УДК 81

**І. Волошиновська**
Національний університет "Львівська політехніка"

# ЛЕКСИКОГРАФІЧНА ОБРОБКА* ТЕКСТОВИХ ДАНИХ ЯК ЗАСІБ ВИЗНАЧЕННЯ СПРЯМОВАНОСТІ ТЕКСТІВ



**Виконано лексикографічний аналіз текстів з прикладної лінгвістики українською та англійською мовами. Визначено найчастотнішу лексику у цих текстах. Зіставний аналіз виділеної лексики дав змогу зробити висновок стосовно різної предметної спрямованості текстів. Відмінність ключових слів українських та англійських текстів вказує на багатоманітність завдань прикладної лінгвістики.**

**The lexicographical analysis of applied linguistics texts in Ukrainian and English languages have been performed. The most frequently used lexical units of these texts have been determined. The comparative analysis of detailed lexic allows to make a conclusion about different object directions of the texts. The distinction in the set of the most frequently used words in Ukrainian and English texts points out variety of applied linguistics objectives.**

Мета даної роботи – виявити різницю між завданнями та цілями прикладної лінгвістики в українському та британському виданнях, застосовуючи лексикографічну обробку* текстових даних. Передбачалось з'ясувати: який зміст стоїть за визначенням терміну "прикладна лінгвістика"; в якій галузі і над якою тематикою працюють лінгвісти з різних країн; чи перетинаються інтереси науковців. Це порівняння проводилось на основі статистичного аналізу, який дає змогу виявити найчастіше вживані слова в тексті. Частотний аналіз лінгвістичних текстів дозволить в майбутньому розробити українсько-англійський словник-мінімум найвживаних слів із прикладної лінгвістики.

Щоб зрозуміти завдання дослідження, треба спочатку визначити, що розуміють під самим терміном "прикладна лінгвістика". Цей термін було офіційно визнано в Мічиганському університеті в 1946 р. На початковій стадії термін вживався в США та Великобританії для позначення перших спроб розробки** наукових підходів до вивчення іноземних мов, розглядаючи англійську як іноземну. У кінці 1950-их на початку 1960-их рр. застосування терміна розширилось, і сюди почали також долучати "автоматизований переклад". Перший міжнародний конгрес із прикладної лінгвістики було проведено в 1964 р., де доповіді стосувались двох чітко визначених напрямків – вивчення іноземної мови та автоматизованого перекладу. Зараз прикладна лінгвістика зосереджується навколо всіх сфер людської діяльності, в яких мова відіграє певну роль. Прикладна лінгвістика швидко

---

* опрацьовування, опрацювання; обробляння, оброблення – *ред.*
** розробляння, розроблення – *ред.*

перетворилася в міждисциплінарну галузь, яка пов'язана з психологією, соціологією, історією, літературою, математичними та комп'ютерними науками.

Математичні та комп'ютерні науки привнесли в лінгвістику статистичні методи для лексикографічної обробки* текстів. Лексикографічна обробка* текстових даних передбачає створення частотних і алфавітно-частотних словників, слововказівників, словників морфем, словників-мінімумів, словників ключових слів тощо. Важливий етап в цьому дослідженні – правильно сформувати вибірку, тобто встановити лінгвістичну та статистичну однорідність, здійснити пошук текстових одиниць, які володіють переліком характеристик для вирішення стилістичних і граматичних проблем тощо [1].

Згідно із законом переваги Дьюї мова і мовлення надають перевагу невеликій кількості одиниць, які часто використовуються і становлять ядро будь-якої мовної чи овленнєвої підсистеми, тоді як кількісно переважають низькочастотні одиниці [2]. Це явище можна спостерігати і під час обробки* матеріалу з певної тематики. Відповідно до спрямування тексту спостерігатиметься зміна складу частотного словника.

Усі рівні мовної системи підвладні дії статистичних законів, підраховувати можна одиниці будь-якого рівня – фонеми, звуки, літери, склади, слова, словосполуки тощо. Важливу умову вибору одиниць підрахунків становить вимога їх точного структурного визначення, що спирається на ознаки і критерії, які не вносять двозначності в їх трактування і доступні для перевірки.

Текст – це високоорганізоване явище, у якому всі одиниці й категорії переплітаються між собою в системі різного роду залежностей, взаємозумовленостей і взаємовиключень. Тому, щоб судити про текст як про ціле, необхідно вивчати ті статистичні залежності, які існують між частотами і розподілами різних одиниць.

Для проведення частотного аналізу текстів використовувалася програма, написана на мові С++. Вона дає змогу обробляти масиви розміром до 100 000 слів, видає інформацію про кількість слів у тексті, кількість слів, яких ужито один раз, а також частоту вживання того чи іншого слова у процентах та за абсолютною величиною. Програма дозволяє** сортувати слова за частотним фактором, а також в алфавітному порядку.

Статистичний аналіз слів проводився*** на матеріалі текстів англійською мовою ("Journal of Applied Linguistics" [3]; загальна кількість слововживань – 61657, кількість різних слововживань – 6948) та українською мовою ("Нариси з комп'ютерної лінгвістики" [4]; загальна кількість слововживань – 51434, кількість різних слововживань – 12898). Вибір цих джерел зумовлено робочою гіпотезою про спільність ключової лексики у текстах з однаковим фаховим спрямуванням.

**Результати статистичного аналізу текстів**

На перших позиціях за результатами проведеного статистичного аналізу для заданих текстів стоять сполучники, частки, прислівники, прийменники. Такого результату очікувано. Проте ці одиниці не несуть термінологічного навантаження, і тому їх далі не аналізували. Решта найвживаніших слів розташовано в Таблиці 1 згідно із частотою їх появи у тексті. Саме аналіз цих слів використовуватиметься для вивчення лінгвістичної термінології.

---

* опрацьовування, опрацювання; обробляння, оброблення – *ред.*

** дає змогу – *ред.*

*** виконано – *ред.*

Виділивши та проаналізувавши двадцять найвживаніших слів (таблиця 1), можна стверджувати, що серед ключових слів досліджуваних текстів є такі спільні: "language" – "мова", "English" – "Українська". Навіть у невідомому тексті за наявності таких ключових слів можна зробити висновок, що ці тексти – мовного спрямування і стосуються вивчення відповідної мови. Усі інші слова мають різне значення і показують різницю у спрямуваннях текстів, незважаючи на те, що це тексти з одної галузі.

У переліку найчастіше вживаних слів українського тексту присутні: "слово", "мова", "текст", "частина", "одиниця", "структура", "граматика". Перелік цих слів указує, що текст присвячений вивченню будови речення в мові. Такі слова, як "словник", "слово", "значення", "семантичний", "термінологічний", "автоматизований" стосуються побудови термінологічних та комп'ютерних словників.

З аналізу слів англійського тексту видно, що до найчастіше вживаних слів відносяться: "learner", "learning", "social", "practice", "applied". Цей набір слів дозволяє* зробити висновок, що досліджуваний англійський текст із прикладної лінгвістики пов'язаний з методами вивчення іноземної мови.

*Таблиця 1*

**Список найчастіше вживаних українських та англійських слів у тестах двох видань з прикладної лінгвістики**

| Ключові слова лінгвістичних текстів англійською мовою | Кількість слововживань / Частота вживання*100 | Ключові слова лінгвістичних текстів українською мовою | Кількість слововживань / Частота вживання*100 |
|---|---|---|---|
| language | 862/1,40 | словник | 776/1,51 |
| learner | 396/0,642 | слово | 606/1,18 |
| research | 235/0,381 | мова | 518/1,00 |
| social | 226/0,367 | лексика (лексична) | 467/0,908 |
| performance | 171/0,277 | українська | 271/0,537 |
| linguistics | 161/0,261 | значення | 266/0,517 |
| model | 143/0,232 | текст | 250/0,486 |
| english | 137/0,222 | частина | 227/0,441 |
| practice | 136/0,221 | система | 211/0,410 |
| applied | 130/0,211 | реєстрова | 201/0,391 |
| work | 130/0,210 | одиниця | 197/0,383 |
| formulaic | 125/0,203 | форма | 169/0,329 |
| university | 120/0,195 | структура | 149/0,290 |
| learning | 119/0,193 | семантична | 144/0,280 |
| development | 117/0,190 | стаття | 143/0,278 |
| example | 117/0,190 | термінологічний | 139/0,270 |
| resources | 108/0,175 | автоматизований | 132/0,256 |
| interaction | 104/0,169 | граматика | 131/0,254 |

Зіставляючи переліки найчастіше вживаних слів, виявлено, що зі 100 слів лише 17 спільні (у таблиці 2 наведено список найчастіше вживаних спільних слів в українській та англійській мовах, їх позиція та частота вживання у тексті). Однак частота їх вживання в англійських та українських текстах суттєво різна. Так, слово "linguistics" займає 6-ту позицію з частотою вживання – 0,26%, а еквівалентне йому "лінгвістика" – тільки 67-му позицію з частотою – 0,07%. Слова "meaning" та "значення" мають такі частоти вживання –

0.15% та 0.52% 0.15%, відповідно. Ці результати ще більшою мірою підтверджують попередній висновок про різну спрямованість досліджуваних лінгвістичних текстів. Отримані результати показують різні тенденції, які реалізуються в межах єдиного підходу – прикладна лінгвістика.

*Таблиця 2*

**Список найчастіше вживаних спільних українських та англійських слів у тестах двох видань з прикладної лінгвістики**

| № | Ключові слова лінгвістичних текстів англійською мовою | Позиція | Кількість слововживань / Частота вживання *100) | Ключові слова лінгвістичних текстів українською мовою | Позиція | Кількість слововживань / Частота вживання *100) |
|---|---|---|---|---|---|---|
| 1. | language | 1 | 862/1,40 | мова | 3 | 518/1,00 |
| 2. | research | 3 | 235/0,381 | дослідження | 58 | 43/0,083 |
| 3. | performance | 5 | 171/0,277 | представлення | 69 | 35/0,068 |
| 4. | linguistics | 6 | 161/0,261 | лінгвістика | 67 | 35/0,068 |
| 5. | english | 8 | 137/0,222 | українська | 5 | 271/0,528 |
| 6. | work | 11 | 130/0,210 | робота | 68 | 35/0,068 |
| 7. | learning | 14 | 119/0,193 | вивчення | 69 | 33/0,064 |
| 8. | development | 15 | 117/0,189 | розвиток | 42 | 66/0,128 |
| 9. | knowledge | 23 | 100/0,162 | знання | 66 | 37/0,072 |
| 10. | meaning | 28 | 90/0,145 | значення | 6 | 266/0,517 |
| 11. | acquisition | 29 | 88/0,143 | сприймання | 51 | 56/0,108 |
| 12. | analysis | 35 | 78/0,127 | аналіз | 54 | 53/0,103 |
| 13. | communicative | 36 | 78/0,127 | спілкування | 61 | 59/0,116 |
| 14. | term | 53 | 52/0,084 | термін | 18 | 86/0,167 |
| 15. | word | 58 | 49/0,079 | слово | 2 | 606/1,06 |
| 16. | grammar | 61 | 47/0,076 | граматика | 20 | 131/0,254 |
| 17. | understanding | 62 | 47/0,076 | розуміння | 75 | 32/0,062 |

Щоб впевнитись в отриманих результатах, було проведено ще одне зіставлення текстів, але на цей раз обидва тексти були на українській мові і відносились до галузі прикладної лінгвістики. Було проаналізовано такі джерела: М. М. Пещак "Нариси з комп'ютерної лінгвістики" (загальна кількість слововживань: 51434, кількість різних слововживань: 12898) [4], З. В. Партико "Образна концепція теорії інформації" (загальна кількість слововживань: 22382, кількість різних слововживань: 5404) [5]. Таких найчастіше вживаних у [4] слів, як "реєстрова", "автоматизований", "відмінок", "правопис", "флексія", "висловлювання", "комунікативний", "іншомовний", нема в тексті [5]. І навпаки, слів "сентенція", "номен", "реципієнт", "кібернетика", "вимірювання", "квантор", "біт", "ймовірність", "предикат", "компресування", "модальність", вжитих у [5], нема в [4]. Цей набір ключових слів вказує, що робота [5] присвячена розробці теоретичних методів аналізу мовної інформації в лінгвістиці. Проаналізований частотний склад цих текстів показує, що науковці в Україні в своїх роботах працюють над різними проблемами лінгвістики і в цілому охоплюють весь спектр завдань прикладної лінгвістики.

**Висновки**

Розроблено комп'ютерну програму на мові С++ для проведення лексикографічної обробки* текстових даних з метою встановлення спрямування текстів прикладної лінгвістики на матеріалах текстів українською та англійською мовами. Програма дає змогу статистично аналізувати тексти і визначати частоту вживання слів у тексті. На основі частотного аналізу текстів із прикладної лінгвістики українською та англійською мовами виділено ключові слова текстів, а також спільні найчастіше вживані слова. Виділення переліку спільних слів "Language" – "Мова", "English" – "Українська" дозволяє* стверджувати, що це тексти мовного спрямування і стосуються вивчення відповідної мови. Інші незбіжні ключові слова відображають індивідуальність досліджуваних текстів.

Незважаючи на однакове фахове спрямування досліджуваних матеріалів "The Journal of Applied Linguistics" та "Нариси з комп'ютерної лінгвістики", ключові слова текстів виявились різними. Отже, робоча гіпотеза про спільність ключової лексики у текстах з однаковим фаховим спрямуванням не підтвердилась. Наукові статті "The Journal of Applied Linguistics" присвячені вивченню та розвитку лінгвістичних методів для полегшення вивчення іноземних мов. "Нариси з комп'ютерної лінгвістики" віддають перевагу комп'ютерним методам вивчення мови та укладанню словників. Ця розбіжність не вказує на різні підходи до розуміння терміна "прикладна лінгвістика", а швидше відображає багатоманітність завдань прикладної лінгвістики.

Розширення кола досліджуваних текстів дозволить створити в майбутньому частотний словник – мінімум з прикладної лінгвістики. Однак, виконуючи це завдання, треба чітко розуміти, що наповненість словника залежатиме від того, для якого з напрямків прикладної лінгвістики укладають цей словник.

Авторка вдячна зав. кафедри прикладної лінгвістики національного університету "Львівська політехніка" Н.І.Андрейчук за запропоновану тему досліджень та допомогу в написанні статті.

*1. Гринбаум О.Н. Компютерные аспекты стилеметрии / Прикладное языкознание. Отв. Редактор Герд А.С. Санкт-Петербург. – Из-во С.-Петербургского ун-та, 1998.– с.451-465. 2. Перебийніс В.І. Статистичні методи для лінгвістів. – Вінниця: Нова Книга, 2001.– с.168. 3. Journal of Applied Linguistics" June, 1997. 4. Пещак М.М. Нариси з комп'ютерної лінгвістики. – Ужгород: Вид-во "Закарпаття", 1999.– с.200. 5. .Партико З.В. Образна концепція теорії інформації. – Львів: Видавничий центр ЛНУ ім. І.Франка, 2001.– С. 133.*

* (для) лексикографічного опрацювання – *ред.*