УДК 004.4:81'324

Жуковська Вікторія Вікторівна
кандидат філологічних наук, доцент,
завідувач кафедри міжкультурної комунікації та прикладної лінгвістики
Житомирський державний університет імені Івана Франка, м. Житомир, Україна
ORCID ID 0000-0002-4622-4435
victoriazhukovska@gmail.com

Мосіюк Олександр Олександрович
кандидат педагогічних наук, старший викладач кафедри прикладної математики та інформатики
Житомирський державний університет імені Івана Франка, м. Житомир, Україна
ORCID ID 0000-0003-3530-1359
mosxandrwork@gmail.com

Комаренко Вероніка Василівна
викладач кафедри міжкультурної комунікації та прикладної лінгвістики
Житомирський державний університет імені Івана Франка, м. Житомир, Україна
ORCID ID 0000-0002-6107-3057
vika1922@ukr.net

ЗАСТОСУВАННЯ ПРОГРАМНОГО ПАКЕТУ R У НАУКОВИХ ДОСЛІДЖЕННЯХ МАЙБУТНІХ ФІЛОЛОГІВ

Анотація. Одним із новітніх напрямів прикладного мовознавства є корпусна лінгвістика, яка займається побудовою, обробленням та експлуатацією текстових корпусів. На сьогодні якісний аналіз величезних масивів емпіричного мовного матеріалу, що надає в розпорядження лінгвіста корпус, неможливо здійснити без залучення комп'ютерних технологій і відповідних статистичних методів. Відтак навчання майбутніх філологів ефективно застосовувати прикладні статистичні програми є важливим етапом наукової підготовки спеціалістів цього напряму. Запропонована стаття розкриває можливості використання однієї з найпоширеніших у західній лінгвістиці, але маловідомої в Україні, статистичної системи аналізу даних – програмного комплексу R – у дослідженнях майбутніх філологів. У роботі розкриваються переваги й недоліки цього продукту порівняно з іншими подібними програмними пакетами (SPSS і Statistica), а також надаються посилання на матеріали в мережі Internet для самостійного опанування зазначеним програмним засобом. Гнучкість й ефективність застосування програмного комплексу R для розв'язання мовознавчих завдань продемонстровано на прикладі статистичного аналізу вживання маркерів зменшення категоричності у корпусі американського академічного мовлення. Для правильного розуміння філологами-початківцями особливостей проведення лінгвостатистичного експерименту в R наведено детальний опис кожного етапу здійсненого дослідження. Статистична верифікація вживання маркерів зменшення категоричності висловлення в мовленні студентів і викладачів була здійснена з використанням таких статистичних методів як λ-критерій Колмогорова-Смірнова та U-критерій Манна-Уітні. У статті наводяться розроблені алгоритми для проведення розрахунків за допомогою зазначених критеріїв із використанням вбудованих команд і різних спеціалізованих бібліотечних функцій R, створених співтовариством користувачів для розширення функціональності зазначеного програмного комплексу. Кожен скрипт, написаний на R для проведення статистичних розрахунків, супроводжується детальним описом та характеристикою отриманих результатів обчислень. Серед перспектив подальших досліджень з окресленої проблематики необхідно звернути увагу на реалізацію низки заходів, спрямованих на підвищення обізнаності майбутніх спеціалістів із статистичною системою аналізу даних і навчання їх роботи з R, що є важливим для фахового зростання майбутнього науковця-філолога.

Ключові слова: статистична система аналізу даних R; корпус академічного мовлення; маркери зменшення категоричності; λ-критерій Колмогорова-Смірнова; U-критерій Манна-Уітні.

1. ВСТУП

Постановка проблеми. У сучасному інформатизованому суспільстві відбувається інтенсивне проникнення комп'ютерних технологій та методів математичної статистики в гуманітарні науки. Соціологія, педагогіка та психологія все частіше залучають спеціалізовані програмні засоби та статистичні методи для перевірки та підтвердження результатів наукового аналізу.

Комп'ютерні технології також суттєво змінили природу й розширили інформаційне поле лінгвістичних досліджень, запропонувавши нові технічні можливості для відбору, опрацювання й збереження мовних даних. Взагалі сьогоднішня наука про мову перебуває під впливом так званого «квантитативного повороту» [1], с. 2], парадигмального зсуву у бік емпіричного підходу до аналізу мови. Як наслідок велика кількість лінгвістичних досліджень передбачає оброблення значних за обсягом масивів мовних даних (лінгвістичних корпусів) і застосування різноманітних статистичних методів. Використання методів статистики в лінгвістичному аналізі дає змогу суттєво модифікувати уявлення про систему мови і можливості її функціонування, а також підвищує дескриптивну і пояснювальну надійність й обґрунтованість результатів лінгвістичних розвідок. Необхідність аналізу великих масивів лінгвальних даних спонукає філологів використовувати відповідні комп'ютерні програми статистичного оброблення. Наприклад, однією з найпоширеніших у західній лінгвістиці є статистична система аналізу даних R, яка є потужним вільнопоширюваним програмним забезпеченням, що надає в розпорядження лінгвіста увесь набір методів, необхідних для якісного лінгвостатистичного аналізу, допомагаючи якісно й швидко здійснити необхідні обчислення й розрахунки та візуалізувати отримані результати.

Зважаючи на беззаперечні переваги, які надає застосування R у наукових дослідженнях із лінгвістики, філологи-початківці практично не знайомі з можливостями цієї програми і не мають досвіду її застосування. Отож, в умовах інтенсивного розвитку комп'ютерних технологій і постійного зростання інформаційних потоків проблема підготовки сучасного філолога до здійснення наукових досліджень із застосуванням сучасних комп'ютерних програм статистичного оброблення даних є надзвичайно актуальною. Адже саме статистичні методики з комп'ютерною підтримкою відкривають нові шляхи дослідження мови й мають величезний потенціал для вирішення багатьох теоретичних і практичних аспектів обробки текстових даних [2], с. 4]. У практичній площині це також сприятиме вдосконаленню алгоритмів штучного інтелекту для машинного перекладу.

Аналіз останніх досліджень та публікацій. Застосування статистичних методів у лінгвістиці має доволі довгу історію, зокрема проблеми використання статистичних методів у лінгвістичних дослідженнях вивчали К. Б. Буктаєв, Б. М. Головін, В. В. Левицький, І. А. Носенко, В. І. Перебийніс, Р. Г. Піотровський та інші. Попри це, у сучасному вітчизняному лінгвістичному доробку наявна значна кількість лінгвостатистичних і стилеметричних досліджень, виконаних під керівництвом професора В. В. Левицького і професора В. І. Перебийніс та їхніх учнів.

Водночас аналіз лінгвістичних праць засвідчив, що в українській лінгвістиці відсутні публікації, у яких здійснено лінгвостатистичне дослідження з використанням статистичної системи аналізу даних R. Натомість використання R у дослідженнях із лінгвістики, зокрема корпусної, набуває все більшої популярності у західній лінгвістиці, про що свідчать останні публікації G. Desagulier, M. Dickinson, S. Th. Gries, G. B. Jenset, N. Levshina та інших.

Варто зазначити, що загалом особливості обробки статистичних даних за допомогою мови програмування R описували Є. М. Балдін, П. А. Волкова, І. С. Зарядов, А. І. Коробейніков, Р. Є. Майборода , С. Е. Мастицький, С. А. Назарова, С. В. Петров, О. В. Сугакова, В. Г. Суфіянов, В. К. Шітіков, А. Б. Шипунов.

З-поміж закордонних науковців статистичний аналіз даних за допомогою R здійснено у працях B. S. Everitt, T. Hothorn, J. M. Quik, Y. Cohen.

Отже, **метою** запропонованої статті є продемонструвати можливості використання статистичної системи аналізу даних R у наукових дослідженнях майбутніх філологів.

Ключовими **завданнями статті** є подані нижче.

1. З'ясувати переваги й недоліки використання програмного пакету R у лінгвістичних дослідженнях.

2. Ознайомити майбутніх філологів з можливостями використання R для проведення лінгвостатистичного аналізу, навівши як приклад дослідження функціонування маркерів зменшення категоричності висловлення в американському академічному мовленні.

2. МЕТОДИ ДОСЛІДЖЕННЯ

Для розкриття поставленої мети були використані загальнонаукові і спеціальні теоретичні методи: аналізу і синтезу даних із наукових джерел і ітератури з проблематики використання методів математичної статистики й засобів комп'ютерного оброблення даних у мовознавстві; порівняння стану дослідження проблеми в наукових публікаціях українських і закордонних авторів, їх узагальнення.

Методи математичної статистики (непараметричний критерій Манна-Уітні та модифікація критерію Колмогорова-Смірнова, яка була виконана Х. Лілліфорсом) дали змогу продемонструвати ефективність розв'язання прикладних задач корпусної лінгвістики за допомогою R.

3. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

3.1. Характеристика мови статистичного програмування R і середовище розроблення R Studio

Аналіз великих обсягів текстів уже неможливо якісно виконувати без спеціалізованого програмного забезпечення. Застосування статистичних критеріїв, факторного і дисперсійного аналізу сприяє отриманню нових результатів, які за допомогою звичних підходів довести було б вкрай важко.

До найбільш популярних програм для оброблення статистичної інформації відносять пакети SAS та SPSS, MiniTAB, Statistica, STATGRAPHICS; окремі статистичні розрахунки можна виконувати на базовому рівні в системах комп'ютерної математики Maple, MatLAB, Matematica; обмеженим потенціалом, порівняно із зазначеними програмами, володіють пакети Excel від компанії Microsoft та Calc, що входить до вільнопоширюваного офісного пакету LibreOffice.

Окремо необхідно виділити статистичну систему аналізу даних R. Це вільнопоширюваний програмний продукт, який динамічно розвивається і використовується провідними науковими установами в різних наукових галузях, зокрема і в корпусній лінгвістиці.

R – статистична система аналізу даних, яка створена Росом Іхакою та Робертом Гентлеманом. Програма становить і мову програмування, і спеціалізоване програмне забезпечення одночасно [3]. R розглядають як діалект статистичної мови S, яка була

створена фахівцями AT1T Bell Laboratory. На цей час підтримання й розроблення нових версій забезпечує об'єднана команда розробників з усього світу.

R поширюється за ліцензією GNU GPL, а отже, будь-хто може цю програму використовувати у своїх дослідженнях. Єдине обмеження накладається на модифікацію платформи. Будь-які зміни в коді продукту або ж створення нових модулів на його основі також повинні поширюватися під такою ж ліцензією, як сама програма.

До найважливіших переваг цього програмного пакету R необхідно віднести:

- ефективне опрацювання даних і прості засоби для збереження результатів;
- набір операторів для оброблення масивів, матриць, векторів;
- велику інтегровану колекцію інструментальних засобів для проведення статистичного аналізу;
- багатофункціональні бібліотеки для графічного оформлення даних досліджень;
- ефективну інтерпретовану мову програмування, яка уможливлює самостійно розширити функціонал програмного забезпечення і пристосувати його до специфіки задачі;
- відкритість і доступність самого програмного забезпечення і його модулів [3].

З-поміж недоліків, які пов'язують з аналізованою програмою, варто виділити такі:

- нестабільність роботи нових версій R;
- необхідність виконувати всі запити в командному рядку;
- R – достатньо складний програмний пакет для вивчення початківцям.

Проте не варто вважати зазначені недоліки критичними. Перший недолік пов'язаний із тим, що R розробляється співтовариством програмістів і науковців переважно у вільний від роботи час. Тому інколи не завжди вдається вчасно, до виходу останнього релізу, якісно протестувати всі зміни. Щоправда, зазвичай, дуже швидко знаходять проблеми в розрахунках – і програмісти оперативно вносять корективи, виправляючи в такий спосіб програму.

Останні два недоліки пов'язані з відсутністю розвинутого графічного інтерфейсу, але це не є значним недоліком. Попри це, роботу з командним рядком можна вважати більше перевагою ніж недоліком, оскільки вона дає змогу гнучко пристосовувати систему до реалій поставленої прикладної задачі.

Складність в опануванні R більшою мірою пов'язана з майже повною відсутністю навчальних матеріалів українською мовою. Водночас англомовних статей, посібників і підручників більш ніж достатньо для засвоєння програми. Чимало їх можна віднайти у вільному доступі в мережі Internet, наприклад, таким ресурсом є сайт [4].

Для спрощення роботи з R досить часто використовують додаткове спеціалізоване програмне забезпечення, зокрема R Studio. R Studio є вільним інтегрованим середовищем розроблення, створене Джозефом Аллейре [5], що розповсюджується у двох версіях R Studio Desktop (для персональних комп'ютерів та ноутбуків) і R Studio Server (серверний варіант IDE). Програма дає змогу полегшити роботу з написання коду на мові R, швидко встановлювати і підключати додаткові модулі, оформляти і зберігати графіки статистичних розподілів тощо. Поєднання можливостей R та IDE R Studio допомагає спростити роботу над аналізом статистичних даних.

3.2. Використання R для статистичного аналізу вживання маркерів зменшення категоричності висловлення

Розглянемо можливості застосування програмного пакету R у філологічних дослідженнях. Для прикладу візьмемо дослідження В. В. Комаренко, аспірантки ННІ іноземної філології Житомирського державного університету імені Івана Франка. Одним із завдань дослідження є здійснити статистичну верифікацію соціолінгвістичної

ідентифікації мовця в американському усному академічному дискурсі в аспекті встановлення статистично значущих розбіжностей у вживанні маркерів зменшення категоричності висловлення в мовленні студентів та викладачів.

Матеріалом дослідження є Мічиганський корпус усного академічного мовлення (MICASE – Michigan Corpus of Academic Spoken English) [6]. Корпус містить більше 1,8 млн. слововживань затранскрибованого академічного мовлення (лекції, обговорення в аудиторії, лабораторні заняття, семінари, консультації), зібраного в Мічиганському університеті (США).

Характерною рисою сучасної англо-американської наукової і науково-академічної комунікації є використання різноманітних засобів хеджингу. Термін «хеджинг», або ухилення від прямої відповіді, увів Дж. Лакофф [7] для позначення властивості певних слів надавати висловленням більшої або меншої невизначеності. Як комунікативна стратегія хеджинг на дискурсивному рівні розуміється як фактична невпевненість та припускає модифікацію комунікантом пропозиційного змісту висловлювання з метою забезпечити точність, повноту, автентичність, достовірність і коректність у передаванні фактів, і, відповідно, відображає ступінь поінформованості автора в галузі [8], с. 83]. Хеджинг вербалізується в мові великим спектром засобів, у нашому розумінні – маркерами зменшення категоричності висловлення.

У дослідженні виокремлюється три групи маркерів зменшення категоричності висловлення:

I. Маркери зорієнтовані на слухача *(hearer oriented),* що вживаються для привертання уваги слухачів, залучаючи їх до участі в академічній комунікації (наприклад, *you see/ think/ could/ might/ know/ (all) assume/ may* та ін.).

II. Маркери зорієнтовані на мовця *(speaker oriented)*, що вживаються для вираження особистої думки і знань мовця, орієнтуючи його на передавання інформації аудиторії й одночасно на дотримання мовного етикету (наприклад, *they say…, according to…, it is assumed…* та ін.).

III. Маркери зорієнтовані на організацію дискурсу (discourse organizing), які вживаються задля привернення уваги до певних частин висловлення, уникнення категоричних тверджень та категоричності, прийняття альтернативних варіантів (наприклад, *at least, to put it mildly, at some point, at some level, up to a certain point, so to speak* та ін.).

Кожна із зазначених груп своєю чергою поділяється на декілька підгруп (детальніше використана в роботі класифікація маркерів зменшення категоричності висловлення описана в [9]).

Зважаючи на значну кількість наукових доробків, присвячених лінгвістичному аналізу маркерів зменшення категоричності висловлення, вітчизняні корпусні дослідження зазначених одиниць із соціолінгвістичних позицій практично відсутні. У запропонованій статті зупинимося на аналізі вживання маркерів зменшення категоричності висловлення як соціолінгвістичних змінних, що є факторами ідентифікації учасників академічної комунікації в американському вищому навчальному закладі. Зокрема, проаналізуємо особливості вживання цих маркерів з огляду на ідентифікаційний фактор «академічна роль комунікантів» із використанням статистичного аналізу, здійсненого за допомогою програмного забезпечення R.

Важливо зауважити, що корпусні розвідки проводяться із залученням великих масивів автентичних лінгвістичних даних, що приводить до необхідності враховувати велику кількість факторів під час їх аналізу. Тому подальше дослідження відбувалося в декілька етапів.

Спочатку з аналізованого корпусу методом суцільної вибірки було відібрано 2109 маркерів зменшення категоричності висловлення в мовленні студентів та 4259 – у

мовленні викладачів. Порівнявши отримані вибірки, було виявлено 227 спільних для обох вибірок маркерів. Отримані маркери, які є матеріалом подальшого аналізу, було покласифіковано на 15 підгруп.

Відзначимо, що ці вибірки є незалежними, містять непараметричні дані та не є співрозмірними за кількістю аналізованих одиниць. Це накладає додаткові обмеження на використання статистичних критеріїв і ускладнює аналіз статистичної інформації. Щоб уникнути помилок, під час перевіряння статистичних гіпотез було перевірено, чи відповідають отримані емпіричні дані нормальному розподілу. Для цього використано λ-критерій Колмогорова-Смірнова, що є другим кроком статистичного дослідження. Цей критерій дає змогу порівнювати емпіричний розподіл з теоретичним чи два емпіричних розподіли [10], с. 142-151], [11], с. 72-77]. Сам метод перевіряння статистичних гіпотез є досить відомим і використовується в різних науках, зокрема в математичній лінгвістиці [12], с. 121-124], [13], с. 340-343] та психолого-педагогічних дослідженнях [10], с. 142-151], [11], с. 72-77].

У нашому випадку перевірилися такі статистичні гіпотези:

Нульова гіпотеза H0. Емпіричний розподіл спостережуваних маркерів у студентських текстах не відрізняється від нормального розподілу, а виявлені відмінності є випадковими.

Альтернативна гіпотеза H1. Емпіричний розподіл спостережуваних маркерів у мовленні студентів відрізняється від нормального розподілу, а виявлені відмінності є статистично значущими.

Аналогічні статистичні гіпотези сформульовані й для емпіричного розподілу виявлених маркерів у текстах викладачів.

Для проведення статистичних розрахунків використовувалася команда lillie.test (), яка включена до бібліотеки nortest. Алгоритм, який реалізується представленою командою, базується на модифікації критерію Колмогорова-Смірнова, що була проведена Х. Лілліфорсом [14], [15].

У системі R програма для розрахунків матиме вигляд:

```
library("nortest", lib.loc="~/R/win-library/3.3")
setwd("D:/R_Statistics")

VarTable <- read.table(file = file.choose(), header = TRUE)

V1 <- c(VarTable$Stud) # Емпіричний розподіл спостережуваних словосполучень у
                       # студентських текстах

V2 <- c(VarTable$Teach) # Емпіричний розподіл спостережуваних словосполучень у
                        # текстах викладачів

# Розрахунок значення для студентів
lillie.test(V1) # Перевірка першого емпіричного розподілу на нормальність

# Розрахунок значення для викладачів
lillie.test(V2) # Перевірка другого емпіричного розподілу на нормальність
```

Цей алгоритм імпортує текстовий файл із даними, сформованими в таблицю за правилами: *1. Кожен стовпчик має власну назву, за допомогою якої програма звертається до його даних. 2. Рядок таблиці відповідає окремому рядку текстового файлу. 3. У рядку дані з різних стовпчиків відокремлюються однією табуляцією.*

Результати виконання цього алгоритму будуть такими:

```
# Розрахунок значення для студентів
              Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data:  V1
D = 0.21991, p-value < 2.2e-16

# Розрахунок значення для викладачів

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  V2
D = 0.19778, p-value < 2.2e-16
```

Як бачимо, в обох вибірках значення $D_{крит}$ є великим, а $p < 2{,}2 \cdot 10^{-16}$ – дуже малим. Це означає, що нульова гіпотеза відкидається і приймається альтернативна: емпіричний розподіл спостережуваних маркерів зменшення категоричності висловлення в мовленні студентів відрізняється від нормального розподілу, а виявлені відмінності є статистично значущими. Така ж гіпотеза буде прийнята і для емпіричного розподілу аналізованих маркерів зменшення категоричності висловлення в мовленні викладачів. Отже, до порівняння двох зазначених вибірок неможливо застосовувати критерії, основою яких є нормальний розподіл.

Наступним етапом аналізу даних було статистичне дослідження, яке допомогло відповісти на таке питання: «Якій із вибірок (мовленню студентів чи викладачів) є притаманним уживання маркерів зменшення категоричності висловлення?».

З огляду на результати попереднього етапу й особливості отриманих вибірок, найоптимальнішим кроком є використання непараметричних критеріїв для оцінки відмінностей, зокрема U - критерій Манна-Уітні. Він розрахований для оцінки відмінностей між вибірками за рівнем деякої ознаки, яка виміряна кількісно [10], с. 49-52]. Характерною особливістю методу є те, що він уможливлює виявити відмінності в кількості спостережуваних якостей на рівні, не меншому за 3 в кожній вибірці. Зазначений критерій, переважно в закордонній літературі, використовується під назвами як критерій Манна-Уітні-Уілкоксона або ж критерій суми рангів Уілкоксона [16], с. 182].

Цей критерій запропонував Френк Уілкоксон [17] у 1945 році. У 1947 р. Х. Б. Манн та Д. Р. Уітні удосконалили й розширили його можливості, відповідно метод було названо за прізвищами науковців [18].

У мові статистичного програмування R U-критерій Манна-Уітні реалізується за допомогою функції wilcox.test (X, Y), де X, Y – порівнювані вибірки [19], с. 8 - 11]. Наразі додаткових параметрів не потрібно додавати.

Під час використання зазначеного критерію дотримано такого припущення. За першу вибірку приймемо емпіричний розподіл аналізованих маркерів зменшення категоричності висловлення в мовленні викладачів, оскільки абсолютні частоти спостережуваного явища тут вищі, а отже, можна припустити, що для викладачів є більш характерним використання зазначених маркерів. Відповідно друга вибірка – емпіричний розподіл спостережуваних маркерів зменшення категоричності висловлення в мовленні студентів.

Було сформульовано статистичні гіпотези, які перевірялися за допомогою вибраного критерію.

Нульова гіпотеза H0. Рівень вживання маркерів зменшення категоричності висловлення студентами не нижчий рівня використання цих маркерів викладачами.

Альтернативна гіпотеза H1. Рівень вживання маркерів зменшення категоричності висловлення студентами нижчий використання цих маркерів викладачами.

Скрипт, який дає змогу перевірити ці гіпотези, має вигляд:

```
setwd("D:/R_Statistics")
VarTable <- read.table(file = file.choose(), header = TRUE)
```

```
V1 <- c(VarTable$Teach) # Емпіричний розподіл спостережуваних словосполучень у
                        # текстах викладачів

V2 <- c(VarTable$Stud)  # Емпіричний розподіл спостережуваних словосполучень у
                        # студентських текстах
wilcox.test(V1, V2)
```

Розрахунки, виконані за допомогою цього алгоритму, матимуть такий вигляд:

```
        Wilcoxon rank sum test with continuity correction

data:  V1 and V2
W = 37682, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Отже, нульова гіпотеза була відкинута, а прийнята альтернативна гіпотеза, що уможливила з високою ймовірністю стверджувати, що рівень вживання маркерів зменшення категоричності висловлення в мовленні студентів нижчий за рівень використання цих маркерів викладачами. Можна зробити висновок, що для викладачів використання зазначених одиниць є більш «природнім», ніж для студентів.

На четвертому етапі статистичного дослідження було аналогічно перевірено кожну із 15 виділених підгруп маркерів зменшення категоричності висловлення. Це дало змогу визначити ті з них, у яких відмінності вживання зазначених маркерів студентами та викладачами є статистично значущими. Сформульовані нульова та альтернативна гіпотези є аналогічними до загального випадку, але з урахуванням назви групи маркерів, яка перевірялася.

У результаті зазначеного перевіряння виявилося, що у 8 виділених підгрупах маркерів відмінності вживання підтвердилися з різним ступенем точності ($p < 0,05$ або $p < 0,01$). Зокрема, вживання маркерів особистого вираження ймовірності, маркерів вираження відносного значення (прислівники, дієслова та прикметники), маркерів зменшення є характерним для викладачів, а ймовірність неправильного відхилення нульової гіпотези становить $p < 0,01$. Приклад такої підгрупи надано в табл. 1:

Таблиця 1

Статистичні показники для підгрупи маркерів особистого вираження ймовірності, зорієнтованих на мовця

Phrase	**Students**	**Teachers**
I guess	5	14
I'm not sure	3	4
I hope	7	33
Let's say	6	13
Say	7	15
I would say	4	23
I'm not sure that	20	7
I mean	20	30
I think	5	27
I wonder	3	10
My / our belief	3	7
We find	3	2
I recall	10	8
My own view	4	6
I suggest	10	20
I consider	2	7
I presume	3	21
I consider	15	17
W = 249.5, **p** < 0.0057		

Для підгруп, які представляють атрибутивні маркери вираження особистої оцінки, зорієнтовані на мовця, маркери вираження адаптації, зорієнтованого на мовця, та маркери, зорієнтовані на організацію дискурсу (зв'язувальні ад'юнкти) також підтвердилася притаманність мовленню викладачів. Єдине, що відрізняє ці підгрупи від попередніх, це те, що ймовірність неправильного відхилення нульової гіпотези становить $p < 0,05$. Приклад такої підгрупи наведено в табл. 2:

Таблиця 2

Статистичні показники для підгрупи атрибутивних маркерів вираження особистої оцінки, зорієнтованих на мовця

Phrase	**Students**	**Teachers**
They say	7	35
According to…	3	1
In / on smb's view / point of view	14	20
In / on smb's opinion	2	17
As smb. Observes	2	10
One might consider	16	26
It is often believed	2	1
It is often heard	5	26
It is recognized	4	11
It is widely / commonly assumed	3	22
It is established that…	5	6
It has been suggested	8	14
I t seems	10	21
Some feel that	0	2
It might be considered	2	8
It is debatable	10	17
It must be	21	4
It is probable	10	4
It is thought to	3	24
W = 260, **p** < 0.021		

У решті семи підгрупах (маркери, зорієнтовані на слухача, маркери вираження відносного значення (модальні дієслова), маркери вираження відносного значення (іменники), маркери вираження відносного значення (займенники), маркери вираження частковості, маркери, зорієнтовані на організацію дискурсу (відносні займенники) та маркери зорієнтовані на організацію дискурсу (синтаксичні маркери й деперсоналізовані звороти)) статистично значущої різниці між використанням відповідних словосполучень студентами і викладачами не виявлено. Приклад підгрупи представлено в табл. 3:

Таблиця 3

Статистичні показники для групи маркерів, зорієнтованих на слухача

Phrase	**Students**	**Teachers**
You see	1	14
You know	42	76
Imagine	1	2
If you catch	6	4
As you suggest	4	8
You consider	4	5
You presume that	1	16
W = 37, **p** < 0.1219		

Якщо в організації статистичного оброблення даних спиратися на думку В. В. Левицького [12], с. 65], згідно з якою в лінгвістичних дослідженнях отримані

емпіричні розподіли є близькими до нормального, то у такому разі, доцільним є застосування критерію χ^2–Пірсона. Для цього представлений код необхідно модифікувати так:

```
setwd("D:/R_Statistics")
VarTable <- read.table(file = file.choose(), header = TRUE)

V1 <- c(VarTable$Stud)
V2 <- c(VarTable$Teach)

DMatrix <- matrix(c(V1,V2), ncol = 2)
chisq.test(DMatrix)
```

Під час застосування цього критерію відмінності у використанні маркерів зменшення категоричності висловлення в американському усному академічному дискурсі студентами і викладачами на загальному рівні також були підтверджені. Наразі було підтверджено відмінності в тих же восьми підгрупах маркерів, а також виявлено відмінності (на різному рівні значимості $p < 0,05$ або $p < 0,01$) ще в шести підгрупах (маркери, зорієнтовані на слухача, вираження відносного значення, зорієнтованого на мовця (модальні дієслова, іменники та прикметники), зорієнтовані на організацію дискурсу (синтаксичні маркери та деперсоналізовані звороти) та вираження частковості, зорієнтовані на мовця). Такий результат не є дивним, оскільки вказаний критерій є більш точним, ніж непараметричні методи.

Щоб визначити, хто (студенти чи викладачі) віддає перевагу вживанню конкретного маркера, на думку В. В. Левицького, необхідно розрахувати коефіцієнт спряженості Ф [12], с. 89, 115].

Нижче наведемо алгоритм для обчислення цього коефіцієнта, який написаний мовою R.

```
setwd("D:/R_Statistics ")
library("gmp", lib.loc="~/R/win-library/3.3")
library("Rmpfr", lib.loc="~/R/win-library/3.3")

VarTable <- read.table(file = file.choose(), header = TRUE)
Phrase <-c(as.vector(VarTable$Phrase))

V1 <-c(VarTable$Stud)
V2 <-c(VarTable$Teach)

Sum1 <-sum(V1)
Sum2 <-sum(V2)

FbVec <-vector(mode = "numeric", length = length(V1))
for (i in 1:length(V1))

{
  a <-as.bigz.bigq(V1[i])
  b <-as.bigz.bigq(V2[i])
  c <-as.bigz.bigq(Sum1 - a)
  d <-as.bigz.bigq(Sum2 - b)
  k <-mul.bigz((a+b),(c+d))
  l <-mul.bigz((a+c),(b+d))
  Fb                <-                as.double(mul.bigz(a,d)                -
mul.bigz(b,c))/sqrt(as.double(mul.bigz(k,l)))
  FbVec[i] = round(Fb,6)
}

ResultTable <- data.frame(Phrase = Phrase, V1 = V1, V2 = V2, FB = FbVec)
write.table(RezaltTable,file = "Rezalt_tab.txt", append = TRUE, sep = "\t",
eol = "\n", dec = ".", col.names = TRUE, row.names = FALSE, quote = FALSE)
```

Залежно від того, який знак коефіцієнта Ф, можна зробити висновок про те, хто віддає перевагу вживанню певного маркера. Проведені розрахунки показують, що навіть у підгрупах, у яких було відзначено перевагу використання викладачами маркерів зменшення категоричності висловлення, були такі значення Ф, які вказували на надання переваги окремим маркерам у мовленні студентів.

4. ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

Отже, здійснений огляд можливостей використання R у лінгвістичних дослідженнях свідчить, що ця статистична система аналізу даних є потужним відкритим програмним комплексом для оброблення й аналізу статистичних даних різного обсягу і складності. Його реалізація у вигляді інтерпретованої мови програмування дає змогу швидко створити код для розв'язання різноманітних завдань, що постають перед лінгвістом. Універсальність програмного забезпечення і його безкоштовність є безперечними перевагами R над подібними статистичними програмами, зокрема SPSS та Statistica.

На думку авторів, використання статистичної системи аналізу даних R у практиці мовознавчих досліджень сприятиме формуванню в майбутніх філологів більш цілісного і поглибленого усвідомлення професійної діяльності й наблизить їхні дослідження до сучасного рівня наукових лінгвістичних знань.

До перспектив подальших пошуків з окресленої проблематики необхідно віднести 1) розроблення методики використання програмного пакету R у різноаспектних лінгвістичних розвідках; 2) створення спеціального навчального курсу для майбутніх філологів з використання R для лінгвостатистичного аналізу; 3) розроблення навчальних матеріалів для забезпечення вказаного курсу якісною літературою.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ ТА ЛІТЕРАТУРИ

[1] L. A. Janda, *Quantitative Methods in Cognitive Linguistics. An Introduction*, *Cognitive linguistics. The quantitative turn. The essential reader,* Berlin : De Gruyter Mouton, 2013, 321 p.

[2] С. Н. Бук, *Основи статистичної лінгвістики*, Львів: Видавничий центр ЛНУ імені Івана Франка, 2008.

[3] "What is R? " [Електронний ресурс]. Доступно: https://www.r-project.org/ about.html.

[4] "R resources (free courses, books, tutorials, & cheat sheets) ". [Електронний ресурс]. Доступно: https://paulvanderlaken. com/2017/08/10/r-resources-cheatsheets-tutorials-books/.

[5] "Why RStudio? " [Електронний ресурс]. Доступно: https://www.rstudio.com/ about/.

[6] "Michigan corpus of academic spoken English". [Електронний ресурс]. Доступно: https://quod.lib.umich.edu/m/micase/.

[7] D. Lakoff, Hedges: "A study in meaning criteria and the logic of fuzzy concepts", *Journal of philosophical logic*, №. 2 (4), 1972, p. 458-508.

[8] А. В. Ярхо, "Референційний хеджинг як стратегія етикетизації у дискурсі англомовної науково-дослідницької статті: контрастивний аналіз", *Вісник Харківського національного університету імені В. Н. Каразіна. №930 Серія «Романо-германська філологія. Методика викладанні іноземних мов»*, 2010, Випуск 64, С. 82-90.

[9] В. В. Шилюк, "Класифікація засобів вираження позиції мовця в усній комунікації: порівняльний аналіз", *Вісник Житомирського державного університету,* Вип. 2 (80), 2015, С. 302-308.

[10] Е. В. Сидоренко, *Методы математической обработки в психологии*, СПб., ООО «Речь», 2000.

[11] Л. В. Шелехова, *Математические методы в педагогике и психологи: в схемах и таблицах:* учебное пособие, Майкоп, изд-во АГУ, 2010.

[12] В. В. Левицкий, *Квантитативные методы в лингвистике*, Черновцы, Рута, 2004.

[13] Р. Г. Пиотровский, К. Б. Бектаев, А. А. Пиотровская, *Математическая лингвистика : учеб. пособие для пед. институтов*, М., Высшая школа, 1977.

[14] H. W. Lilliefors, "On the Kolmogorov-Smirnov test for normality with mean and variance unknown", *Journal of the American Statistical Association*, Vol. 62, 1967, p. 399-402.
[15] "Package 'nortest'". [Електронний ресурс]. Доступно: https://cran.r-project.org/web/packages/nortest/nortest.pdf.
[16] R. M. Conroy, "What hypotheses do "nonparametric" two-group tests actually test?", *The Stata Journal*, № 2, 2012, p. 182-190.
[17] F. Wilcoxon, "Individual comparisons by ranling methods", *Biometrics Bull*, vol. 1, 1945, p. 80-83.
[18] H. B. Mann, D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other", *Annals of Mathematical Statistics*, vol. 18, № 1, 1947, p. 50-60.
[19] А. Б. Шипунов, А. И. Коробейников, Е. М. Балдин, "Анализ данных с R (II"). [Електронний ресурс]. Доступно: http://www.inp.nsk.su/~baldin/DataAnalysis/ R/R-05-2var.pdf.

Матеріал надійшов до редакції 21.03.2018 р.

ПРИМЕНЕНИЕ ПРОГРАММНОГО ПАКЕТА R В НАУЧНЫХ ИССЛЕДОВАНИЯХ БУДУЩИХ ФИЛОЛОГОВ

Жуковская Виктория Викторовна
кандидат филологических наук, доцент,
заведующая кафедрой межкультурной коммуникации и прикладной лингвистики
Житомирский государственный университет имени Ивана Франко., г. Житомир, Украина
ORCID ID 0000-0002-4622-4435
victoriazhukovska@gmail.com

Мосиюк Александр Александрович
кандидат педагогических наук,
старший преподаватель кафедры прикладной математики и информатики
Житомирский государственный университет имени Ивана Франко., г. Житомир, Украина
ORCID ID 0000-0003-3530-1359
mosxandrwork@gmail.com

Комаренко Вероника Васильевна
преподаватель кафедры межкультурной коммуникации и прикладной лингвистики
Житомирский государственный университет имени Ивана Франко., г. Житомир, Украина
ORCID ID 0000-0002-6107-3057
vika1922@ukr.net

Аннотация. Одним из новейших направлений прикладного языкознания является корпусная лингвистика, которая занимается построением, обработкой и эксплуатацией текстовых корпусов. Сегодня качественный анализ огромных массивов эмпирического языкового материала, который предоставляет в распоряжение лингвиста корпус, невозможно осуществить без привлечения компьютерных технологий и соответствующих статистических методов. Поэтому обучение будущих филологов эффективно использовать прикладные статистические программы является важным этапом научной подготовки специалистов этого направления. Предложенная статья раскрывает возможности использования одной из наиболее распространенных в западной лингвистике, но малоизвестной в Украине, статистической системы анализа данных – программного комплекса R – в исследованиях будущих филологов. В работе раскрываются преимущества и недостатки этого продукта по сравнению с другими подобными программными пакетами (SPSS и Statistica), а также предоставляются ссылки на материалы в сети Internet для самостоятельного освоения указанного программного средства. Гибкость и эффективность применения программного комплекса R для решения языковедческих задач продемонстрировано на примере статистического анализа употребления маркеров уменьшение категоричности в корпусе американской академической речи. Для правильного понимания начинающими филологами особенностей проведения лингвостатистического эксперимента в R предоставлено подробное описание каждого этапа проведенного исследования. Статистическая верификация употребления маркеров уменьшения категоричности высказывания в речи студентов и преподавателей была осуществлена с использованием таких статистических методов, как λ-критерий Колмогорова-Смирнова и

U-критерий Манна-Уитни. В статье представлены разработанные алгоритмы для проведения вычислений с помощью указанных критериев с использованием уже встроенных команд и различных специализированных библиотечных функций R, созданных сообществом пользователей для расширения функциональности указанного программного комплекса. Каждый скрипт, написанный на R для проведения статистических подсчетов, сопровождается подробным описанием и характеристикой полученных результатов вычислений. Среди перспектив дальнейших исследований по обозначенной проблематике необходимо обратить внимание на реализацию ряда мероприятий, направленных на повышение осведомленности будущих специалистов со статистической системой анализа данных и обучение их работы с R, что является важным для профессионального роста будущего ученого-филолога.

Ключевые слова: статистическая система анализа данных R; корпус академической речи; маркеры уменьшение категоричности; λ-критерий Колмогорова-Смирнова; U-критерий Манна-Уитни.

USING R IN THE RESEARCH BY FUTURE PHILOLOGISTS

Victoriia V. Zhukovska
PhD in Philology, Associate Professor,
Head of the Department of Cross-Cultural Communication and Applied Linguistics
Zhytomyr Ivan Franko State University, Zhytomyr, Ukraine
ORCID ID 0000-0002-4622-4435
victoriazhukovska@gmail.com

Oleksandr O. Mosiiuk
PhD in Pedagogics, Senior Lecturer,
Department of Applied Mathematics and Computer Science
Zhytomyr Ivan Franko State University, Zhytomyr, Ukraine
ORCID ID 0000-0003-3530-1359
mosxandrwork@gmail.com

Veronika V. Komarenko
Lecturer, Department of Cross-Cultural Communication and Applied Linguistics
Zhytomyr Ivan Franko State University, Zhytomyr, Ukraine
ORCID ID 0000-0002-6107-3057
vika1922@ukr.net

Abstract. Corpus linguistics is a newly emerging field of study in applied linguistics that deals with construction, processing, and exploitation of text corpora. To date, a high-quality analysis of vast amounts of empirical language data provided by computerized corpora is impossible without computer technologies and relevant statistical methods. Therefore, teaching future philologists to effectively apply statistical computer programs is an important stage in their research training. The article discusses the possibilities of using one of the leading in Western linguistics, but not well-known in Ukraine, software packages for statistical data analysis – R statistical software environment – in the research by future philologists. The paper reveals the advantages and disadvantages of this program in comparison with other similar software packages (SPSS and Statistica) and provides Internet links to R self-learn tutorials. The flexibility and efficacy of R for linguistic research are demonstrated on the example of a statistical analysis of the use of hedges in the corpus of academic speech. For novice philologists to properly understand the peculiarities of conducting a statistical linguistic experiment with R, a detailed description of each stage of the study is provided. The statistical verification of hedges in the speech of students and lecturers was carried out using such statistical methods as the Kolmogorov–Smirnov test and the Mann-Whitney U Test. The article presents the developed algorithms to calculate the specified tests applying the built-in commands and various specialized library functions, created by R user community to enhance the functionality of this statistical software. Each script for statistical calculations in R is accompanied by a detailed description and interpretation of the results obtained. Further study of the issue will involve a number of activities aimed at raising awareness and improving skills of future philologists in using R statistical software, which is important for their professional development as researchers.

Keywords: R statistical software environment; corpus of academic speech; hedges; the Kolmogorov-Smirnov test; the Mann-Whitney U Test.

REFERENCES (TRANSLATED AND TRANSLITERATED)

[1] L. A. Janda, Quantitative Methods in Cognitive Linguistics. An Introduction, *Cognitive linguistics. The quantitative turn. The essential reader*, Berlin : De Gruyter Mouton, 2013, (in English).
[2] S. N. Buk, The Basics of Statistical Linguistics: educational method. manual, Lviv: Publishing Center of Ivan Franko National University of LNU, 2008, (in Ukrainian).
[3] What is R? : [Online]. Available: https://www.r-project.org/about.html, (in English).
[4] R resources (free courses, books, tutorials, & cheat sheets). [Online]. Available: https://paulvanderlaken.com/2017/08/10/r-resources-cheatsheets-tutorials-books/, (in English).
[5] Why RStudio? [Online]. Available: https://www.rstudio.com/about/, (in English).
[6] Michigan corpus of academic spoken English. [Online]. Available: https://quod.lib.umich.edu/m/ micase/, (in English).
[7] D. Lakoff, Hedges: A study in meaning criteria and the logic of fuzzy concepts, *Journal of philosophical logic*, №. 2 (4), 1972, p. 458 - 508. (in English).
[8] A. V. Yarkho, Referential hedging as an etiquette strategy in the discourse of an anglo-american scientific research paper: a contrastive analysis, *Journal of Kharkiv National University named after V. N. Karazin, №930 Series «Romano-Germanic Philology. Methodology of Teaching Foreign Languages»*, 2010, issue 64, p. 82-90., (in Ukrainian).
[9] V. V. Shiluk, Classification of means of expressing the position of the speaker in spoken communication: comparative analysis, *Bulletin of Zhytomyr State University*, issue 2 (80), 2015, p. 302 - 308, (in Ukrainian).
[10] E. V. Sydorenko, Methods of mathematical processing in psychology, SPb: OOO "Rech", 2000, (in Russian).
[11] L. V. Shelekhova, Mathematical Methods in Pedagogy and Psychologists: in Schemes and Tables: Textbook, Maykop: ASU Publishing house, 2010, (in Russian).
[12] V. V. Levitsky, Quantitative methods in linguistics, Chernivtsi: Ruta, 2004, (in Russian).
[13] R. G. Piotrovsky, K. B. Bektaev, A. A. Piotrovskaya, Mathematical Linguistics: Textbook for pedagogical institutes, Moscow: «Higher School», 1977, (in Russian).
[14] H. W. Lilliefors, On the Kolmogorov-Smirnov test for normality with mean and variance unknown, *Journal of the American Statistical Association*, vol. 62,1967, p. 399 - 402, (in English).
[15] Package 'nortest'. [Online]. Available: https://cran.r-project.org/web/packages/nortest/nortest.pdf, (in English).
[16] R. M. Conroy, What hypotheses do "nonparametric" two-group tests actually test?, *The Stata Journal*, № 2, 2012, p. 182 - 190, (in English).
[17] F. Wilcoxon, Individual comparisons by ranling methods, *Biometrics Bull*, vol. 1, 1945, p. 80 - 83, (in English).
[18] H. B. Mann, D. R. Whitney, On a test of whether one of two random variables is stochastically larger than the other, *Annals of Mathematical Statistics*, vol. 18, № 1, 1947, p. 50 - 60, (in English).
[19] A. B. Shipunov, A. I. Korobeinikov, E. M. Baldin, Analysis of data with R (II). [Online]. Available: http://www.inp.nsk.su/~baldin/DataAnalysis/R/R-05-2var.pdf, (in Russian).