	*крит-ик-ан-ств-ува-ти*
70	>***-*-**-*	тальнича	*сентимент-аль-н-ич-а-ти*
71	>***-*-**-**	тальнішаю	*сентимент-аль-н-іш-аю*

ДОДАТОК 2. Морфологічні коди у системі АГАТ (У списку кодів частково використано дані за монографією Дарчук Н. П. Комп'ютерне анотування тексту: результати і перспективи: монографія. [Дарчук 2013: 356]**)**

Для дієслів:

1) ГФ — інфінітив;
2) ГП — інфінітив з постфіксом -ся;
3) ГR — 1 особа однини теперішнього часу;
4) ГV — 1 особа однини теперішнього часу з постфіксом -ся;
5) ГW — 2 особа однини теперішнього часу;
6) Г0 — 2 особа однини теперішнього часу з постфіксом -ся;
7) ГЯ — 3 особа однини теперішнього часу;
8) ГА — 3 особа однини теперішнього часу з постфіксом -ся;
9) ГЖ — 1 особа множини теперішнього часу;
10) ГК — 1 особа множини теперішнього часу з постфіксом -ся;
11) ГЦ — 2 особа множини теперішнього часу,
12) ГО — 2 особа множини теперішнього часу з постфіксом -ся;
13) ГЮ — 3 особа множини теперішнього часу;
14) ГУ — 3 особа множини теперішнього часу з постфіксом -ся;
15) ГЙ — минулий час жіночого роду;
16) ГИ — минулий час жіночого роду з постфіксом -ся;

17) ГР — минулий час середнього роду;
18) ГЛ — минулий час середнього роду з постфіксом -ся;
19) ГЕ — минулий час чоловічого роду;
20) ГЭ — минулий час чоловічого роду з постфіксом -ся;
21) ГН — множина минулого часу;
22) ГМ — множина минулого часу з постфіксом -ся;
23) ГZ — 1 особа однини майбутнього часу;
24) Г6 — 1 особа однини майбутнього часу з постфіксом -ся;
25) ГS — 2 особа однини майбутнього часу;
26) Г9 — 2 особа однини майбутнього часу з постфіксом -ся;
27) ГD — 3 особа однини майбутнього часу;
28) ГG — 3 особа однини майбутнього часу з постфіксом -ся;
29) Г4 — 1 особа множини майбутнього часу з постфіксом -ся
30) Г2 — 2 особа множини майбутнього часу;
31) Г7 — 2 особа множини майбутнього часу з постфіксом -ся;
32) Г3 — 3 особа множини майбутнього часу;
33) Г8 — 3 особа множини майбутнього часу з постфіксом -ся;
34) Г5 — 2 особа однини наказового способу;
35) ГІ — 2 особа однини наказового способу з постфіксом -ся;
36) ГС — 1 особа множини наказового способу;
37) ГХ — 1 особа множини накзового способу з постфіксом -ся;
38) ГШ — 2 особа множини наказового способу;
39) ГЩ — 2 особа множини наказового способу з постфіксом –ся;
40) ГД – безособові дієслова типу *холоднішає* (співвідносне з 3 ос.одн. без-ся);
41) Г К– безособові дієслова типу *похолодало* (співвідносне з с.р. одн. без -ся);
42) ГҮ – безособові дієслова типу *хотілося* (співвідносне з с.р.одн. з -ся)

Для іменників чоловічого роду (Й), жіночого (К), середнього (Л) і pluralia tantum друга літера двоелементного коду позначає:

И - ім.наз.в. одн. А - ім.наз.в. мн.
Р - ім.род.в. одн. Е - ім.род.в. мн.
Д - ім.дав.в. одн. О - ім.дав.в. мн.
В - ім.знах.в. одн. У - ім.зн.в. мн.
Т - ім.ор.в. одн. Ю - ім.ор.в. мн.
П - ім.місц.в. одн. Я - ім.місц.в. мн.
К – клична форма

Для ад'єктивних класів (код А) – власне прикметників, займенників-прикметників, порядкових прикметників пропонуються такі граматичні коди (друга літера двоелементного коду):

И – наз.в.ч.р.одн.	Ж - наз.в.ж.р.одн.	С - наз.в.с.р.одн.	А – наз. в.мн.
Р - род.в.ч.р.одн.	З - род.в.ж.р.одн.	Ч - род.в.с.р.одн.	Е – род. в.мн.
Д – дав.в.ч.р.одн.	К - дав.в.ж.р.одн.	Ф - дав.в.с.р.одн.	О – дав. в.мн.
В - зн.в.ч.р.одн.	Л - зн.в.ж.р.одн.	Х - зн.в.с.р.одн.	У – зн. в.мн.
Т - ор.в.ч.р.одн.	М - ор.в.ж.р.одн.	Ц - ор.в.с.р.одн.	Ю – ор. в.мн.
П – місц.в.ч.р.одн.	Н - місц.в.ж.р.одн.	Ш- місц.в.с.р.одн.	Я – місц. в.мн.

Для займенників-іменників (код М):

И – наз.в.ч.р.одн.	Ж - наз.в.ж.р.одн.	С - наз.в.с.р.одн.	А – наз. в.мн.
Р - род.в.ч.р.одн.	З - род.в.ж.р.одн.	Ч - род.в.с.р.одн.	Е – род. в.мн.
Д – дав.в.ч.р.одн.	К - дав.в.ж.р.одн.	Ф - дав.в.с.р.одн.	О – дав. в.мн.
В - зн.в.ч.р.одн.	Л - зн.в.ж.р.одн.	Х - зн.в.с.р.одн.	У – зн. в.мн.

Т - ор.в.ч.р.одн. М - ор.в.ж.р.одн. Ц - ор.в.с.р.одн. Ю – ор. в.мн.
П – місц.в.ч.р.одн. Н - місц.в.ж.р.одн. Ш - місц.в.с.р.одн. Я – місц. в.мн.

Для кількісних числівників (код Ч):
И – наз.в.ч.р.одн. Ж - наз.в.ж.р.одн. С - наз.в.с.р.одн. А – наз. в.мн.
Р - род.в.ч.р.одн. З - род.в.ж.р.одн. Ч - род.в.с.р.одн. Е – род. в.мн.
Д – дав.в.ч.р.одн. К - дав.в.ж.р.одн. Ф - дав.в.с.р.одн. О – дав. в.мн.
В - зн.в.ч.р.одн. Л - зн.в.ж.р.одн. Х - зн.в.с.р.одн. У – зн. в.мн.
Т - ор.в.ч.р.одн. М - ор.в.ж.р.одн. Ц - ор.в.с.р.одн. Ю – ор. в.мн.
П – місц.в.ч.р.одн. Н - місц.в.ж.р.одн. Ш - місц.в.с.р.одн. Я – місц. в.мн.

Для присвійних невідмінюваних займенників :
ЪЗ - ***її***
ЪS - ***його***
ЪL - ***їх***

Для прийменників, які беруть участь у керуванні:
родовим відмінком – код ПР;
давальним – ПД;
знахідним – ПВ;
орудним – ПТ;
місцевим - ПП.

Сполучники одержали коди: СС – сполучник сурядний; СП – сполучник підрядний.

Для часток код [Ь0].

ДОДАТОК 2.1. Фрагмент БД спільнокореневих слів: корінь -голод // голодь-

morfvoc		
ID	wrd	morfem
15159	виголодатися	PCRHSIFKXM
15161	виголоджуватися	PCRISLFNXP/голод
15162	виголодніти	PCRHSISJFL
15163	виголоднілий	PCRHSISJSKFM
15164	виголодуватися	PCRHSKFMXO
24361	впроголодь	PBPERKSK/голод
29065	голод	RFFG
29066	голодний	RFSGFI
29067	голодненький	RFSGSKFM
29068	голоднеча	RFSGSIFJ
29069	голодніший	RFSGSIFK
29070	голодніше	RFSGSISJ
29071	голодно	RFSGSH
29072	голодовка	RFSHSIFJ
29073	голодомор	RFIGRJSKFL/мор2/
29074	голодоморний	RFIGRJSKFM/ мор2/
29078	голодувати	RFSIFK
29079	голодування	RFSISKFL
29080	голодуючий	RFSHSJFL
50664	зголодніти	PBRGSHSIFK
50665	зголоднілий	PBRGSHSISJFL
50666	зголодніло	PBRGSHSISJSK
54724	ізголодніти	PCRHSISJFL
78106	наголодуватися	PCRHSKFMXO
78397	надголодь	PDRJSP/голод/
79479	найголодніший	PDRISJSLFN
81624	напівголодний	PFRKSLFN
85381	неголодний	PCRHSIFK
85382	неголодно	PCRHSISJ
188691	обголодь	PCRISI/голод
114187	поголодніти	PCRHSISJFL
114188	поголодувати	PCRHSKFM
124399	приголодніти	PDRISJSKFM
127767	проголодувати	PDRISLFN
163353	упроголодь	PBPERKSK/голод/

ДОДАТОК 3. Лексична вибірка спільноафіксальних слів із суфіксом -тв-

1. антимистецтво,Л,PERISKSMFN/
2. бавовництво,Л,RGSISKFL
3. баришництво,Л,RFSGSISKFL
4. басмацтво,Л,RGSIFJ/басмач/
5. батрацтво,Л,RGSIFJ/батрак/
6. баштанництво,Л,RGSHSJSLFM
7. бджільництво,Л,RGSHSJSLFM/бджол/
8. безбожництво,Л,PDRGSHSJSLFM/Бог/
9. бешкетництво,Л,RGSHSJSLFM
10. битва,К,RCSEFF
11. бідацтво,Л,RDSFSHFI
12. бідняцтво,Л,RDSESGSIFJ
13. благодійництво,Л,REIFRISJSLSNFO/благ3, дij/
14. богобудівництво,Л,RDIERHSJSMSOFP/Бог, буд2/
15. боговідступництво,Л,RDIEPHRLSMSOSQFR/Бог, ступ4/
16. бортництво,Л,RESFSHSJFK/борть/
17. боягузтво,Л,RDSESHSJFK
18. бражництво,Л,RESFSHSJFK/браг/
19. бродяжництво,Л,RESGSJSLFM/брод1/
20. будівництво,Л,RDSFSGSISKFL/буд2/
21. бурлацтво,Л,RESGSIFJ/бурл2/
22. бурсацтво,Л,RESGSIFJ
23. буряківництво,Л,RFIHSKSMFN
24. вбозтво,Л,RESGFH/убог/
25. великодержавництво,Л,RDSFIGRMSNSPSRFS/вел1, держав/
26. вигадництво,Л,PCRFSISKFL/гад2/
27. вигнанництво,Л,PCRESFSGSJSLFM/ган1/
28. видавництво,Л,RESFSISKFL/вида1/
29. винахідництво,Л,PCRHSKSMFN/наход/
30. виробництво,Л,PCRFSISKFL/роб1/
31. витівництво,Л,PCRESFSISKFL
32. відлюдництво,Л,PDRGSHSJSLFM
33. відступництво,Л,PDRHSKSMFN/ступ4/
34. відхідництво,Л,PDRGSJSLFM/ход1/
35. відходництво,Л,PDRGSJSLFM/ход1/
36. візництво,Л,RDSGSIFJ/віз1/
37. віровідступництво,Л,RDIEPHRLSOSQFR/вір, ступ4/
38. віршомазництво,Л,REIFRISLSNFO
39. владицтво,Л,RESGSIFJ/волод/
40. власництво,Л,RFSHSJFK
41. вовноткацтво,Л,REIFRHSISJSLFM/вовн, тк2/
42. волоцюзтво,Л,RHSJFK//

43.вояцтво,Л,RDSFSHFI
44.гайдамацтво,Л,RISKFL/гайдамак-/
45.гідробудівництво,Л,REIFRISKSLSNSPFQ/гідр2, буд2/
46.гірництво,Л,RDSESGSIFJ/гор2/
47.головотество,Л,RFIGRJSLFM/голов1, тес/
48.голубівництво,Л,RFIHSKSMFN/голуб2/
49.гонитва,К,RDSESGFH/ган1/
50.городництво,Л,RFSISKFL/город1/
51.грабіжництво,Л,RESGSJSLFM/граб3/
52.градобудівництво,Л,REIFRISKSNSPFQ/город2, буд2/
53.градоначальництво,Л,REIFRLSOSQFR//
54.грибівництво,Л,REIGSJSLFM
55.гуральництво,Л,RGSHSJSLFM/
56.гуральництво,Л,RGSJSLFM
57.гусівництво,Л,RDIFSISKFL/гус1/
58.гутництво,Л,RDSGSIFJ/гут1/
59.дворушництво,Л,RCIDRGSJSLFM/дв, рук/
60.державництво,Л,RGSHSJSLFM
61.дивацтво,Л,RDSFSHFI/див2/
62.дівоцтво,Л,RDSFSHFI
63.діляцтво,Л,RDSFSHFI/діл2/від діляга
64.догідництво,Л,PCRFSGSISKFL/год1/
65.домобудівництво,Л,RDIERHSJSKSMSOFP/дім, буд2/
66.дослідництво,Л,RGSHSJSLFM
67.другорічництво,Л,REIFRISJSLSNFO/друг1, рік1/
68.духівництво,Л,RDSFSGSISKFL/дух3/
69.енергобудівництво,Л,RFIGRJSLSMSOSQFR/енерг, буд2/
70.єретицтво,Л,RFSHSJFK/jересь/
71.жебрацтво,Л,RESFSGSIFJ
72.женоненависництво,Л,RDIEPGPIRLSMSOSQFR/жін, вид1/
73.житлобудівництво,Л,RCSDSEIFRISKSLSNSPFQ/жи, буд2/
74.жінконенависництво,Л,RDSEIFPHRMSNSPSRFS
75.жіноцтво,Л,RDSFSHFI
76.жорства,К,RESGFH
77.жрецтво,Л,RCSESGFH
78.завойовництво,Л,PCRFSHSKSMFN/вој/
79.загарбництво,Л,PCRGSJSLFM/гарб2/
80.замісництво,Л,PCRFSISKFL/міст2/
81.занепадництво,Л,PCPERHSKSMFN/пас, па/
82.засновництво,Л,PCRGSJSLFM/сну/
83.заставництво,Л,PCRGSJSLFM/став12/
84.заступництво,Л,PCRGSHSJSLFM/ступ6/
85.затвірництво,Л,PCRGSJSLFM/твор2/
86.затворництво,Л,PCRGSJSLFM/твор2/

87.західництво,Л,RFSISKFL/заход/
88.збиральництво,Л,PBRESFSISKSMFN/бр/
89.звідництво,Л,PBRESHSJFK/вод15/
90.звірівництво,Л,RESJSLFM
91.здирництво,Л,PBRESHSJFK/дер1/
92.земляцтво,Л,RESGSIFJ
93.зерновиробництво,Л,REIFPHRKSNSPFQ/зерн, роб1/
94.змовництво,Л,PBRESHSJFK/мов1/
95.знавецтво,Л,RDSESGSISJ
96.зрадництво,Л,RESHSJFK
97.каліцтво,Л,RFSHFI/калік/
98.каракулівництво,Л,RHSJSMSOFP
99.катеробудівництво,Л,RFIGRJSMSOSQFR
100. католицтво,Л,RHSJFK/католик/
101. каучуківництво,Л,RGSISLSNFO
102. квітництво,Л,RESHSJFK//
103. керівництво,Л,RDSFSGSISKFL
104. килимоткацтво,Л,RFIGRISJSKSMFN
105. кіновиробництво,Л,REIEPGRJSMSOFP
106. кіномистецтво,Л,RERISKSMFN
107. клятва,К,RDSFFG
108. клятвений,А,RDSFSHFJ
109. клятвено,Н,RDSFSHSI
110. клятвопорушення,Л,RDSFIGPIRLSOFP/рух/
111. клятвопорушний,А,RDSFIGPIRLSMFO/рух/
112. клятвопорушник,Й,RDSFIGPIRLSOFP/рух/
113. клятвопорушниця,К,RDSFIGPIRLSOFP/рух/
114. кляузництво,Л,RFSGSISKFL
115. книговидавництво,Л,REIFPHRJSKSNSPFQ
116. козацтво,Л,RDSFSHFI
117. козівництво,Л,RDSISKFL
118. колісництво,Л,RFSGSISKFL
119. кон’юнктурництво,Л,RKSLSNSPFQ//
120. кочівництво,Л,RDSFSISKFL//
121. кравецтво,Л,RGSIFJ
122. кріпацтво,Л,RESGSIFJ
123. кріпосництво,Л,RESHSJSLFM//
124. кролівництво,Л,RESGSJSLFM//
125. кукурудзівництво,Л,RISNSPFQ
126. культурництво,Л,RHSKSMFN//
127. кунацтво,Л,RFSHFI/кунак/
128. купецтво,Л,RDSFSHFI
129. курівництво,Л,RDSISKFL
130. либацтво,Л,RDSFSHFI//

131. лихацтво,Л,RDSFSHFI//
132. лівацтво,Л,RDSFSHFI
133. лівонародництво,Л,RDIERJSKSMSOFP//
134. лісівництво,Л,RDSFSISKFL
135. лісництво,Л,RDSESGSIFJ
136. ліярництво,Л,RDSFSISKFL//
137. ловецтво,Л,RDSFSHFI//
138. ловитва,К,RDSESGFH
139. луківництво,Л,RDSFSISKFL
140. людиноненависництво,Л,RDSFIGPIRNSOSQSSFT/навид/
141. льоноткацтво,Л,REIFRHSISJSLFM
142. мазурництво,Л,RFSISKFL
143. мандрівництво,Л,RFSHSISKSMFN
144. маніяцтво,Л,RFSHSJFK
145. маралівництво,Л,RFIHSKSMFN
146. мертвенність,К,RDSFSISMFN//
147. мертвецька,К,RDSFSISJFK//
148. мертвеччина,К,RDSFSISKFL//
149. мертвовід,Й,RDSFIGRJSKFL//
150. мертворожденний,А,RDSFIGRKSNFP//
151. мистецтво,Л,RESGSIFJ
152. мистецтвознавець,Й,RESGSIIJRMSNSQFR
153. мистецтвознавство,Л,RESGSIIJRMSNSQFR
154. мистецтвознавчий,А,RESGSIIJRMSNSOFQ
155. мірошництво,Л,RFSISKFL
156. місництво,Л,RDSGSIFJ
157. містобудівництво,Л,REIFRISKSLSNSPFQ
158. мішечництво,Л,RDSFSGSISKFL/міх/
159. молитва,К,RDSESGFH
160. молитвеник,Й,RDSESGSISKFL
161. молитвениця,К,RDSESGSKFL//
162. молитвослов,Й,RDSESGIHRLFM//
163. молитвувати,Г,RDSESGSJFL//
164. молодецтво,Л,RFSHSJFK
165. мосяжництво,Л,RFSGSISKFL
166. мрійництво,Л,RESFSHSJFK
167. мужицтво,Л,RDSFSHFI
168. мужолозтво,Л,RDIERHSJFK/лож/
169. музейництво,Л,RFSGSISKFL
170. мучеництво,Л,RDSFSHSJFK/мук/
171. надвиробництво,Л,PDPFRISLSNFO
172. надомництво,Л,PCRFSISKFL//
173. наложництво,Л,RFSISKFL
174. намісництво,Л,PCRFSISKFL/місц/

175. народництво,Л,RFSGSISKFL
176. насінництво,Л,RFSGSISKFL
177. наставництво,Л,PCRGSJSLFM/став2/
178. наступництво,Л,PCRGSHSJSLFM/ступ1/
179. начальництво,Л,RGSHSJSLFM/начал/
180. начотництво,Л,PCRFSISKFL
181. невігластво,Л,RISKFL
182. невільництво,Л,PCRGSHSJSLFM/вол3/
183. негідництво,Л,PCRFSGSISKFL/гід2/
184. незаможництво,Л,PCPERHSISKSMFN
185. неклятвений,А,PCRFSHSJFL
186. немертвонароджений,А,PCRFSHIIPKROSQFS/род/
187. ненависництво,Л,PCRHSISKSMFN
188. неуцтво,Л,PCRESGFH/уч/
189. низькопоклонництво,Л,RESFFGPIRMSPSRFS/низ1,клон/
190. новобудівництво,Л,RDIERHSJSKSMSOFP/нов,буд2/
191. областництво,Л,RFSGSISKFL/область/
192. обмежництво,Л,PCRFSISKFL//
193. овочівництво,Л,RESJSLFM
194. огородництво,Л,RGSJSLFM//
195. одиноцтво,Л,RESGSIFJ//
196. одноосібництво,Л,RDIERISJSLSNFO/особ/
197. окольництво,Л,RISKFL/окольник/
198. олійництво,Л,RESFSHSJFK
199. осадництво,Л,PBRESHSJFK//
200. осібництво,Л,RESFSHSJFK/особ/
201. отроцтво,Л,RFSHFI/отрок/
202. паломництво,Л,RFSISKFL
203. паркобудівництво,Л,REIFRISKSLSNSPFQ/парк,буд2/
204. партацтво,Л,RGSIFJ/партач/
205. партбудівництво,Л,RERHSJSKSMSOFP/партij, буд2/
206. парубоцтво,Л,RFSHSJFK/парубj/
207. пасічництво,Л,RFSGSISKFL/пасік/
208. паства,К,RDSFFG/пас3/віруючі одної парафії
209. перевиробництво,Л,PEPGRJSMSOFP/роб1/
210. передвижництво,Л,PERISLSNFO/двиг1/
211. письмацтво,Л,RESFSHSJFK/пис/
212. письменництво,Л,RESFSHSISKSMFN/пис/
213. пияцтво,Л,RDSFSHFI/пи/
214. підлабузництво,Л,PDRISLSNFO
215. підмитва,К,PDRFSHFI//
216. підсобництво,Л,PDRGSHSJSLFM
217. підступництво,Л,PDRHSISKSMFN/ступ5/
218. плодівництво,Л,RESGSJSLFM/плід/

219. плодоовочівництво,Л,REIFRJSLSOSQFR/плід, овоч/
220. подвижництво,Л,RGSJSLSM/подвиг/
221. позадництво,Л,PCRFSISKFL
222. політкерівництво,Л,RFRISKSLSNSPFQ/політик, кер/
223. полковництво,Л,RESGSJSLFM
224. поміщицтво,Л,RFSHSJFK
225. поп-мистецтво,Л,RDRISKSMFN/попул
226. попутництво,Л,PCRFSISKFL/путь/
227. поручництво,Л,PCRFSISKFL/рук/
228. посадництво,Л,RFSISKFL
229. посередництво,Л,PCRHSKSMFN
230. пособництво,Л,PCRFSISKFL
231. потурнацтво,Л,RFSISKFL//
232. правництво,Л,RESHSJFK/прав9/
233. правозаступництво,Л,REIFPHRLSMSOSQFR/прав9, ступ4/
234. правонаступництво,Л,REIFPHRLSMSOSQFR/прав9, ступ4/
235. представництво,Л,PERISLSNFO/став7/
236. пригодництво,Л,PDRGSJSLFM/год4/
237. прислужництво,Л,PDRHSKSMFN/слуг,служ/
238. пролазництво,Л,PDRGSJSLFM/лаз1/
239. проповідництво,Л,PDRISLSNFO
240. пророцтво,Л,PDRGSIFJ/рек/
241. просвітництво,Л,PDRHSISKSMFN/світ4/
242. простацтво,Л,RFSHSJFK/прост1/
243. пруссацтво,Л,RFSHSJFK/Пруссij/
244. псевдомистецтво,Л,RGRMSOFP/мистець/
245. птаство,Л,RCSESGFH
246. птахівництво,Л,RCSESJSLFM
247. пустельництво,Л,RESHSISKSMFN/пуст1/
248. пустинництво,Л,RESGSHSJSLFM/пуст1/
249. рабівництво,Л,RDSFSISKFL//
250. рабовласництво,Л,RDIERJSLSNFO/раб, власн/
251. радбудівництво,Л,RDRGSISJSLSNFO/рад3, буд2/
252. рахівництво,Л,RDSFSISKFL
253. рвацтво,Л,RCSDSESGFH/рв1/
254. ремество,Л,RFSHFI/ремес/
255. ремествувати,Г,RFSHSKSM//
256. ремісництво,Л,RFSGSISKFL/ремес/
257. рибацтво,Л,RDSFSHFI
258. рибництво,Л,RDSESGSIFJ
259. рисівництво,Л,RDSFSISKFL/рис1/
260. ритвина,К,RCSESGFH
261. рільництво,Л,RESFSHSJFK/рілл/
262. робітництво,Л,RDSFSISKFL/роб1/

263. розбишацтво,Л,RGSISKFL
264. розбійництво,Л,PDRGSHSJSLFM/би/
265. розжитво,Л,PDRFSHFI//
266. розкольництво,Л,PDRHSKSMFN/коло/
267. розпорядництво,Л,PDPFRISLSNFO//
268. рослинництво,Л,RDSESGSHSJSLFM
269. рукодільництво,Л,RDIERISJSLSNFO/рук, діл2/
270. саботажництво,Л,RFSHSKSMFN
271. садівництво,Л,RDSFSISKFL/сад1/
272. самітництво,Л,RDSFSGSISKFL/сам1/
273. самотництво,Л,RDSFSGSISKFL/сам1/
274. свідоцтво,Л,RESGSIFJ/свідок/
275. свояцтво,Л,RESGSIFJ/свій/
276. святенництво,Л,RESGSJSLFM/свят2/
277. селюцтво,Л,RDSFSHFI
278. середняцтво,Л,RFSGSISKFL
279. сільгоспвиробництво,Л,RERIPKRNSQSSFT/сел, госп, роб1/
280. сіпацтво,Л,RDSESFSHFI
281. склочництво,Л,RFSGSISKFL/склок/
282. скоморошство,Л,RISKFL/скоморох/
283. скотолозтво,Л,REIFRISKFL/скот, лож1/
284. смертництво,Л,PBRESFSISKFL/мер2/
285. снохацтво,Л,RESGSIFJ
286. собаківництво,Л,RFSKSMFN
287. соболівництво,Л,RFSKSMFN/соболь/
288. сонцепоклонництво,Л,RDSEIFPHRLSOSQFR/сон1, клан2/
289. соромітництво,Л,RFSHSISKSMFN
290. соцбудівництво,Л,RDSDIDRGSISJSLSNFO/соціал, буд2/
291. союзництво,Л,RFSISKFL
292. спиртовиробництво,Л,RFIGPIRLSOSQFR/спирт, роб1/
293. співробітництво,Л,PERHSJSMSOFP/роб1/
294. співтовариство,Л,PERLSNFO/товариш/
295. спільництво,Л,RFSGSISKFL
296. споживацтво,Л,PBPDRFSHSISKFL/жи1/
297. спорттовариство,Л,RFIFRMSOFP/спорт, товариш/
298. старецтво,Л,RESGSIFJ/стар1/
299. старообрядництво,Л,REIFPHRKSNSPFQ/стар1, ряд2/
300. страдництво,Л,RFSGSISKFL
301. стрілецтво,Л,RFSHSJFK/стріл1/
302. сумісництво,Л,PCRFSGSISKFL/міст2/
303. суперництво,Л,RFSISKFL//
304. супірництво,Л,PCRFSISKFL/пер2
305. сутяжництво,Л,PCRFSGSISKFL/тяг4
306. схимництво,Л,RESHSJFK//

307. тваринництво,Л,RESGSHSJSLFM//
308. телемистецтво,Л,RERISKSMFN/
309. ткацтво,Л,RCSDSESGFH/тк2/
310. товариство,Л,RHSJFK/товариш/
311. трудівництво,Л,RESJSLFM
312. трюкацтво,Л,RESGSIFJ
313. тютюнництво,Л,RFSISKFL
314. убозтво,Л,RESGFH/убог/
315. угадництво,Л,PBRESHSJFK/гад2
316. ударництво,Л,RESFSHSJFK
317. ухильництво,Л,PBRFSISKFL/хил1/
318. уходництво,Л,PBRESHSJFK/хід1/
319. учениництво,Л,RCSESGSIFJ
320. фальшивомонетництво,Л,RFSHIIRNSQSSFT/
321. фільмовиробництво,Л,RFIGPIRLSOSQFR/
322. фокусництво,Л,RFSISKFL/фокус2/
323. фотомистецтво,Л,RERKSMFN/фото, мистец/
324. хабарництво,Л,RFSISKFL
325. характерництво,Л,RISLSNFO/характер1/
326. харлацтво,Л,RESGSIFJ
327. харпацтво,Л,RESGSIFJ
328. харцизтво,Л,RGSIFJ/
329. харцизяцтво,Л,RGSISKFL
330. хвацтво,Л,RESGFH/хват/
331. хижацтво,Л,RDSFSHFI/хиж1/
332. хлоп’яцтво,Л,RFSHSJFK/хлоп1/
333. хмільництво,Л,RFSISKFL
334. хороство,Л,RFSHFI/хорош
335. храмобудівництво,Л,REIFRISKSLSNSPFQ/храм, буд2/
336. хробацтво,Л,RGSIFJ/хробак/
337. циркацтво,Л,RESGSIFJ
338. цитатництво,Л,RDSFSGSISKFL/цит2/
339. цитрусівництво,Л,RGSISLSNFO
340. чарівництво,Л,RDSFSISKFL/чар1/
341. черевоугодництво,Л,RFIGPHRKSNSPFQ/черев, год1/
342. чернецтво,Л,RESGSIFJ
343. чиновництво,Л,RDSGSISKFL/чин1/
344. чоловіцтво,Л,RHSJFK/чоловік/
345. чорнокнижництво,Л,REIFRJSKSMSOFP/чорн, книг/
346. чортяцтво,Л,RESGSIFJ/
347. чудацтво,Л,RDSFSHFI
348. чудернацтво,Л,RDSFSISKFL
349. чужолозтво,Л,RDIERHSJFK/леж
350. чумацтво,Л,RFSHFI/чумак/

351. швацтво,Л,RCSESGFH/ши/
352. шерстеткацтво,Л,RFIGRISJSKSMFN/шерсть, тк/
353. шитво,Л,RCSEFF
354. шістдесятництво,Л,RERJSMSOFP/шість, десять/
355. шкідництво,Л,RESHSJFK/шкод1/
356. шкільництво,Л,RFSISKFL/школ/
357. шкурництво,Л,RESFSHSJFK
358. шовківництво,Л,RESGSJSLFM
359. шовкоткацтво,Л,REIFRHSISJSLFM/шовк, тк/
360. юнацтво,Л,RDSFSHFI
361. ябедництво,Л,RFSISKFL
362. ягідництво,Л,RFSISKFL/jагод/
363. язицтво,Л,RFSHFI/jазич
364. язичництво,Л,RFSISKFL//

ДОДАТОК 4. База даних електронного словника афіксальних морфем (аломорф -тв- морфеми -ств- ’діяльність’)

SUFF								
suff	word	forml	znachennia	motuvuyuche slovo	usichennia	nakladannya	cherguvannya	interfiksacia
тв	бавовництво	RSSF	діяльність	бавововник		ц-ц	к//ц, с//ц	
тв	баришництво	RSSSF	діяльність	баришник		ц-ц	к//ц, с//ц	
тв	батрацтво	RSF	діяльність	батрак		ц-ц	к//ц, с//ц	
тв	баштанництво	RSSSF	діяльність	баштанник		ц-ц	к//ц, с//ц	
тв	бджільництво	RSSSF	діяльність	бджільник		ц-ц	о//і, к//ц, с//ц	
тв	благодійництво	RIRSSSF	діяльність	благодійник		ц-ц	к//ц, с//ц	
тв	бортництво	RSSSF	діяльність	бортник		ц-ц	к//ц, с//ц	
тв	бражництво	RSSSF	діяльність	бражник		ц-ц	к//ц, с//ц	
тв	будівництво	RSSSSF	діяльність	будівник		ц-ц	к//ц, с//ц	
тв	буряківництво	RISSF	діяльність	буряківник		ц-ц	к//ц, с//ц	
тв	вигадництво	PRSSF	діяльність	вигадник		ц-ц	к//ц, с//ц	
тв	видавництво	RSSSF	діяльність	видавник		ц-ц	к//ц, с//ц	
тв	винахідництво	PRSSF	діяльність	винахідник		ц-ц	к//ц, с//ц	
тв	виробництво	PRSSF	діяльність	виробник		ц-ц	і//о, к//ц, с//ц	
тв	витівництво	PRSSSF	діяльність	витівник		ц-ц	к//ц, с//ц	
тв	вовноткацтво	RIRSSSF	діяльність	вовноткач		ц-ц	ч//ц, с//ц	о
тв	візництво	RSSF	діяльність	візник		ц-ц	к//ц, с//ц	
тв	віршомазництво	RIRSSF	діяльність	віршомазник		ц-ц	к//ц, с//ц	о
тв	голубівництво	RISSF	діяльність	голубівник		ц-ц	к//ц, с//ц	
тв	городництво	RSSF	діяльність	городник		ц-ц	к//ц, с//ц	
тв	грабіжництво	RSSSF	діяльність	грабіжник		ц-ц	к//ц, с//ц	
тв	грибівництво	RISSF	діяльність	грибівник		ц-ц	к//ц, с//ц	
тв	гутництво	RSSF	діяльність	гутник		ц-ц	к//ц, с//ц	
тв	гідробудівництво	RIRSSSSF	діяльність	гідробудівник		ц-ц	к//ц, с//ц	о
тв	гірництво	RSSSF	діяльність	гірник		ц-ц	к//ц, с//ц	
тв	домобудівництво	RIRSSSSF	діяльність	домобудівник		ц-ц	і//о, к//ц, с//ц	о
тв	дослідництво	RSSSF	діяльність	дослідник		ц-ц	к//ц, с//ц	
тв	енергобудівництво	RIRSSSSF	діяльність	енергобудівник		ц-ц	к//ц, с//ц	о
тв	житлобудівництво	RSSIRSSSSF	діяльність	житлобудівник		ц-ц	к//ц, с//ц	
тв	завойовництво	PRSSSF	діяльність	завойовник		ц-ц	к//ц, с//ц	
тв	загарбництво	PRSSF	діяльність	загарбник		ц-ц	к//ц, с//ц	
тв	засновництво	PRSSF	діяльність	засновник		ц-ц	к//ц, с//ц	
тв	заступництво	PRSSSF	діяльність	заступник		ц-ц	к//ц, с//ц	
тв	збиральництво	PRSSSSF	діяльність	збиральник		ц-ц	к//ц, с//ц	
тв	звідництво	PRSSF	діяльність	звідник		ц-ц	к//ц, с//ц	
тв	здирництво	PRSSF	діяльність	здирник		ц-ц	к//ц, с//ц	

SUFF								
suff	word	forml	znachennia	motuvuyuche slovo	usichennia	nakladannya	cherguvannya	interfiksacia
тв	зерновиробництво	RIPRSSF	діяльність	зерновиробник		ц-ц	к//ц, с//ц	о
тв	каракулівництво	RSSSF	діяльність	каракулівник		ц-ц	к//ц, с//ц	
тв	катеробудівництво	RIRSSSF	діяльність	катеробудівник		ц-ц	к//ц, с//ц	о
тв	каучуківництво	RSSSF	діяльність	каучуківник		ц-ц	к//ц, с//ц	
тв	квітництво	RSSF	діяльність	квітник		ц-ц	к//ц, с//ц	
тв	килимоткацтво	RIRSSSF	діяльність	килимоткач		ц-ц	ч//ц, с//ц	о
тв	книговидавництво	RIPRSSSF	діяльність	книговидавник		ц-ц	к//ц, с//ц	о
тв	колісництво	RSSSF	діяльність	колісник		ц-ц	к//ц, с//ц	
тв	кравецтво	RSF	діяльність	кравець		ц-ц	ц'//ц, с//ц	
тв	кіновиробництво	RIPRSSF	діяльність	кіновиробник		ц-ц	к//ц, с//ц	
тв	либацтво	RSSF	діяльність	либак		ц-ц	к//ц, с//ц	
тв	ловецтво	RSSF	діяльність	ловець		ц-ц	ц'//ц, с//ц	
тв	луківництво	RSSSF	діяльність	луківник		ц-ц	к//ц, с//ц	
тв	льоноткацтво	RIRSSSF	діяльність	льоноткач		ц-ц	ч//ц, с//ц	
тв	лівацтво	RSSF	діяльність	лівак		ц-ц	к//ц, с//ц	
тв	лійарництво	RSSSF	діяльність	лійарник		ц-ц	к//ц, с//ц	
тв	мазурництво	RSSF	діяльність	мазурник		ц-ц	к//ц, с//ц	
тв	маралівництво	RISSF	діяльність	маралівник		ц-ц	к//ц, с//ц	
тв	мосяжництво	RSSSF	діяльність	мосяжник		ц-ц	к//ц, с//ц	
тв	музейництво	RSSSF	діяльність	музейник		ц-ц	к//ц, с//ц	
тв	мірошництво	RSSF	діяльність	мірошник		ц-ц	к//ц, с//ц	
тв	містобудівництво	RIRSSSSF	діяльність	містобудівник		ц-ц	к//ц, с//ц	о
тв	мішечництво	RSSSSF	діяльність	мішечник		ц-ц	к//ц, с//ц	
тв	надвиробництво	PPRSSF	діяльність	надвиробник		ц-ц	к//ц, с//ц	
тв	надомництво	PRSSF	діяльність	надомник		ц-ц	і//о, к//ц, с//ц	
тв	наставництво	PRSSF	діяльність	наставник		ц-ц	к//ц, с//ц	
тв	нaйізництво	PRSSF	діяльність	нaйізник		ц-ц	к//ц, с//ц	
тв	новобудівництво	RIRSSSSF	діяльність	новобудівник		ц-ц	к//ц, с//ц	
тв	обмежництво	PRSSF	діяльність	обмежник		ц-ц	к//ц, с//ц	
тв	овочівництво	RSSF	діяльність	овочівник		ц-ц	к//ц, с//ц	
тв	огородництво	RSSF	діяльність	огородник		ц-ц	к//ц, с//ц	
тв	одноосібництво	RIRSSSF	діяльність	одноосібник		ц-ц	к//ц, с//ц	о
тв	паркобудівництво	RIRSSSSF	діяльність	паркобудівник		ц-ц	к//ц, с//ц	
тв	партацтво	RSF	діяльність	партач		ц-ц	ч//ц, с//ц	
тв	партбудівництво	RRSSSSF	діяльність	партбудівник		ц-ц	к//ц, с//ц	
тв	пасічництво	RSSSF	діяльність	пасічник		ц-ц	к//ц, с//ц	
тв	перевиробництво	PPRSSF	діяльність	перевиробник		ц-ц	к//ц, с//ц	
тв	письмацтво	RSSSF	діяльність	письмак		ц-ц	к//ц, с//ц	
тв	письменництво	RSSSSSF	діяльність	письменник		ц-ц	к//ц, с//ц	
тв	подвижництво	RSSS	діяльність	подвижник		ц-ц	г//ж, к//ц, с//ц	
тв	позадництво	PRSSF	діяльність	позадник		ц-ц	к//ц, с//ц	
тв	посередництво	PRSSF	діяльність	посередник		ц-ц	к//ц, с//ц	
тв	пособництво	PRSSF	діяльність	пособник		ц-ц	к//ц, с//ц	
тв	правдошукацтво	RIRSSF	діяльність	правдошукач		ц-ц	ч//ц, с//ц	о
тв	проповідництво	PRSSF	діяльність	проповідник		ц-ц	к//ц, с//ц	
тв	птахівництво	RSSSF	діяльність	птахівник		ц-ц	к//ц, с//ц	
тв	підсобництво	PRSSSF	діяльність	підсобник		ц-ц	к//ц, с//ц	
тв	рабівництво	RSSSF	діяльність	рабівник		ц-ц	к//ц, с//ц	
тв	радбудівництво	RRSSSSF	діяльність	радбудівник		ц-ц	к//ц, с//ц	
тв	рахівництво	RSSSF	діяльність	рахівник		ц-ц	к//ц, с//ц	
тв	ремісництво	RSSSF	діяльність	ремісник		ц-ц	к//ц, с//ц	
тв	рибацтво	RSSF	діяльність	рибак		ц-ц	к//ц, с//ц	
тв	розбійництво	PRSSSF	діяльність	розбійник		ц-ц	к//ц, с//ц	
тв	розпорядництво	PPRSSF	діяльність	розпорядник		ц-ц	к//ц, с//ц	
тв	рукодільництво	RIRSSSF	діяльність	рукодільник		ц-ц	к//ц, с//ц	о
тв	склочництво	RSSSF	діяльність	склочник		ц-ц	к//ц, с//ц	
тв	скоморошство	RSF	діяльність	скоморох		с-с	х//с	
тв	скотолозтво	RIRSF	діяльність	скотолоз		з-з	с//з	
тв	снохацтво	RSSF	діяльність	снохач		ц-ц	ч//ц, с//ц	
тв	собаківництво	RSSF	діяльність	собаківник		ц-ц	к//ц, с//ц	
тв	соболівництво	RSSF	діяльність	соболівник		ц-ц	к//ц, с//ц	
тв	соцбудівництво	RSIRSSSSF	діяльність	соцбудівник		ц-ц	к//ц, с//ц	

SUFF								
suff	word	forml	znachennia	motuvuyuche slovo	usichennia	nakladannya	cherguvannya	interfiksacia
тв	спиртовиробництво	RIPRSSF	діяльність	спиртовиробник		ц-ц	к//ц, с//ц	о
тв	сумісництво	PRSSSF	діяльність	сумісник		ц-ц	к//ц, с//ц	
тв	сутяжництво	PRSSSF	діяльність	сутяжник		ц-ц	к//ц, с//ц	
тв	сільгоспвиробництво	RRPRSSF	діяльність	сільгоспвиробник		ц-ц	к//ц, с//ц	
тв	ткацтво	RSSSF	діяльність	ткач		ц-ц	ч//ц, с//ц	
тв	трудівництво	RSSF	діяльність	трудівник		ц-ц	к//ц, с//ц	
тв	трюкацтво	RSSF	діяльність	трюкач		ц-ц	ч//ц, с//ц	
тв	тютюнництво	RSSF	діяльність	тютюнник		ц-ц	к//ц, с//ц	
тв	ударництво	RSSSF	діяльність	ударник		ц-ц	к//ц, с//ц	
тв	ухильництво	PRSSF	діяльність	ухильник		ц-ц	к//ц, с//ц	
тв	фокусництво	RSSF	діяльність	фокусник		ц-ц	к//ц, с//ц	
тв	хабарництво	RSSF	діяльність	хабарник		ц-ц	к//ц, с//ц	
тв	характерництво	RSSF	діяльність	характерник		ц-ц	к//ц, с//ц	
тв	харцизяцтво	RSSF	діяльність	харцизяка		ц-ц	к//ц, с//ц	
тв	хмільництво	RSSF	діяльність	хмільник		ц-ц	к//ц, с//ц	
тв	храмобудівництво	RIRSSSSF	діяльність	храмобудівник		ц-ц	к//ц, с//ц	о
тв	циркацтво	RSSF	діяльність	циркач		ц-ц	ч//ц, с//ц	
тв	цитрусівництво	RSSSF	діяльність	цитрусівник		ц-ц	к//ц, с//ц	
тв	чарівництво	RSSSF	діяльність	чарівник		ц-ц	к//ц, с//ц	
тв	чаївництво	RSSSF	діяльність	чаївник		ц-ц	к//ц, с//ц	
тв	чорнокнижництво	RIRSSSF	діяльність	чорнокнижник		ц-ц	к//ц, с//ц	о
тв	швацтво	RSSF	діяльність	швач		ц-ц	ч//ц, с//ц	
тв	шерстеткацтво	RIRSSSF	діяльність	шерстеткач		ц-ц	ч//ц, с//ц	
тв	шкідництво	RSSF	діяльність	шкідник		ц-ц	к//ц, с//ц	
тв	шовкоткацтво	RIRSSSF	діяльність	шовкоткач		ц-ц	ч//ц, с//ц	о
тв	їздецтво	RSSF	діяльність	їздець		ц-ц	ц'//ц, с//ц	
тв	ґуральництво	RSSF	діяльність	ґуральник		ц-ц	к//ц, с//ц	

ДОДАТОК 5. Лексична вибірки слів із кореневим морфом -берег-

1. безберегий,А,PDRISIFK/берег1/
2. берегтися,Г,RFFHXJ/берег2
3. берегти,Г,RFFH/берег2/
4. берег,Й,RFFG/берег1/
5. берегиня,К,RFSHFI/ берег2
6. берегівка,К,RFSHSIFJ/берег1/
7. берегівський,а,RFSHSKFM/берег3/
8. береговий,А,RFSHFJ/берег1/
9. берегово,л,RFSHFI/берег3/
10. берегове,л,RFSHFI/берег3/
11. береговий1,й,RFSHFJ/берег3/
12. берегова,к,RFSHFI/берег3/
13. береговина,К,RFSHSJFK/берег1/
14. берегозахисний,А,RFIGRLSMFO/берег1, захист/
15. берегоукріплення,Л,RFIGPHRMSPFQ/берег1, кріп2/
16. берегоукріплювальний,А,RFIGPHRMSPSSFU/берег1, кріп2/
17. вберегти,Г,PBRGFI/берег2/
18. вберегтися,Г,PBRGFIXK/берег2/
19. зберегтися,Г,PBRGFIXK/берег2/
20. зберегти,Г,PBRGFI/берег2/

21.крутоберегий,А,REIFRKSKFM/крут1, берег1
22.оберегти,Г,PBRGFI/берег2/
23.оберегтися,Г,PBRGFIXK/берег2/
24.оберег,Й,PBRGSHFI/берег2/
25.поберегти,Г,PCRHFJ/берег2/
26.поберегтися,Г,PCRHFJXL/берег2/
27.приберегти,Г,PDRIFK/берег2/
28.уберегти,Г,PBRGFI/берег2/
29.уберегтися,Г,PBRGFI/берег2/
30.уздовжбереговий,А,PCRGRLSNFP/довг, берег1

ДОДАТОК 6. Лексична вибірка слів із кореневим морфом -береж-

1. безбережність,К,PDRISJSNFO/берег1/
2. безбережний,А,PDRISJFL/берег1/
3. безбережно,Н,PDRISJSK/берег1/
4. бережок,Й,RFSHFI/берег1/
5. бережани,и,RFSHFI/берег3/
6. бережанський,А,RFSHSKFM/берег3/
7. бережений,А,RFSHFJ/берег2/
8. бережений1,А,RFSHFJ/берег2/
9. бережечок,Й,RFSHSJFK/берег1/
10.бережина,К,RFSHFI/берег1
11.бережина1,К,RFSHFI/берег1/
12.бережистий,А,RFSIFK/берег1/
13.бережіння,Л,RFSIFJ/берег2/
14.бережкий,А,RFSGFI/берег2/
15.бережливий,А,RFSIFK/берег2/
16.бережливість,К,RFSISMFN/берег2/
17.бережливо,Н,RFSISJ/берег2/
18.бережной,й,RFSGFI/берег3/
19.бережна,к,RFSGFH/берег3/
20.бережність,К,RFSGSKFL/берег2/
21.бережний1,А,RFSGFI/берег2/
22.бережний,й,RFSGFI/берег3/
23.бережно,Н,RFSGSH/берег2/
24.бережняк,Й,RFSGSIFJ/берег1/
25.вбережено,@,PBRGSISJ/берег2/
26.відбережний,А,PDRISJFL/берег1/
27.енергозбереження,Л,RFIGPHRMSPFQ/енерг, берег2/
28.забережень,Й,PCRHSKFL/берег1/
29.збережений,А,PBRGSIFK/берег2/
30.збереженість,К,PBRGSISMFN/берег2/
31.збереження,Л,PBRGSJFK/берег2/

32.збережено,@,PBRGSISJ/берег2/
33.крутобережний,А,REIFRKSLFN/крут1, берег1/
34.лівобережець,Й,RDIERJSMFN/лів1, берег1/
35.лівобережний,А,RDIERJSKFM/лів1, берег1/
36.лісозбереження,Л,RDIEPFRKSNFO/ліс, берег2/
37.набережна,К,PCRHSIFJ/берег1/
38.набережний,А,PCRHSIFK/берег1/
39.надбережний,А,PDRISJFL/берег1/
40.надобережність,К,PDPERJSKSOFP/берег2/
41.найобережніший,А,PDPERJSKSMFO/берег2/
42.найобережніше,Н,PDPERJSKSMSN/берег2/
43.небезбережний,А,PCPFRKSLFN/берег1/
44.незбереження,Л,PCPDRISLFM/берег2/
45.необережний,А,PCPDRISJFL/берег2/
46.необережність,К,PCPDRISJSNFO/берег2/
47.необережно,Н,PCPDRISJSK/берег2/
48.неприбережний,А,PCPFRKSLFN/берег1/
49.обережний,А,PBRGSHFJ/берег2/
50.обережність,К,PBRGSHSLFM/берег2/
51.обережненько,Н,PBRGSHSLSM/берег2/
52.обережніший,А,PBRGSHSJFL/берег2/
53.обережніше,Н,PBRGSHSJSK/берег2/
54.обережно,Н,PBRGSHSI/берег2/
55.південноузбережний,А,RGSHIIPKRPSQFS/південь,берег1/
56.підбережний,А,PDRISJFL/берег1/
57.побережець,Й,PCRHSKFL/берег1/
58.побережанин,Й,PCRHIJSLFM/берег1/
59.побережина,К,PCRHSJFK/берег1/
60.побережний,А,PCRHSIFK/берег1/
61.побережник1,Й,PCRHSISKFL/берег2/
62.побережник,Й,PCRHSISKFL/берег1/
63.побережниця,К,PCRHSISKFL/берег1/
64.правобережець,Й,REIFRKSNFO/прав1, берег1/
65.правобережний,А,REIFRKSLFN/прав1, берег1/
66.правобережці,И,REIFRKSLFM/прав1, берег1
67.прибережений,А,PDRISKFM/берег2/
68.прибережено,@,PDRISKSL/берег2/
69.прибережний,А,PDRISJFL/берег1/
70.прибережник,Й,PDRISJSLFM/берег1/
71.прибережниця,К,PDRISJSLFM/берег1/
72.ресурсозбереження,Л,RGIHPIRNSQFR/ресурс, берег2/
73.самозбереження,Л,RDIEPFRKSNFO/сам1, берег2/
74.теплозбережність,К,REIFPGRLSMSQFR/тепл, берег2/
75.убережений,А,PBRGSIFK/берег2/

76.убереження,Л,PBRGSJFK/берег2
77.убережено,@,PBRGSISJ/берег2/
78.узбережний,А,PCRHSIFK/берег1/

ДОДАТОК 7. Лексична вибірка слів із кореневим морфом -беріг-

1. вберігатися,Г,PBRGSHFJXL/берег2/
2. вберігати,Г,PBRGSHFJ/берег2/
3. вберігання,Л,PBRGSHSJFK/берег2/
4. енергозберігаючий,А,RFIGPHRMSOSQFS/енерг, берег2/
5. зберігатися,Г,PBRGSHFJXL/берег2/
6. зберігати,Г,PBRGSHFJ/берег2/
7. зберігання,Л,PBRGSHSJFK/берег2/
8. зберігач,Й,PBRGSHSIFJ/берег2/
9. лісозберігання,Л,RDIEPFRKSLSNFO/ліс, берег2/
10. оберіг,Й,PBRGSHFI/берег2/
11. оберігати,Г,PBRGSHFJ/берег2/
12. оберігатися,Г,PBRGSHFJXL/берег2/
13. оберігання,Л,PBRGSHSJFK/берег2/
14. приберігати,Г,PDRISJFL/берег2/
15. ресурсозберігаючий,А,RGIHPIRNSPSRFT/ресурс, берег2/
16. уберігатися,Г,PBRGSHFJXL/берег2
17. уберігати,Г,PBRGSHFJ/берег2/
18. уберігання,Л,PBRGSHSJFK/берег2/

ДОДАТОК 8. Лексична вибірка слів із кореневим морфом -бережж-

1. безбережжя,Л,PDRJSJFK/берег1/
2. крутобережжя,Л,REIFRLSLFM/крут1, берег1/
3. лівобережжя,Л,RDIERKSKFL/лів1, берег1/
4. надбережжя,Л,PDRJSJFK/берег1/
5. побережжя,Л,PCRISIFJ/берег1/
6. правобережжя,Л,REIFRLSLFM/прав1, берег1/
7. прибережжя,Л,PDRJSJFK/берег1/
8. узбережжя,Л,PCRISIFJ/берег1/

ДОДАТОК 9. XML-представлення анотації фрагмента тексту НКРЯ: ***Цены в них ниже, чем в обычных магазинах*** (Подано за монографією О.М.Ляшевської «Корпусные инструменты в грамматических исследованиях русского языка» [Ляшевская 2016])

```
<word text="Цены">
 <ana>
  <el name="lex">
   <el-group>
    <el-atom>цена</el-atom>
   </el-group>
  </el>
  <el name="gramm">
   <el-group>
    <el-atom>S</el-atom>
    <el-atom>inan</el-atom>
    <el-atom>f</el-atom>
    <el-atom>pl</el-atom>
    <el-atom>nom</el-atom>
   </el-group>
  </el>
 </ana>
 <ana>
  <el name="sem">
   <el-group>
    <el-atom>r:abstr</el-atom>
    <el-atom>t:param</el-atom>
   </el-group>
  </el>
 </ana>
 <ana>
  <el name="flags">
   <el-group>
    <el-atom>animred</el-atom>
    <el-atom>capital</el-atom>
    <el-atom>first</el-atom>
    <el-atom>numred</el-atom>
    <el-atom>posred</el-atom>
   </el-group>
  </el>
 </ana>
```

```
<word text="в">
 <ana>
  <el name="lex">
   <el-group>
    <el-atom>в</el-atom>
   </el-group>
  </el>
  <el name="gramm">
   <el-group>
    <el-atom>PR</el-atom>
   </el-group>
  </el>
 </ana>
 <ana/>
</word>
<text> </text>
```

```
<word text="них">
 <ana>
  <el name="lex">
   <el-group>
    <el-atom>они</el-atom>
   </el-group>
  </el>
  <el name="gramm">
   <el-group>
    <el-atom>SPRO</el-atom>
    <el-atom>3p</el-atom>
    <el-atom>pl</el-atom>
    <el-atom>loc</el-atom>
   </el-group>
  </el>
 </ana>
 <ana>
  <el name="sem">
   <el-group>
    <el-atom>r:pers</el-atom>
   </el-group>
  </el>
 </ana>
 <ana/>
</word>
<text> </text>
```

ДОДАТОК 10. Фрагмент морфологічної анотації тексту (2 речення балади «Причинна» Т. Шевченка)

```
word_Id,word,code,lemm,sentence_number
22645522,Реве,ГЯ,ревти,1
22645523,та,СС,та,1
22645524,стогне,ГЯ,стогнати,1
22645525,Дніпр,йИ,Дніпр,1
22645526,широкий,АИ,широкий,1
22645527,",",,",",1
22645528,Сердитий,АИ,сердитий,1
22645529,вітер,ЙИ,вітер,1
22645530,завива,ГЯ,завивати,1
22645531,",",,",",1
22645532,Додолу,Н0,додолу,1
22645533,верби,КУ,верба,1
22645534,гне,ГЯ,гнути,1
22645535,високі,АУ,високий,1
22645536,",",,",",1
22645537,Горами,Н0,горами,1
22645538,хвилю,КВ,хвиля,1
22645539,підійма,ГЯ,підіймати,1
22645540,.,,.,1
22645541,І,СС,і,2
22645542,блідний,АИ,блідний,2
22645543,місяць,ЙИ,місяць,2
22645544,на,ПВ,на,2
22645545,ту,ОЛ,той,2
22645546,пору,КВ,пора,2
22645547,Із,ПР,із,2
22645548,хмари,КР,хмара,2
22645549,де-де,Н0,де-де,2
22645550,виглядав,ГЕ,виглядати,2
22645551,",",,",",2
22645552,Неначе,СП,неначе,2
22645553,човен,ЙИ,човен,2
22645554,в,ПП,в,2
22645555,синім,АШ,синій,2
22645556,морі,ЛП,море,2
22645557,То,СС,то,2
22645558,виринав,ГЕ,виринати,2
22645559,",",,",",2
22645560,то,СС,то,2
22645561,потопав,ГЕ,потопати,2
22645562,.,,.,2
```

ДОДАТОК 11. Фрагмент таблиці "temp_freq" (2 речення балади «Причинна» Т. Шевченка)

temp_freq					
Код	wrd	cls	vib	tid	sentnum
1	ревти	Г	1	10295	1
2	та	С	1	10295	1
3	стогнати	Г	1	10295	1
4	Дніпр	й	1	10295	1
5	широкий	А	1	10295	1
6	сердитий	А	1	10295	1
7	вітер	Й	1	10295	1
8	завивати	Г	1	10295	1
9	додолу	Н	1	10295	1
10	верба	К	1	10295	1
11	гнути	Г	1	10295	1
12	високий	А	1	10295	1
13	горами	Н	1	10295	1
14	хвиля	К	1	10295	1
15	підіймати	Г	1	10295	1
16	і	С	1	10295	2
17	блідний	А	1	10295	2
18	місяць	Й	1	10295	2
19	на	П	1	10295	2
20	той	О	1	10295	2
21	пора	К	1	10295	2
22	із	П	1	10295	2
23	хмара	К	1	10295	2
24	де-де	Н	1	10295	2
25	виглядати	Г	1	10295	2
26	неначе	С	1	10295	2
27	човен	Й	1	10295	2
28	в	П	1	10295	2
29	синій	А	1	10295	2
30	море	Л	1	10295	2
31	то	С	1	10295	2
32	виринати	Г	1	10295	2
33	то	С	1	10295	2
34	потопати	Г	1	10295	2

ДОДАТОК 12. Фрагмент таблиці "freq" (5076 - 5123)

freq									
Код	wrd	cls	xabs	xsred	sigma	v	R	d	morfem
5076	ревнитель	Й	1	**1,63934426229508E-02**[39]	0,126983060531391	7,74596669241483	1	1,110223E-15	RSF
5077	ревносно	Н	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	RSS
5078	ревнути	Г	3	4,91803278688525E-02	0,216244359971687	4,39696865275764	3	4,323538	RSF
5079	ревти	Г	22	0,360655737704918	0,923869036301992	2,56163687338279	14	6,692941	RF
5080	ревти-завивати	Г	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
5081	ревучий	А	2	3,27868852459016E-02	0,1780783687082	5,43139024560011	2	2,988105	RSF
5082	регот	Й	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
5083	реготатися	Г	3	4,91803278688525E-02	0,216244359971687	4,39696865275764	3	4,323538	RSFX
5084	реготня	К	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	RSF
5085	реєстер	Й	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
5086	ректи	Г	4	6,55737704918033E-02	0,356156737416401	5,43139024560011	2	2,988105	RF
5087	ремиґати	Г	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	RSS
5088	ремінь	Й	2	3,27868852459016E-02	0,1780783687082	5,43139024560011	2	2,988105	RF
5089	репетувати	Г	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	RSF
5090	реп’ях	Й	2	3,27868852459016E-02	0,1780783687082	5,43139024560011	2	2,988105	RF
5091	республіка	К	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	RF
5092	ретязь	Й	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	RF
5093	решотка	К	3	4,91803278688525E-02	0,282043451378447	5,73488351136175	3	2,596297	
5094	Ржавиця	к	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
5096	рибалка	Й	2	3,27868852459016E-02	0,1780783687082	5,43139024560011	2	2,988105	RSSF
5097	рибалонька	Й	5	8,19672131147541E-02	0,521507335101209	6,36238948823475	2	1,786191	
5098	рибка	К	2	3,27868852459016E-02	0,1780783687082	5,43139024560011	2	2,988105	RSF
5099	рибонька	К	2	3,27868852459016E-02	0,1780783687082	5,43139024560011	2	2,988105	RSF
5100	рибчина	К	2	3,27868852459016E-02	0,1780783687082	5,43139024560011	2	2,988105	RSSF
5101	ридати	Г	33	0,540983606557377	0,879153290347011	1,62510153670205	22	7,902003	RSF
5102	ридати-молитися	Г	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
5103	риза	К	5	8,19672131147541E-02	0,416664427119306	5,08330601085553	5	3,43748	RF
5104	рик	Й	2	3,27868852459016E-02	0,253966121062781	7,74596669241483	1	1,110223E-15	RSSF
5105	Рим	й	11	0,180327868852459	0,93226670217134	5,16984262113197	7	3,325762	RF
5106	римлянин	Й	4	6,55737704918033E-02	0,306693228424094	4,67707173346743	4	3,961926	RISF
5107	римський	А	2	3,27868852459016E-02	0,1780783687082	5,43139024560011	2	2,988105	RSF
5108	риплючий	А	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
5109	рити	Г	3	4,91803278688525E-02	0,282043451378447	5,73488351136175	2	2,596297	RF
5110	рити-рити	Г	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
5111	рифма	К	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
5112	риштовати	Г	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
5113	риштувати	Г	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	RSF
5114	рівно	Н	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	RF
5115	рівня	К	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	RF
5116	ріг	Й	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	RF
5117	рід	Й	10	0,163934426229508	0,450148531729376	2,7459060435492	8	6,45505	RF
5118	рідний	А	4	6,55737704918033E-02	0,247535555254779	3,77491721763537	4	5,126603	
5119	рідня	К	5	8,19672131147541E-02	0,274314762798058	3,3466401061363	4	5,679506	RSF
5120	рідонький	А	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
5121	різати	Г	17	0,278688524590164	0,925612735311023	3,32131628552779	7	5,712199	RSF
5122	різатися	Г	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	RSFX
5123	різник	Й	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	RSF

[39] Символи E-02 у кінці числа вказують, що кому необхідно перенести на 2 знаки вліво (4,91803278688525E-02 → 0,0491803278688525), кількість перенесення знаків позначається числом після E (1,110223E-15 → 0,000000000000001110223). Конвертація числа в інтерфейсі частотних словників здійснюється автоматично.

ДОДАТОК 13. Фрагмент вибірки (А-Б) таблиці "freq": слова із невизначеною морфемною будовою

freq									
Код	**wrd**	**cls**	**xabs**	**xsred**	**sigma**	**v**	**R**	**d**	**morfem**
18	**а**	**С**	**1224**	**20,0655737704918**	**5,44380788997174**	**0,271300883405454**	**230**	**9,649752**	
19	Ааронія	й	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
22	абичий	О	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
25	ав	Ь	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
28	агу	В	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
29	ад	Й	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
31	**аж**	**ь**	**126**	**2,0655737704918**	**1,66810830779213**	**0,807576244248571**	**60**	**8,957423**	
32	ажеж	Ь	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
34	аксамит	Й	4	6,55737704918033E-02	0,247535555254779	3,77491721763537	4	5,126603	
35	аксамитний	А	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
36	аксамитовий	А	2	3,27868852459016E-02	0,1780783687082	5,43139024560011	2	2,988105	
37	але	С	5	8,19672131147541E-02	0,274314762798058	3,3466401061363	5	5,679506	
39	Алкід	й	6	9,83606557377049E-02	0,761898363188344	7,74596669241483	4	-2,220446E-15	
40	Алкідовий	а	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
42	Альбано	л	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
43	Альта	к	2	3,27868852459016E-02	0,1780783687082	5,43139024560011	2	2,988105	
45	Амон	й	4	6,55737704918033E-02	0,507932242125563	7,74596669241483	1	1,110223E-15	
47	Анафан	й	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
50	**ані**	**С**	**21**	**0,344262295081967**	**0,765768571418142**	**2,22437537411936**	**12**	**7,128344**	
60	Аппієвий	а	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
63	аренда	К	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
64	арестант	Й	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
67	армянин	Й	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
71	архістратиг	Й	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
73	Аскоченський	а	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
74	ась	Ь	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
78	б		1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
80	бо́сий	А	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
89	багато	Ч	2	3,27868852459016E-02	0,1780783687082	5,43139024560011	2	2,988105	
98	базиліанин	Й	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
106	бакаляр	Й	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
107	Бакчисарай	й	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
116	баня-прохолода	К	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
119	барило	Й	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
133	батько-отаман	Й	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
138	бачиться	Ь	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
139	бачиш	Ь	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
140	бачся	Ь	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
141	бачте	Ь	5	8,19672131147541E-02	0,274314762798058	3,3466401061363	5	5,679506	
142	баша	Й	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
147	**без**	**П**	**110**	**1,80327868852459**	**1,45781862789611**	**0,80842669365148**	**60**	**8,956326**	
155	беззаконіє	Л	3	4,91803278688525E-02	0,216244359971687	4,39696865275764	3	4,323538	
160	Безрукий	й	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
165	безчестіє	Л	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
168	Бендерський	а	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
170	бенкетовати	Г	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
174	берегами	Н	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
184	бисть	U	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
188	Бихів	й	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	

freq									
Код	**wrd**	**cls**	**xabs**	**xsred**	**sigma**	**v**	**R**	**d**	**morfem**
190	бігма	Н	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
192	Біда	й	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
194	**бідний**	**А**	**10**	**0,163934426229508**	**0,485201593014713**	**2,95972971738975**	**8**	**6,179005**	
203	білохатий	А	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
208	біс	Й	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
210	бісноватий	А	2	3,27868852459016E-02	0,1780783687082	5,43139024560011	2	2,988105	
218	благоволеньє	Л	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
222	благоденствіє	Л	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
223	благодушне	Н	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
225	благосклонний	А	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
233	ближченько	Н	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
245	**бо**	**С**	**150**	**2,45901639344262**	**1,78861399304733**	**0,727369690505912**	**71**	**9,060969**	
256	боже	В	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
267	Борзна	к	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
268	Бориспóль	й	2	3,27868852459016E-02	0,253966121062781	7,74596669241483	1	1,110223E-15	
269	Боровиків	а	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
278	босфоровий	А	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
284	браво	В	2	3,27868852459016E-02	0,253966121062781	7,74596669241483	1	1,110223E-15	
292	брат-запорожець	Й	2	3,27868852459016E-02	0,1780783687082	5,43139024560011	2	2,988105	
293	**брати**	**Г**	**34**	**0,557377049180328**	**0,932554928087075**	**1,67311325333269**	**25**	**7,84002**	
298	братолюбіє	Л	2	3,27868852459016E-02	0,1780783687082	5,43139024560011	2	2,988105	
315	Брут	й	3	4,91803278688525E-02	0,380949181594172	7,74596669241483	1	-2,220446E-15	
319	буває	Ь	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
320	бувало	Ь	5	8,19672131147541E-02	0,274314762798058	3,3466401061363	5	5,679506	
327	Будище	л	2	3,27868852459016E-02	0,253966121062781	7,74596669241483	1	1,110223E-15	
328	будовати	Г	1	1,63934426229508E-02	0,126983060531391	7,74596669241483	1	1,110223E-15	
339	бунчуг	Й	4	6,55737704918033E-02	0,399542872359146	6,09302880347697	2	2,133934	

ДОДАТОК 14. Фрагмент таблиці "morfems" (5076 - 5123)

morfems			
morfemid	wrdFK	mname	mseg
12336	5076	R	ревн
12337	5076	S	итель
12341	5078	R	рев
12342	5078	S	ну
12343	5078	F	ти
12344	5079	R	рев
12345	5079	F	ти
12346	5081	R	рев
12347	5081	S	уч
12348	5081	F	ий
12349	5083	R	регот
12350	5083	S	а
12351	5083	F	ти
12352	5083	X	ся
12353	5084	R	регот
12354	5084	S	н
12355	5084	F	я
12356	5086	R	рек
12357	5086	F	ти
12358	5087	R	ремиґ
12359	5087	S	а
12360	5087	S	ти
12361	5088	R	ремінь
12362	5089	R	репет
12363	5089	S	ува
12364	5089	F	ти
12365	5090	R	репјах
12366	5091	R	республік
12367	5091	F	а
12368	5092	R	ретязь
12369	5096	R	риб
12370	5096	S	ал
12371	5096	S	к
12372	5096	F	а
12373	5098	R	риб
12374	5098	S	к
12375	5098	F	а
12376	5099	R	риб
12377	5099	S	оньк

morfems			
morfemid	wrdFK	mname	mseg
12378	5099	F	а
12379	5100	R	риб
12380	5100	S	ч
12381	5100	S	ин
12382	5100	F	а
12383	5101	R	рид
12384	5101	S	а
12385	5101	F	ти
12386	5103	R	риз
12387	5103	F	а
12390	5105	R	рим
12391	5106	R	римл
12392	5106	S	ян
12393	5106	S	ин
12394	5107	R	рим
12395	5107	S	ськ
12396	5107	F	ий
12397	5109	R	ри
12398	5109	F	ти
12399	5113	R	ришт
12400	5113	S	ува
12401	5113	F	ти
12402	5114	R	рівн
12403	5114	F	о
12404	5115	R	рівн
12405	5115	F	я
12406	5116	R	ріг
12407	5117	R	рід
12408	5119	R	рід
12409	5119	S	н
12410	5119	F	я
12414	5122	R	різ
12415	5122	S	а
12416	5122	F	ти
12417	5122	X	ся
12418	5123	R	різ
12419	5123	S	ник

ДОДАТОК 15.
Фрагмент таблиці "morfemstat"

morfemstat			
Код	mname	mseg	freq
1	R	аби	10
2	X	-то	16
3	R	або	54
4	R	щ	87
5	F	о	1711
6	R	авраам	1
7	R	агар	1
8	S	ян	21
9	S	ськ	85
10	F	ий	3918
11	R	адам	1
12	P	а	2
13	R	кафіст	2
14	R	алілуја	7
15	R	алмаз	2
16	R	амінь	5
17	R	амфор	2
18	F	а	4539
19	R	ангел	6
20	S	ят	20
21	S	очк	23
22	R	а	14
23	R	ні	3
24	R	к	1
25	I	ог	1
26	S	ісіньк	12
27	R	хт	340
28	R	ну	30
29	S	мо	1
30	S	те	6
31	R	аполлон	2
32	R	апостол	14
33	R	апостоль	1
34	R	арал	2
35	R	арен	4
36	R	арій	1
37	R	армій	1
38	P	архі	3
39	R	јерей	1
40	R	мандрит	2
41	R	архімед	1
42	R	аршин	1

morfemstat			
Код	mname	mseg	freq
43	R	аул	1
44	R	афін	1
45	F	и	1157
46	R	ач	3
47	R	б	167
48	R	ба	11
49	R	баб	19
50	S	ус	35
51	S	еньк	200
52	F	я	1085
53	R	бав	4
54	S	и	2460
55	F	ти	11728
56	R	багат	57
57	S	ир	4
58	S	о	576
59	R	-багат	1
60	R	багатт	1
61	R	багн	5
62	R	баг	3
63	S	овинн	3
64	R	багр	4
65	S	иц	166
66	R	баж	5
67	S	а	3600
68	R	базар	21
69	R	байдак	13
70	R	байдар	1
71	R	байдуж	25
72	S	е	229
73	R	байкал	1
74	R	байрак	11
75	R	байстр	19
76	S	юк	17
77	R	бал	2
78	R	балак	2
79	R	балк	3
80	R	балкан	1
81	R	бандур	2
82	S	ист	4
83	R	банкет	1
84	R	банк	1
85	S	ір	70
86	R	барвін	6
87	R	ок	113

morfemstat			
Код	mname	mseg	freq
88	R	барвіноч	2
89	S	ок	338
90	R	барил	1
91	R	бар	12
92	X	ся	1836
93	R	бариш	1
94	R	баркас	1
95	R	бат	8
96	S	еч	67
97	S	к	885
98	S	іг	1
99	R	батіж	1
100	S	ур	6
101	S	ин	588
102	R	бать	121
103	S	ів	40
104	S	щ	4
105	R	бахур	1
106	R	бач	178
107	R	бг	4
108	R	бебех	1
109	R	бебр	1
110	R	бевзь	1
111	P	без	52
112	R	бож	80
113	S	н	899
114	R	верх	8
115	S	0	768
116	R	віч	7
117	R	голов	21
118	S	ј	113
119	R	дих	6
120	S	нн	9
121	R	дон	28
122	R	конеч	1
123	R	краї	12
124	F	ји	4
125	R	люд	306
126	R	пер	15
127	R	слав	163

ДОДАТОК 16. Таблиця "morfstructs"

morfstructs		
Код	mfstr	freq
4	RF	16323
11	RSF	7482
1	R	6605
17	PRSF	3804
6	PRF	1631
5	RSSF	898
14	RSFX	683
30	PRSFX	672
12	RS	604
34	PRS	499
33	PR	491
22	RSS	359
35	PRSS	254
18	PRSSF	212
43	PPRSF	186
47	PRFX	150
21	RIRSF	125
20	RFX	110
56	PRSSX	99
15	RSSSF	77
32	RIRF	72
7	RR	70
62	RSSFX	58
3	RRF	51
71	PRSSFX	48
2	RX	38
29	RISF	37
46	PRSSS	30

morfstructs		
Код	mfstr	freq
49	PPRF	27
10	RRS	27
40	RSSS	23
61	PPRSFX	22
19	PPRSSF	22
42	PRSSSF	19
92	RSX	18
36	PPRS	17
25	RRSF	17
16	RSSSSF	12
48	RSR	11
13	RSRS	10
24	RIRISF	9
90	RFPR	8
41	PPRSS	8
38	RSIRSF	7
72	PPR	6
70	PRISF	5
55	RIR	5
54	PRSX	5
28	RSSSS	5
80	PRIRF	4
57	RSISF	4
52	RISS	4
60	RISSF	3
58	PRIRS	3
45	PSSFX	3
37	RSRF	3
31	PPRFX	3
27	RIRSSSF	3

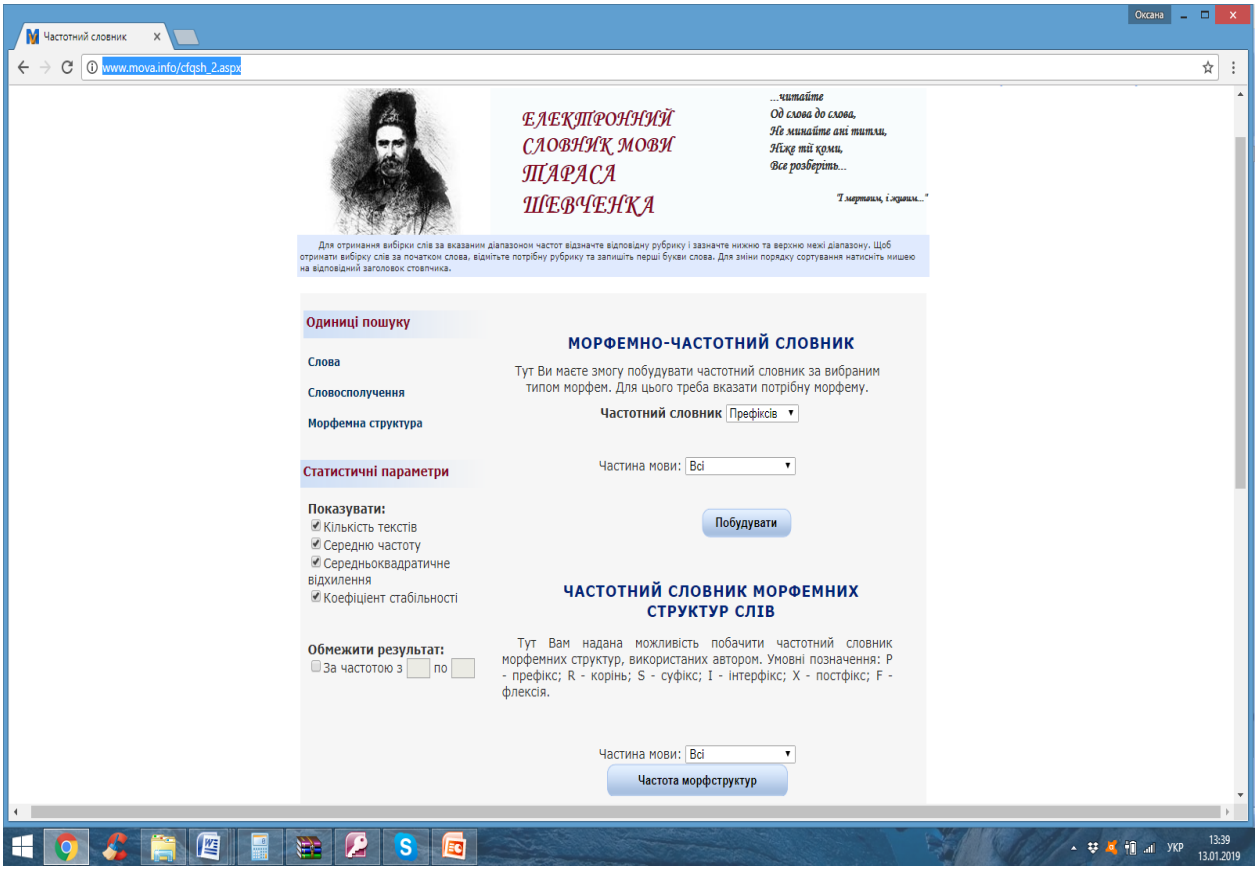
morfstructs		
Код	mfstr	freq
26	RRSSF	3
23	RIRSSF	3
87	RSFF	2
86	RIRS	2
81	PRIRSF	2
78	PRRSS	2
68	RSRSF	2
50	RIF	2
39	RSIRSSF	2
94	RSXX	1
93	RRR	1
89	PRSRS	1
84	RSPRSS	1
83	PRSSSS	1
82	PRSSSSF	1
79	RIPRSF	1
76	PRX	1
75	PRISSF	1
74	PPPRSF	1
69	PRRS	1
63	RSIRF	1
59	RSSX	1
53	RIS	1
51	RSSRSS	1
44	RRSS	1
9	RRRF	1
8	RRRISF	1

ДОДАТОК 17. Таблиця розшифрування граматичних кодів частин мови

classes	
sub	kont
?	невідоме
@	форма на -но/-то
А	прикметник
Б	скорочення
В	вигук
Г	дієсл.
Ґ	присудкова форма
Д	дієприсл.
И	ім. множ.
Й	ім. ч. р.
К	ім. ж. р.

classes	
sub	kont
Л	ім. с. р.
М	займ.-ім.
Н	прислівн.
О	займ.-прикм.
П	прийм.
С	сполучн.
Ф	не укр. алф.
Ч	числ.
Ъ	займ.
Ь	част.
Э	абревіатура

ДОДАТОК 18. Інтерфейс електронного словника мови Т. Шевченка



ДОДАТОК 19. Фрагмент кореневої вибірки таблиці "morfemstat" БД частотних словників

morfemstat			
Код	mname	mseg	freq
1600	R	не	1715
2657	R	ja	1051
924	R	з	776
246	R	бу	531
1488	R	м	518
534	R	вон	505
430	R	пі	380
562	R	ста	340
27	R	хт	340
480	R	він	335
393	R	ми	334
300	R	й	315
794	R	до	312
195	R	бог	306
125	R	люд	306

morfemstat			
Код	mname	mseg	freq
320	R	світ	287
811	R	зна	283
814	R	каз	280
169	R	да	280
768	R	див	273
2185	R	св	260
1452	R	мат	252
748	R	де	243
144	R	би	242
405	R	плак	238
1015	R	син	235
445	R	ход	235
2490	R	хат	231
356	R	в	229
2167	R	сам	209
463	R	т	209
805	R	дол	208
796	R	добр	193
1259	R	коз	191
1905	R	по	191
2641	R	ще	190

morfemstat			
Код	mname	mseg	freq
1420	R	люб	189
2204	R	сер	185
305	R	гор	181
1784	R	так	180
1740	R	пан	180
106	R	бач	178
775	R	дів	177
168	R	вол	170
808	R	жи	168
47	R	б	167
127	R	слав	163
1514	R	молод	153
301	R	гад	149
1031	R	тих	147
223	R	брат	145
1086	R	мал	145
2270	R	соб	145
1099	R	нов	137
1063	R	земл	133
1555	R	над	131
398	R	нес	130

ДОДАТОК 20. Фрагмент кореневої вибірки таблиці "morfems" БД частотних словників

morfems			
morfemid	wrdFK	mname	mseg
1	20	R	аби
2	21	R	аби
4	23	R	або
5	24	R	або
6	24	R	щ
8	26	R	авраам
9	27	R	агар
13	30	R	адам
15	33	R	кафіст
16	38	R	алілуя
17	41	R	алмаз
18	44	R	амінь
19	46	R	амфор
21	48	R	ангел
22	49	R	ангел
26	51	R	а
27	51	R	ні
28	52	R	а
29	52	R	ні
30	52	R	к
34	53	R	а
35	53	R	ні
36	53	R	хт
38	54	R	а
39	54	R	ну
40	55	R	а
41	55	R	ну
43	56	R	а
44	56	R	ну
46	57	R	аполлон
47	58	R	апостол
48	59	R	апостоль

morfems			
morfemid	wrdFK	mname	mseg
51	61	R	арал
52	62	R	арен
54	65	R	арій
56	66	R	армій
59	68	R	јерей
61	69	R	мандрит
62	70	R	архімед
63	72	R	аршин
64	75	R	аул
65	76	R	афін
67	77	R	ач
68	79	R	б
69	81	R	ба
70	82	R	баб
72	83	R	баб
76	84	R	баб
79	85	R	бав
82	86	R	багат
84	87	R	багат
86	88	R	багат
88	90	R	багат
90	90	R	-багат
92	91	R	багатт
94	92	R	багн
96	93	R	баг
99	94	R	багр
102	95	R	багр
106	96	R	баж
109	97	R	базар
110	99	R	байдак
111	100	R	байдар
113	101	R	байдуж
115	102	R	байкал
116	103	R	байрак
117	104	R	байстр
119	105	R	байстр
121	108	R	бал
122	109	R	балак
125	110	R	балк

morfems			
morfemid	wrdFK	mname	mseg
127	111	R	балкан
129	112	R	бандур
131	113	R	бандур
133	114	R	банкет
134	115	R	банк
136	117	R	барвін
137	117	R	ок
138	118	R	барвіноч
140	120	R	барил
142	121	R	бар
146	122	R	бариш
147	123	R	баркас
148	124	R	бат
152	125	R	бат
154	126	R	батіж
156	127	R	бат
159	128	R	бать
162	129	R	бать
165	130	R	бать
170	131	R	бать
176	132	R	бать
179	134	R	бахур
180	135	R	бач
181	136	R	бач
184	137	R	бач
188	143	R	бг
191	144	R	бебех
192	145	R	бебр
195	146	R	бевзь
197	148	R	бож
201	149	R	верх
205	150	R	віч

ДОДАТОК 21. Фрагмент конкорданса до слова *люди*

ЧАСТОТНИЙ СЛОВНИК ЗБІРКИ "ТВОРИ В П'ЯТИ ТОМАХ". ТАРАС ШЕВЧЕНКО

Конкорданс до слова: **люди** (ім. множ.)
Морфемна структура: **RF /люд/и/**

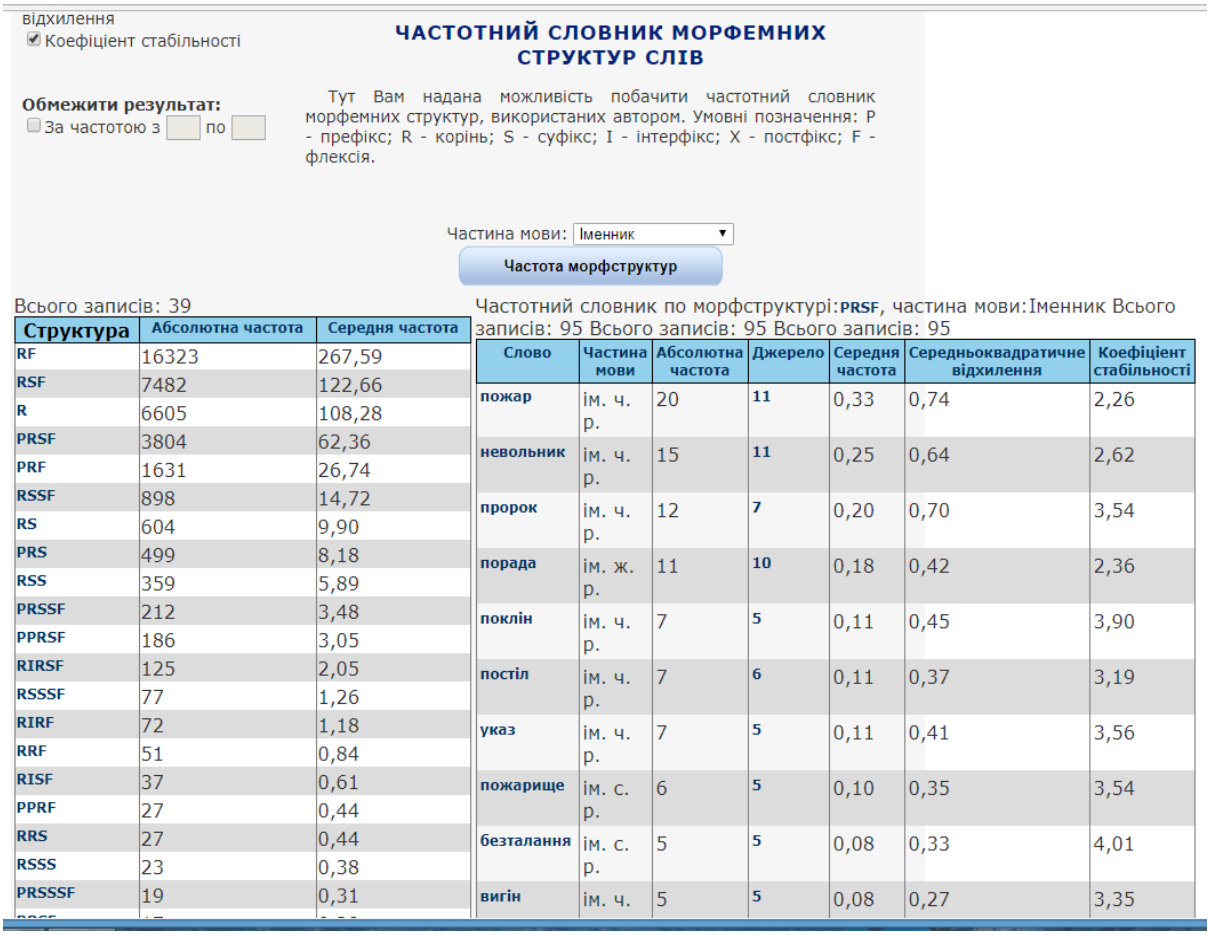
Контекст	Джерело
Пошли ж ти їй долю — вона молоденька , Бо люде чужії її засміють .	>>
На чужині не ті люде — Тяжко з ними жити !	>>
Кохайтеся , чорнобриві , Та не з москалями , Бо москалі — чужі люде , Роблять лихо з вами .	>>
Серце в'яне співаючи , Коли знає , за що ; Люде серця не побачать , А скажуть — ледащо !	>>
Кохайтеся ж , чорнобриві , Та не з москалями , Бо москалі — чужі люде , Згнущаються вами .	>>
Нехай собі тії люде , Що хотять , говорять : Вона любить , то й не чує , Що вкралося горе .	>>
Тойді Катерина Буде собі московкою , Забудеться горе ; А поки що , нехай люде , Що хотять , говорять .	>>
Батько , мати — чужі люде , Тяжко з ними жити !	>>
Пішла б в садок поплакати , Так дивляться люде , Зайде сонце — Катерина По садочку ходить , На рученьках носить сина , Очиці поводить : « Отут з муштри виглядала , Отут розмовляла , А там ... а там ... сину , сину !»	>>
Іди ж їх шукати , Та не кажи добрим людям , Що є в тебе мати .	>>
Іди доню , найди її , Найди , привітайся , Будь щаслива в чужих людях , До нас не вертайся !	>>
А до того — Московщина , Кругом чужі люде ... « Не потурай » , — може , скажеш , Та що з того буде ?	>>
А хто грає , того знають І дякують люде : Він їм тугу розганяє , Хоть сам світом нудить .	>>
Старий заховавсь В степу на могилі , щоб ніхто не бачив , Щоб вітер по полю слова розмахав , Щоб люде не чули , бо то Боже слово , То серце по волі з Богом розмовля , То серце щебече Господнюю славу , А думка край світа на хмарі гуля .	>>
І знову на небо , бо на землі горе , Бо на їй , широкій , куточка нема Тому , хто все знає , тому , хто все чує : Що море говорить , де сонце ночує — Його на сім світі ніхто не прийма ; Один він між ними , як сонце високе , Його знають люде , бо носить земля ; А якби почули , що він , одинокий , Співа на могилі , з морем розмовля , — На Божеє слово вони б насміялись , Дурним би назвали , од себе б прогнали .	>>
Ходи собі , мій голубе , Поки не заснуло Твоє серце , та виспівуй , Щоб люде не чули .	>>
Лягла спочить ... А тим часом Виросла могила , А над нею орел чорний Сторожем літає , І про неї добрим людям Кобзарі співають , Все співають , як діялось , Сліпі небораки , Бо дотепні ... А я ... А я	>>
А надто той , що дивиться На людей душою — Пекло йому на сім світі , А на тім ...	>>
Нехай думка , як той ворон , Літає та кряче , А серденько соловейком Щебече та плаче Нишком — люди не побачуть , То й не засміються ...	>>
Не щебече соловейко В лузі над водою , Не співає чорнобрива , Стоя під вербою , Не співає — сиротою Білим світом нудить : Без милого батько , мати — Як чужії люди , Без милого сонце світить — Як ворог сміється , Без милого скрізь могила ...	>>
Сама колись дівувала — Теє лихо знаю ; Минулося — навчилася : Людям помагаю .	>>
Що сміються люди , Скажи йому , що загину , Коли не прибуде !	>>

ДОДАТОК 22. Фрагмент списку джерел, в яких вживаються речення із словом *люди*

ЧАСТОТНИЙ СЛОВНИК ЗБІРКИ "ТВОРИ В П'ЯТИ ТОМАХ". ТАРАС ШЕВЧЕНКО

Джерело
ПРИЧИННА ("Реве та стогне Дніпр широкий...")
ДУМКА ("Тече вода в синє море...")
КАТЕРИНА ("Кохайтеся, чорнобриві...")
ДО ОСНОВ'ЯНЕНКА ("Б'ють пороги; місяць сходить...")
ПЕРЕБЕНДЯ ("Перебендя старий, сліпий...")
ТОПОЛЯ ("По діброві вітер виє...")
"Думи мої..."
Н. МАРКЕВИЧУ ("Бандуристе, орле сизий...")
МАР'ЯНА-ЧЕРНИЦЯ ("Вітер в гаї нагинає...")
УТОПЛЕНА ("Вітер в гаї не гуляє...")
ГАЙДАМАКИ ("Все йде, все минає - і краю немає...")
ТИТАР ("У гаю, гаю...")
СВЯТО В ЧИГИРИНІ ("Гетьмани, гетьмани, якби то ви встали...")
ТРЕТІ ПІВНІ ("Ще день Украйну катували...")
ГУПАЛІВЩИНА ("Зійшло сонце; Україна...")
ЕПІЛОГ ("Давно те минуло, як, мала дитина...")
"Вітер з гаєм розмовляє..."
ДІВИЧІЇ НОЧІ ("Розплелася густа коса...")
СОВА ("Породила мати сина...")
"У неділю не гуляла..."
43. "Боже, нашими ушима..."
52. "Пребезумний в серці каже..."
81. "Меж царями-судіями..."
93. "Господь Бог лихих карає..."
149. "Псалом новий Господеві..."
І МЕРТВИМ, І ЖИВИМ, І НЕНАРОЖДЕННИМ ЗЕМЛЯКАМ МОЇМ В УКРАЇНІ І НЕ В УКРАЇНІ МОЄ ДРУЖНЄЄ ПОСЛАНІЄ ("І смеркає, і світає...")
ХОЛОДНИЙ ЯР (У всякого своє лихо...)
МАЛЕНЬКІЙ МАР'ЯНІ ("Рости, рости, моя пташко...")
"Минають дні, минають ночі..."
ПРОЛОГ ("У неділю вранці-рано...")

ДОДАТОК 23. 1-ий екран інтерфейсу ЧС морфемних структур слів



відхилення
Коефіцієнт стабільності

ЧАСТОТНИЙ СЛОВНИК МОРФЕМНИХ СТРУКТУР СЛІВ

Обмежити результат:
За частотою з по

Тут Вам надана можливість побачити частотний словник морфемних структур, використаних автором. Умовні позначення: P - префікс; R - корінь; S - суфікс; I - інтерфікс; X - постфікс; F - флексія.

Частина мови: Іменник

Частота морфструктур

Всього записів: 39

Структура	Абсолютна частота	Середня частота
RF	16323	267,59
RSF	7482	122,66
R	6605	108,28
PRSF	3804	62,36
PRF	1631	26,74
RSSF	898	14,72
RS	604	9,90
PRS	499	8,18
RSS	359	5,89
PRSSF	212	3,48
PPRSF	186	3,05
RIRSF	125	2,05
RSSSF	77	1,26
RIRF	72	1,18
RRF	51	0,84
RISF	37	0,61
PPRF	27	0,44
RRS	27	0,44
RSSS	23	0,38
PRSSSF	19	0,31

Частотний словник по морфструктурі:PRSF, частина мови:Іменник Всього записів: 95 Всього записів: 95 Всього записів: 95

Слово	Частина мови	Абсолютна частота	Джерело	Середня частота	Середньоквадратичне відхилення	Коефіцієнт стабільності
пожар	ім. ч. р.	20	11	0,33	0,74	2,26
невольник	ім. ч. р.	15	11	0,25	0,64	2,62
пророк	ім. ч. р.	12	7	0,20	0,70	3,54
порада	ім. ж. р.	11	10	0,18	0,42	2,36
поклін	ім. ч. р.	7	5	0,11	0,45	3,90
постіл	ім. ч. р.	7	6	0,11	0,37	3,19
указ	ім. ч. р.	7	5	0,11	0,41	3,56
пожарище	ім. с. р.	6	5	0,10	0,35	3,54
безталання	ім. с. р.	5	5	0,08	0,33	4,01
вигін	ім. ч.	5	5	0,08	0,27	3,35

Наукове видання

ЗУБАНЬ Оксана Миколаївна

КОМП'ЮТЕРНЕ ЛЕКСИКОГРАФІЧНЕ МОДЕЛЮВАННЯ МОРФЕМНОЇ СИСТЕМИ УКРАЇНСЬКОЇ МОВИ

Монографія

Друкується за авторською редакцією

Оригінал-макет виготовлено ВПЦ "Київський університет"

Формат 60x84$^{1/16}$. Ум. друк. арк. 15,1. Наклад 100. Зам. № 220-9611
Гарнітура Nimes New Roman. Папір офсетний. Друк офсетний.
Підписано до друку 10.02.20

Видавець і виготовлювач
ВПЦ "Київський університет"
Б-р Тараса Шевченка 14, м. Київ, 01601
(38044) 239 32 22; (38044) 239 31 72; тел./факс (38044) 239 31 28
e-mail: vpc_div.chief@univ.net.ua; redaktor@univ.net.ua
http://vpc.univ.kiev.ua
Свідоцтво суб'єкта видавничої справи ДК № 1103 від 31.10.02