РОЗДІЛ VII
Новітні тенденції лексикографії

УДК 81’322

Оксана Зубань

Електронні частотні морфемні словники в Корпусі української мови

У статті описано частотні морфемні словники (www.mova.info), укладені автоматично на базі текстів Т. Шевченка в Корпусі української мови. Морфемний сегментатор українського тексту в Корпусі української мови – це система, на вході якої знаходяться лексеми аналізованого тексту, представлені у вигляді алфавітно-частотного словника, а на виході – ті ж самі лексеми, індексовані кодами граматичної належності до певної частини мови та розчленовані на морфи – кореневі й афіксальні – з відповідними індексами. Модуль морфемного аналізу в Корпусі української мови – це зручний лінгвістичний інструмент, який допоможе користувачеві в автоматичному режимі проводити дослідження з морфеміки та словотвору на базі величезного ілюстративного текстового матеріалу Корпусу української мови, що дозволить отримати нові знання про семантичну та формальну структуру українського слова, проводити різноманітні класифікаційні аналізи лексики за кількісно-морфними моделями; створювати кореневі, афіксальні та словотвірні словники різних стилів і дискурсів.

Ключові слова: морфемно-словотвірна база даних, морфемний аналіз, Корпус української мови, морфемний сегментатор української мови, електронний частотний словник.

Постановка наукової проблеми та її значення. Наукову значущість результатів дослідження визначає насамперед репрезентабельність матеріалу: чим більше мовних фактів, тим достовірніші спостережувані закономірності. Проте традиційна форма збирання і систематизації інформації (переважно паперова картотека та використання різноманітних паперових словників) сьогодні не задовольняє потреб дослідників-філологів. Потрібні нові інформаційні технології, які б оптимізували роботу дослідника. Тому в українському мовознавстві на сьогодні нагальна проблема укладання електронних лінгвістичних словників, які мають формат параметризованих електронних баз даних, оснащених пошуково-класифікаційними програмними аналізаторами, що забезпечують: ефективне й оперативне проведення лінгвістичного аналізу; можливість аналізу великих лексичних масивів; отримання точних формальних характеристик мовних одиниць різних рівнів як у системі мови, так і реляційно-функціональних особливостей їх у тексті.

Аналіз досліджень цієї проблеми. У сучасній комп’ютерній лінгвістиці досить плідно розвивається лексикографічний напрям. Зокрема, в українському мовознавстві сьогодні створено чимало різногалузевих електронних словників, з різноманітними системами навігацій та інтерфейсами, частина з яких стала вже лінгвістичним комерційним продуктом і доступна для широкого кола користувачів. У галузі морфеміки та словотвору плідно розвивається практика і теорія створення автоматизованих баз даних: Морфемно-словотвірний фонд української мови [5], створений у відділі структурно-математичної лінгвістики Інституту мовознавства ім. О. О. Потебні АН України; автоматизована система морфемно-словотвірного аналізу (АСМСА) [1], [2], створена в лабораторії комп’ютерної лінгвістики Інституту філології КНУ ім. Т. Шевченка.

Ці дві морфемно-словотвірні бази мають різні завдання і, відповідно, створені на основі різної методики. Морфемно-словотвірний фонд української мови – це база даних, спрямована на виконання функції своєрідного довідника для лінгвіста-дослідника, і, без сумніву, надзвичайно важлива для організації повномасштабного дослідження мови, але вона статична, її не можна використовувати в режимі автоматизованого аналізу тексту. На базі Морфемно-словотвірного фонду української мови було укладено відомі морфемні та словотвірні словники української мови, зокрема «Словник афіксальних

морфем української мови» [6], «Кореневий гніздовий словник української мови» [4], а також проведено велику кількість глибоких лінгвістичних досліджень, представлених у статтях та монографіях співробітників відділу структурно-математичної лінгвістики. Автоматизована система морфемно-словотвірного аналізу (АСМСА) – це електронний лінгвістичний продукт, спрямований на аналіз тексту і має статус динамічної пошукової системи, що здатна в автоматичному або автоматизованому режимі вилучати інформацію про морфемні та словотвірні одиниці мови з будь-якого параметризованого тексту, представленого в Корпусі української мови (www.mova.info).

Мета – описати методику укладання та структуру електронних частотних морфемних словників, автоматично укладених на матеріалі текстів Т. Шевченка в Корпусі української мови.

Виклад основного матеріалу й обґрунтування отриманих результатів дослідження. У межах наукового проекту «Корпус української мови» колектив лабораторії комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка розробив методику комп'ютерного моделювання структурних відношень одиниць різних мовних рівнів, на основі якої було створено комп'ютерні інструменти – пакети програм, що забезпечують отримання лінгвістичної інформації з тексту і створення різноманітних електронних словників та картотек в автоматичному або автоматизованому режимі.

Корпус українських текстів має модульну будову:

1. Модуль-текст, який репрезентує корпус параметризованих текстів (30 млн слововживань).

2. Модуль-аналізатор – інструмент лінгвістичних досліджень великих текстових масивів, що містить пакети програм, які можуть виконувати такі функції: забезпечують зв'язок корпусу текстів із лінгвістичними базами даних (морфологічною, морфемно-словотвірною, синтаксичною); забезпечують роботу в автоматичному режимі пошуку і класифікації лексики за різними параметрами, а також проведення морфологічного, морфемного і статистичного аналізів; формують словник-конкорданс контекстових слововживань.

3. Модуль-словник, у якому систематизується різнотипна лінгвістична інформація внаслідок аналізу текстів.

Структура Корпусу українських текстів відображає таку логіку лінгвістичного аналізу: модуль-текст – текстовий матеріал дослідження, на базі якого за допомогою модуля-аналізатора в автоматичному або автоматизованому режимі укладають різноманітні лінгвістичні словники (модуль-словник).

Одним із модулів автоматичного аналізу тексту Корпусу є морфемний модуль, що виконує функцію автоматичного морфемного сегментування словоформ тексту.

Автоматичний морфемний аналіз у Корпусі української мови – це зручний комп'ютерний інструмент, який на базі великого масиву текстового матеріалу може здійснювати різноманітні завдання:

1) автоматичне морфемне сегментування початкових словоформ;

2) автоматичне групування лексики у спільнокореневі та спільноафіксальні вибірки;

3) автоматичне групування лексики за моделями морфемних структур слів у спільноструктурні вибірки;

4) автоматичні статистичні обчислення одиниць морфемного рівня тексту.

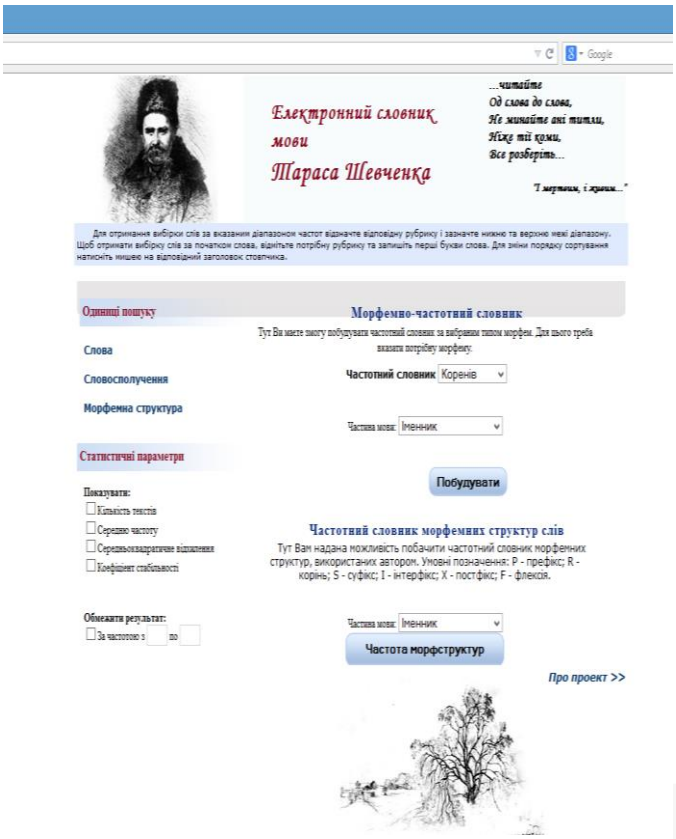
Перше завдання виконується з метою визначення морфемної будови початкових форм слововживань тексту. На етапі морфологічного модуля всі слововживання тексту лематизуються у початкові форми і потрапляють у морфемний модуль у вигляді реєстру початкових форм із визначеною інформацією про частиномовну належність. Морфемний сегментатор українського тексту – це система, на вході якої знаходяться лексеми (початкові словоформи) аналізованого тексту, представлені у вигляді алфавітно-частотного словника, а на виході – ті ж самі лексеми, індексовані кодами граматичної належності до певної частини мови та розчленовані на морфи – кореневі, афіксальні – з відповідними індексами. Після завершення процедури морфного сегментування словоформ морфна структура слова запам'ятовується у символах лінгвістичної моделі, яка визначає тип і межі кожного морфа словоформи, дозволяє автоматично описати кожну морфну структуру через програмну процедуру. Морфемна будова кожної початкової форми слова автоматично моделюється за допомогою латинських літер, що символізують функціональні типи морфем: P – префікс, R – корінь, S – суфікс, F – флексія, I – інтерфікс, X – постфікс. Таким чином, кожній лексемі (початковій словоформі) тексту автоматично приписується кількісно-функціональна модель морфемної структури слова, наприклад: RIRSF – *льод-о-різ-∅-∅*. Автоматичне визначення морфемної структури слова у Корпусі здійснюється за процедурою автоматичного зіставлення з лінгвістичною моделлю морфемної структури слова в Морфемно-словотвірній базі даних (МСБД), де за загально-

прийнятими теоретичними принципами методики морфемного аналізу І. Т. Яценка [7] в автоматизованому режимі просегментовано ≈ 190 тис. початкових словоформ і визначено модель морфемної структури кожного слова.

Наступні три завдання виконуються з метою укладання морфемних частотних словників. На сьогодні у Корпусі (www.mova.info) укладено два типи частотних морфемних словників поетичного мовлення Лесі Українки, Л. Костенко, В. Стуса, М. Матіос та Т. Шевченка: 1) Частотні словники кореневих та афіксальних морфем; 2) Частотні словники морфемних структур слів.

У статті представлено морфемні частотні словники поетичного мовлення Т. Шевченка, автоматично укладені на базі ≈ 80 тис. слововживань поетичних текстів Т. Шевченка.

Ці словники структуровані на три зони, поєднані зручною навігацією: 1. Інвентар одиниць (морфструктур, коренів, афіксів). 2. Реалізація морфструктури / морфеми у словах тексту за такими характеристиками: а) інвентар і кількість слів (лексем) кожної морфструктури, кореня, афікса; б) частиномовна характеристика слова; в) абсолютна частота вживання слова (кількість слововживань лексеми у текстах); г) кількість текстів (творів), у яких уживається слово; ґ) середня частота слова; д) середньоквадратичне відхилення; е) коефіцієнт стабільності; 3. Контексти вживання аналізованого слова.



Частотний словник морфем

1 ЗОНА. Інвентар одиниць частотного словника морфем укладають автоматично за вибором користувача. У випадних списках користувач за двома фільтрами: 1) тип морфеми; 2) частина мови; може вибрати:1) одиниці аналізу, у нашому випадку – корені (рис. 1); 2) та морфологічне поле вибірки слів: всі слова тексту або слова окремої частини мови, у демонстраційному варіанті – іменники.

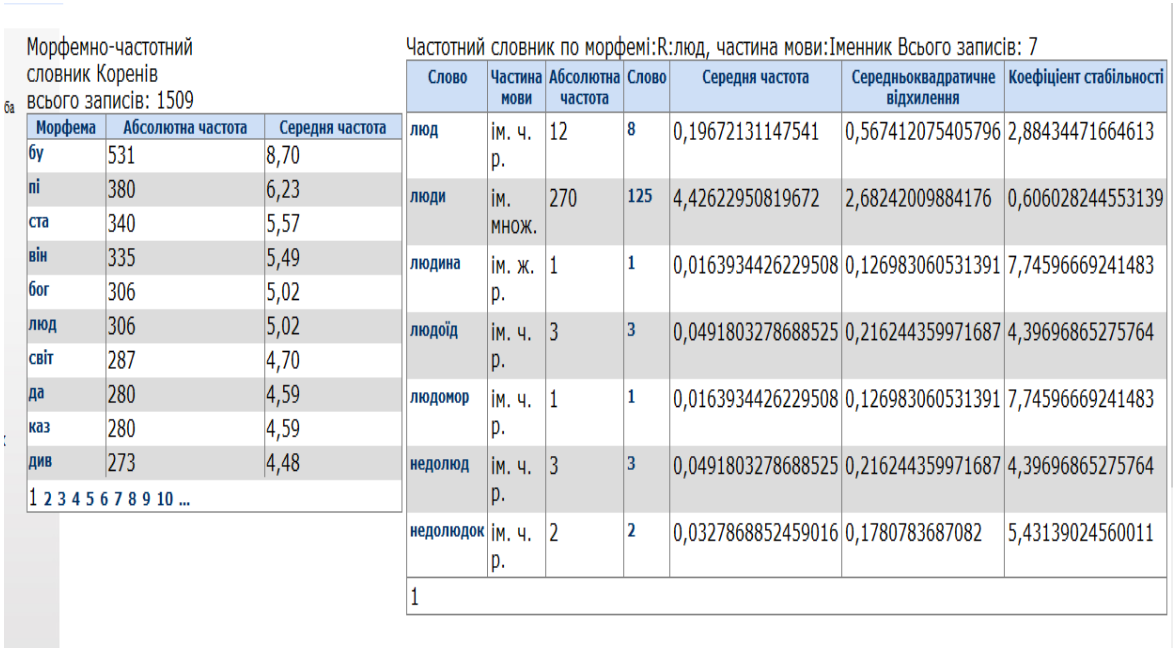
Фрагмент частотного словника коренів (рис. 2) показує, що словник коренів іменникової лексики поетичного мовлення Т. Шевченка становить 1509 коренів з інформацією про абсолютну та середню частоту вживання кожного кореня у текстових слововживаннях (в усіх слововживаннях тексту без обмеження частиномовної характеристики слів). Формування інвентаря коренів можливе за спадом або за зростанням абсолютних частот.

Морфемно-частотний словник Коренів
всього записів: 1509

Морфема	Абсолютна частота	Середня частота
бу	531	8,70
пі	380	6,23
ста	340	5,57
він	335	5,49
бог	306	5,02
люд	306	5,02
світ	287	4,70
да	280	4,59
каз	280	4,59
див	273	4,48

Рис. 2. **Фрагмент морфемно-частотного словника коренів**

2 ЗОНА. У другій зоні словника подано лексичну реалізацію конкретного кореня (рис. 3), вибраного в першій зоні: фрагмент демонструє реалізацію частотного кореня *-люд-* (з абсолютною частотою 306 слововживань) в іменникових слововживаннях тексту. Цей корінь реалізується у 7-ми словах із різною продуктивністю. Найчастотнішим є слово *люди*, яке реалізоване в поетичному мовленні у 270 слововживаннях тексту.



Морфемно-частотний словник Коренів всього записів: 1509

Морфема	Абсолютна частота	Середня частота
бу	531	8,70
пі	380	6,23
ста	340	5,57
він	335	5,49
бог	306	5,02
люд	306	5,02
світ	287	4,70
да	280	4,59
каз	280	4,59
див	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	Слово	Середня частота	Середньоквадратичне відхилення	Коефіцієнт стабільності
люд	ім. ч. р.	12	8	0,19672131147541	0,567412075405796	2,88434471664613
люди	ім. множ.	270	125	4,42622950819672	2,68242009884176	0,606028244553139
людина	ім. ж. р.	1	1	0,0163934426229508	0,126983060531391	7,74596669241483
людоїд	ім. ч. р.	3	3	0,0491803278688525	0,216244359971687	4,39696865275764
людомор	ім. ч. р.	1	1	0,0163934426229508	0,126983060531391	7,74596669241483
недолюд	ім. ч. р.	3	3	0,0491803278688525	0,216244359971687	4,39696865275764
недолюдок	ім. ч. р.	2	2	0,0327868852459016	0,1780783687082	5,43139024560011

1

Рис. 3. **Лексична реалізація кореня**

З ЗОНА. У третій зоні подають контексти вживання вибраного в реєстрі 2-ї зони слова. До кожного фрагмента тексту подано джерело – твір, із якого взята цитата. Навігація до цієї інформації здійснюється через активізацію позначки – >>.

Біліє місяць ; люде сплять , І титар спить ...	>>
Із-за лісу , з-за туману Місяць випливає , Червоніє , круглолиций , Горить , а не сяє , Неначе зна , що не треба Людям його світу , Що пожари Україну Нагріють , освітять .	>>
Людей не чуть ; через базар Кажан костокрилий Перелетить ; на вигоні Сова завиває .	>>
А де ж люде ?..	>>
Та ні , то люде гомонять Старшина	>>
А тим часом місяць пливе оглядать І небо , і зорі , і землю , і море , Та глянуть на люде , що вони моторять , Щоб Богові вранці про те розказать .	>>
Кругом його Мов вимерли люде .	>>
А за віщо , За що люде гинуть ?	>>
Ходив я , та плакав , та людей шукав ,	>>
Що письменні люде	>>
Їх люде минули ;	>>
Вибачайте , люде добрі ,	>>
Погаються добрі люди , Як холодні хвилі , Потім собі подивляться , Як сирота плаче , Потім спитай , де сирота , — Не чув і не бачив .	>>
А мені ще й завидують , Гордою і злою Злії люди нарікають .	>>
Дивуйтеся , люди !	>>

1 2 3 4 5 6

Рис.4. **Конкорданс до лексеми *люди***

Частотний словник морфемних структур слів

У 1-й зоні частотного словника морфемних структур користувач вибирає лише морфологічне поле вибірки слів, у демонстраційному фрагменті – іменник (рис. 1). Автоматично укладається інвентар із 39 морфемних структур іменникової лексики з інформацією про абсолютну та середню частоту вживання кожної морфструктури в текстових слововживаннях (у всіх слововживаннях тексту без обмеження частиномовної характеристики слова). Формування інвентаря морфструктур можливе за спадом або за зростанням абсолютних частот. Наприклад, морфструктура PRSF є продуктивною і реалізується у 3804 слововживаннях тексту.

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
RRSF	17	0,28

Частотний словник по морфструктурі:PRSF, частина мови:Іменник Всього записів: 95 Всього записів: 95 Всього записів: 95

Слово	Частина мови	Абсолютна частота	Слово	Середня частота	Середньоквадратичне відхилення	Коефіцієнт стабільності
пожар	ім. ч. р.	20	11	0,327868852459016	0,740432117418768	2,25831795812724
невольник	ім. ч. р.	15	11	0,245901639344262	0,644160088668475	2,61958436058513
пророк	ім. ч. р.	12	7	0,19672131147541	0,697058741818665	3,54338193757822
порада	ім. ж. р.	11	10	0,180327868852459	0,424966603174286	2,35663298123922
поклін	ім. ч. р.	7	5	0,114754098360656	0,447153498261831	3,89662334199596
постіл	ім. ч. р.	7	6	0,114754098360656	0,366568520901605	3,1943828249997
указ	ім. ч. р.	7	5	0,114754098360656	0,408851275847211	3,56284683238284
пожарище	ім. с. р.	6	5	0,0983606557377049	0,348529370909333	3,54338193757822
безталання	ім. с. р.	5	5	0,0819672131147541	0,328687502553499	4,00998753115268
вигін	ім. ч. р.	5	5	0,0819672131147541	0,274314762798058	3,3466401061363

Рис. 5. **Інвентар морфструктур та лексичне наповнення морфструктури PRSF**

У другій зоні словника морфструктур подано лексичну реалізацію вибраної у першій зоні морфемної структури, наприклад, морфструктура PRSF реалізована у 95 іменниках текстів Т. Шевченка. Кожен іменник із морфемною структурою PRSF характеризується своєю абсолютною частотою вживання. Одним із найчастотніших є іменник *по-рад-?-а* з абсолютною частотою 11 слововживань.

У третій зоні словника морфемних структур подаються контексти вживання вибраного в реєстрі 2-ї зони слова: 11 слововживань слова *порада* представлено в 11-ти контекстах. Як і в попередньому словнику, до кожного фрагмента тексту подано джерело – твір, із якого взята цитата. Навігація до цієї інформації здійснюється через активізацію позначки – >>.

Висновки та перспективи подальшого дослідження. Вивчення статистичних параметрів морфемного рівня текстів у Корпусі української мови відкриває широкі можливості й перспективи для глибоких стилеметричних розвідок у вивченні різних функціональних стилів та ідіостилів. Зокрема, дослідження статистичних параметрів морфемного рівня організації поетичного тексту Т. Шевченка [3] показують, що кількісні та статистичні характеристики морфемних структур слів, які формують відносно невеликий інвентар одиниць, виявляють закономірності будови тексту ідіостилю поета. Електронні частотні морфемні словники такого типу можуть бути укладені для всіх авторів і тестів Корпусу української мови за запитом користувача, якому потрібно отримати статистичну інформацію для різноманітних стилеметричних досліджень.

Джерела та література

1. Алексієнко Л. А. Методика створення автоматизованої системи морфемно-словотвірного аналізу (АСМСА) слів української мови / Л. А. Алексієнко, Н. П. Дарчук, О. М. Зубань // Наукова спадщина професора С. В. Семчинського : зб. наук. пр. – К. : [б. в.], 2001. – Ч. 1. – С. 38–49.
2. Дарчук Н. П. Комп'ютерне анотування тексту: результати і перспективи : монографія / Н. П. Дарчук. – К. : [б. в.], 2013. – 543 с.
3. Зубань О. М. Особливості морфемної будови слів у поетичних текстах Т. Шевченка (на матеріалі Корпусу української мови) / О. М. Зубань // Українське мовознавство : міжвідомчий наук. зб. – К. : [б. в.], 2014. – С. 123–133.

4. Карпіловська Є. А. Кореневий гніздовий словник української мови / Є. А. Карпіловська. – К. : Укр. енцикл., 2002. – 683 с.
5. Клименко Н. Ф. Морфемно-словотвірний фонд української мови як дослідницька та інформаційно-довідкова система / Н. Ф. Клименко, Є. А. Крапіловська, Л. І. Комарова, Т. І. Недозим, Т. В. Іванова // Клименко Н. Ф. Вибрані праці. – К. : [б. в.], 2014. – С. 545–558.
6. Словник афіксальних морфем української мови / за ред. Н. Ф. Клименко ; [укл.: Н. Ф. Клименко, Є. А. Карпіловська, В. С. Карпіловський, Т. І. Недозим]. – К. : [б. в.], 1998. – 434 с.
7. Яценко І. Т. Морфемний аналіз : Словник-довідник / І. Т. Яценко. – К. : [б. в.], 1980. – 1981. Т. 12.

Зубань Оксана. Электронные частотные морфемные словари в Корпусе украинского языка. В статье анализируются частотные морфемные словари (www.mova.info), составленные автоматически на базе текстов Т. Шевченко в Корпусе украинского языка. Морфемный сегментатор украинского текста в Корпусе украинского языка – это система, на входе которой находятся лексемы текста, представленные в алфавитно-частотном порядке, а на выходе те же лексемы, индексированные кодами части речи и сегментированные на морфы (корневые, аффиксальные) с соответствующими индексами. Модуль морфемного анализа в Корпусе украинского языка – это удобный лингвистический инструмент, который поможет пользователю в автоматическом режиме проводить исследования в области морфемики и словообразования на базе большого текстового материала Корпуса украинского языка, получать новые данные о семантической и формальной структуре украинского слова, проводить разнообразные классификации лексики по количественно-морфным моделям, а также создавать частотные корневые, аффиксальные словари разных стилей и дискурсов.

Ключевые слова: морфемно-словообразовательная база данных, морфемный анализ, Корпус украинского языка, морфемный сегментатор украинского текста, электронный частотный словарь.

Zuban Oksana. Electronic Dictionary of Frequency of Morphemic in the Corpus of the Ukrainian Language. The article presents Electronic dictionary of frequency of morphemic (www.mova.info), which was automatically constructed based on texts by T. Shevchenko in the Corpus of the Ukrainian language. The morphic segmentator of the Ukrainian text – is a system, on the input of which there are lexemes (or word forms) of an analysed text. They are presented in a form of an alphabetic-frequency dictionary. On its output there are the same lexemes (word forms)that are index-linked by the codes of grammatical belonging to a definite part of speech and are split into morphs – root morphs, affixal morphs with a proper index. The morphemic analys of the Corpus of the Ukrainian language – is a convenient linguistic tool, which in an online mode helps the user to carry out the research on the study of morphemes and derivation on the basis of a great number of illustrative textual materials of the corpus of the Ukrainian language, which enables to get new knowledge about the semantic and formal structure of the Ukrainian word. It also enables: to carry out various classifying analyses of vocabulary according to the quantitative and morphic models; to form root dictionaries, dictionary of affixes and derivational dictionaries of different styles and discourses.

Key words: Morphemic-Derivational Data Base, morphemic analys, Corpus of the Ukrainian language, the morphic segmentator of the Ukrainian text, Electronic dictionary of frequency.

Стаття надійшла до редколегії
12.03.2015 р.

УДК 81’22:81’42

Тетяна Приступа

Положення прагматичної лінгвістики у проекції на двомовні галузеві словники

Актуальність статті зумовлена потребою створити методологічну базу укладання двомовних галузевих словників, а саме українсько-іноземних, з урахуванням практичних та теоретичних результатів різних лінгвістичних дисциплін. Мета і завдання роботи – висвітлити основні положення прагматичної лінгвістики у проекції на проблеми лексикографії. У статті проаналізовано наукову літературу з лексикографії та прагматичної лінгвістики, висвітлено основні принципи лінгвопрагматики, визначено типи та специфіку прагматичної інформації, схарактеризовано прагматичні значення та розкрито, як вони можуть бути застосовані у процесі укладання двомовних галузевих словників, визначати їх макро- і мікроструктуру, наповнення, формат. Вимога досліджувати живу мову в її дії спричиняє потребу враховувати прагматичну інформацію під час укладання словників.