УДК 811.161.2'32

КВАНТИТАТИВНІ МЕТОДИ В ЛІНГВІСТИЦІ: НОВІТНІ ТЕНДЕНЦІЇ

Краснобаєва-Чорна Ж. В.
Донецький національний університет імені Василя Стуса

Визначено новітні тенденції застосування квантитативних методів у сучасній лінгвопарадигмі з проектуванням на освітній процес вищої школи. Схарактеризовано сучасні мовознавчі студіювання з актуалізацією квантитативних методів у рамах лінгвоперсонології, лексико-семантичної макротипології мов, принципу іконічності в мові, конструкційної граматики, стилеметрії з елементами психолінгвістики. Окреслено структурну специфіку курсу «Квантитативна лінгвістика» (галузь знань 03 «Гуманітарні науки», спеціальність 035 «Філологія»), що належить до циклу математичної та природничо-наукової підготовки навчального плану бакалаврів філології (прикладної лінгвістики) (освітня програма «Прикладна лінгвістика» / «Applied Linguistics») Донецького національного університету імені Василя Стуса.

Ключові слова: *квантитативна лінгвістика, лінгвоперсонологія, лексико-семантична макротипологія мов, принцип іконічності в мові, конструкційна граматика, стилеметрія.*

Krasnobaieva-Chorna Zh. Quantitative methods in linguistics: the latest trends. *Interest in the quantitative paradigm in linguistics has been going on for more than one century. In the early XX century attempts were made to statistically process the text aimed at establishing authorship or plagiarism (A. Markov, M. Morozov). Activation of domestic and foreign linguistic statistics takes place in the second half of the XX century with the appearance of a number of fundamental works that led to the beginning of new directions in linguistics, for example: glottochronology (M. Swodesh), linguistic typology (J. Greenberg), structural-mathematical and applied linguistics (V. Perebyjnis), etc. At the end of the XX century – the beginning of the XXI century there are attempts to describe the constitutive principles of quantitative linguistics (M. Arapov, S. Buck, N. Darchuk, P. Fletcher, A. Hughes, V. Levitsky, Yu. Tuldava, A. Woods etc.). The relevance of the article is conditioned by the need to characterize the specifics of quantitative methods in the linguistic works of recent years.*

The purpose is to determine the latest trends in the application of quantitative methods in modern linguistic paradigm with designing for the higher school educational process.

Linguistic works of recent years have intensified the active use of quantitative methods in various linguistic studies on language personality (I. Danyliuk), the lexico-semantic macro-typology of languages (A. Kretov, A. Voevudsky, I. Merkulova, V. Titov), the principle of iconicity in the language (T. Kozlova), constructive grammar (H. Sytar), stylometry with elements of psycholinguistics (O. Pavlyshenko).

The active use of quantitative methods in linguistics makes it possible to develop the university course «Quantitative linguistics» (field of knowledge 03 «Humanities», specialty 035 «Philology») which consists of two content modules «Quantitative linguistics: qualification bases» and «Experience in applying quantitative methods in linguistics».

Thus, the use of quantitative methods in domestic and foreign linguistics has its own traditions and modern development within the framework of language personality, the lexico-semantic macro-typology of languages, the principle of iconicity in the language, constructive grammar, stylometry with elements of psycholinguistics that can be effectively applied in the educational process of higher education, in scientific topics and projects.

We see the perspective in determining the features of modern systemic quantification of phonetics and phonology, vocabulary and phraseology, grammar, etc.

Key words: *quantitative linguistics, language personality, lexico-semantic macro-typology of the language, principle of iconicity in the language, constructive grammar, stylometry.*

Постановка проблеми та обґрунтування актуальності її розгляду. Зацікавлення квантитативною парадигмою в лінгвістиці триває вже не одне століття (див. праці В. Богородицького, О. Пєшковського, Л. Шермана та ін.). На початку XX ст. здійснено спроби статистичного опрацювання тексту, що полягають у підрахунках слововживань у творах різних авторів або творах одного автора та порівнянні отриманих даних (див.: А. Марков «Приклад статистичного дослідження над текстом «Євгенія Онєгіна», який ілюструє зв'язок випробувань у ланцюг» («Пример статистического исследования над текстом «Евгения Онегина», иллюстрирующий связь испытаний в цепь», 1913), М. Морозов «Лінгвістичні спектри: засіб для відрізнення плагіатів від справжніх творів того чи того відомого автора. Стилеметричний етюд» («Лингвистические спектры: средство для отличения плагиатов от истинных произведений того или иного известного автора. Стилеметрический этюд», 1915). Одним із завдань таких досліджень є встановлення авторства чи плагіату. Активізація вітчизняних і закордонних лінгвостатистичних досліджень відбувається в другій половині XX ст. із появою ряду фундаментальних праць, що зумовили започаткування нових напрямів мовознавства, серед яких: *глотохронологія* (М. Сводеш «Лексикостатистичне датування доісторичних етнічних контактів на матеріалі племен північноамериканських індійців та ескімосів» («Lexicostatistic Dating of Prehistoric Ethnic Contacts with Special Reference to North American Indians and Eskimos», 1952), «До питання про підвищення точ-

ності в лексикостатистичному датуванні» («Towards Greater Accuracy in Lexicostatistic Dating», 1955)); *лінгвістична типологія* (Дж. Грінберг «Квантитативний підхід до морфологічної типології мов» («A Quantitative Approach to the Morphological Typology of Language», 1960)); *структурно-математична та прикладна лінгвістика*, зокрема створення теоретичної та методико-процедурної бази для впровадження комп'ютерних технологій в українське мовознавство (В. Перебийніс «Статистичні параметри стилів», 1967), а також перший опис фонологічної системи української мови за допомогою структурно-лінгвістичних методів «Кількісні та якісні характеристики системи фонем української мови», 1970). Наприкінці XX ст. з'являються спроби опису конститутивних засад квантитативної лінгвістики: О. Надточий «Статистична діагностика авторських відмінностей у синтаксисі» («Статистическая диагностика авторских различий в синтаксисе», 1983), Ю. Тулдава «Проблеми та методи квантитативно-системного дослідження лексики» («Проблемы и методы квантитативно-системного исследования лексики», 1987); М. Арапов «Квантитативна лінгвістика» («Квантитативная лингвистика», 1988); Е. Вудс, П. Флетчер, А. Х'юз «Статистика в лінгвістичних дослідженнях» («Statistics in Language Studies», 1996). Актуальність статті зумовлена потребою схарактеризувати специфіку квантитативних методів у лінгвістичних працях останніх років.

Аналіз останніх досліджень і публікацій. Початок XXI ст. визначуваний появою теоретико-прикладних і лексикографічних праць із квантитативної лінгвістики: Ш. Еверт «Кембриджський словник статистики» («The Cambridge Dictionary of Statistics», 2002), В. Перебийніс «Статистичні методи для лінгвістів» (2002; 2013), В. Левицький «Квантитативні методи в лінгвістиці» («Квантитативные методы в лингвистике», 2007), С. Бук «Основи статистичної лінгвістики», 2008), Н. Дарчук «Комп'ютерна лінгвістика (автоматичне опрацювання тексту)», 2008) та ін.

Основним осередком опрацювання проблем квантитативної лінгвістики у вітчизняній науці є відділ структурно-математичної лінгвістики (завідувач – проф. Є. Карпіловська) Інституту української мови НАН України. В історії відділу виділяють три етапи, кожен із яких позначений полівекторністю досліджень мови: 1) утвердження структурних, статистичних методів, методу моделювання у вивченні мов; 2) підготовка узагальнювальних праць зі структурної граматики української мови, розбудова принципів статистичного дослідження українськомовних текстів різних функціональних стилів, укладання частотних словників різних стилів української мови; 3) розроблення систем автоматизованого аналізу російсько- та українськомовних текстів, створення комп'ютерного морфемно-словотвірного фонду української мови та укладання на їхній основі нових словників української мови. Нині відділ працює над дослідженням інноваційних процесів у сучасній українській мові, їх комп'ютерним моделюванням, вивченням впливу соціодинаміки на систему та структуру мови, тенденцій мовних змін.

Формулювання мети і завдань статті. Мета статті полягає у визначенні новітніх тенденцій застосування квантитативних методів у сучасній лінгвопарадигмі з проектуванням на освітній процес вищої школи. Успішна реалізація мети передбачає розв'язання таких **завдань**: 1) розглянути найновіші лінгвістичні дослідження з використанням квантитативних методів; 2) окреслити місце цих розвідок у структурі вишівського курсу «Квантитативна лінгвістика».

Дослідження виконано у рамах наукового проекту «Об'єктивна і суб'єктивна мовносоціумна граматика : комунікативно-когнітивний та прагматико-лінгвокомп'ютерний виміри» (0118U003137) Донецького національного університету імені Василя Стуса.

Виклад основного матеріалу дослідження. Лінгвістичні розвідки останніх років активізують застосування квантитативних методів у різних мовознавчих студіюваннях.

1. *Лінгвоперсонологія.* Опрацювання методологійних засад лінгвоперсонології пов'язане з використанням методу машинного навчання[1] та поняттям 'кольорової мапи', що передбачають апелювання до певних квантитативних параметрів. Кольорову мапу І. Данилюк вважає автоматизованою інфографікою для візуалізації мовленнєвих даних [2, 84] й інтерпретує як множину кольорових квадратиків (або інших фігур), кожен із яких представляє конкретну кольороназву в оригінальному тексті [1, 78]. Певне використання прикметника на позначення кольору – «білий», «чорний», «червоний», «золотий» – у кольоровій мапі буде передано квадратиком відповідного кольору. Це дасть змогу отримати повне й абсолютно об'єктивне представлення лексики конкретного тексту і певні риси «картини світу», концепти окремих кольорів у літературних творах. Одним із завдань для автоматичного генерування кольорової мапи мовознавець визначає отримання всієї можливої статистичної інформації з тексту для подальшого аналізу (а також створити процедури (мовою для Mathematica), щоб працювати з окремими словами та реченнями; побудувати мовну модель для називання кольору з урахуванням відмінювання й окремих випадків словотвору; використати модель для генерування кольорової мапи певного тексту). На сьогодні І. Данилюк у середовищі Mathematica репрезентував кольорові мапи тексту «Кобзаря» Тараса Шевченка [14], поетичного спадку Василя Стуса [1], роману «Криничар» Мирослава Дочинця [2], трилогії «Волинь» Уласа Самчука [3].

2. *Лексико-семантична макротипологія мов.* Важливий теоретичний і прикладний характер для формування нового розділу теоретичної лінгвістики – лексико-семантичної макротипології мов як частини лінгвістичної типології – визначає монографію

[1]Метод машинного навчання опрацьовувано в рамах наукового проекту «Об'єктивна і суб'єктивна мовносоціумна граматика : комунікативно-когнітивний та прагматико-лінгвокомп'ютерний виміри» (0118U003137).

О. Кретова, О. Воєвудської, І. Меркулової, В. Титова «Єдність Європи за даними лексики»[2] (2016), що присвячена лексико-семантичному устрою державних мов Європи і має на меті виокремлення ядер лексико-семантичних систем мов світу. Об'єктивність і достовірність результатів дослідження забезпечувані джерельною базою – малі, середні та великі одно- й двомовні словники 35 державних мов Європи, що містять основну комунікативно значущу лексику, – і адекватним комплексом наукових методів, провідне місце з-поміж яких посідає параметричний метод аналізу лексики. Отже, мовознавці цілком переконливо констатують, що «в дзеркалі «ядерних» ексклюзем центром і домінантою ментального простору Європи постають германські мови, які наділені безпрецедентною внутрішньою згуртованістю, що в чотири рази перевищує внутрішню згуртованість слов'янських мов, і високою аттрактивністю: поза їхнім впливом виявилася одна мова з 26 можливих – португальська» [7, 394].

3. *Принцип іконічності в мові.* Наукові надбання з квантитативної лінгвістики В. Левицького та В. Кушнерика (див. фоносемантичний принцип) активно використовувані нині Т. Козловою під час студіювання мовної іконічності в діахронії [5], у якому проаналізовано структуру й семантику індоєвропейських коренів, установлено відповідності між їх матеріальною оболонкою та внутрішнім наповненням, розкрито еволюційно-адаптивну роль іконічності в процесах мовної еволюції, висвітлено мотивованість фонологічного контрасту кентум-сатем, а також виявлено важливість звукозображання в системній організації мовної картини світу. Дослідження принципу іконічності ґрунтовано на залученні обширного фактичного матеріалу, що охоплює більше двохсот мов і різних діалектів та наріч усіх груп індоєвропейських мов.

Розгляд фонетичної структури коренів здійснювано за допомогою встановлення частотності кореневих ініціалей, з'ясування синтагматики фонем у праіндоєвропейських коренях, вияву гомогенності синтагм і простеження гомогенності синтагм у співвідношенні з порядком розташування маргінальних приголосних у досліджуваних праіндоєвропейських коренях. Мотивованою, на думку А. Загнітка, постає теза про те, що «…структура праіндоєвропейського кореня тяжіє до симетричної консонантної рамки навколо вокалічного ядра» (с. 181 дис.) із визначенням основних напрямків структурної модифікації кореня і/чи праіндоєвропейських коренів та окресленням залежності останніх від ступеня узагальненості значень» [4, 129]. Заявлений аналіз ґрунтовано на встановленні моделей фонетичної структури етимонів і констатуванні, що «послідовність {CVC} забезпечує високу частотність коренів» [6, 202].

4. *Конструкційна граматика*. Г. Ситар здійснює студіювання синтаксичних фразеологізмів у розрізі конструкційної граматики [13]. Мовознавець опрацювала статистичні критерії аналізу синтаксичних фразеологізмів; апробувала статистичний аналіз фразеологізованих речень (показник асоціації 'mutual information'), обчислення показника асоціації MI log Freq як статистичний метод дослідження синтаксичних фразеологізмів тощо. Статистичний аналіз синтаксичних фразеологізмів здійснювано за даними Українського національного лінгвістичного корпусу. Загалом використано 180 мільйонів слововживань.

'Mutual information' (або MI-score) Г. Ситар розуміє як коефіцієнт, який відбиває невипадковість (залежність) певної послідовності слів у тексті [11, 106]. Показник асоціації MI (обчислення MI ввели американські дослідники К. В. Чарч і П. Хенкс), на думку лінгвіста, видається придатним для визначення коректності виділення фразеологізованої моделі речення та вірогідності встановлення стійкості поєднання двох або більше словоформ у межах незмінного компонента моделі речення. Статистичний критерій MI відповідає таким вимогам: дає змогу визначити коефіцієнт невипадковості поєднання двох і більше слів у корпусі текстів, враховує частоту конструкції, частоту її компонентів, розмір корпусу та має формулу в узагальненому вигляді для конструкцій з будь-якою кількістю компонентів [12, 114]. Статистичний аналіз забезпечив правильність висунутої гіпотези про наявність високого ступеня (>>3) невипадковості поєднання словоформ у межах незмінного компонента всіх обстежених моделей фразеологізованих речень. Друга гіпотеза [11, 104] підтвердилася частково: відмінності в показниках невипадковості появи словоформ виявлено в різних групах мовних одиниць – лексичних фразеологізмів, нефразеологізованих речень і фразеологізованих речень – тільки для трикомпонентних конструкцій.

5. *Стилеметрія з елементами психолінгвістики.* У стилеметричному студіюванні О. Павлишенко [8] успішно реалізовано мету – дослідити за допомогою кількісно-дистрибутивних методик частотну параметризацію лексико-семантичних полів дієслова англійської мови та виявити характерні статистичні особливості розподілу лексем цих полів у текстах. Методологійним підґрунтям дослідження стали запропоновані в лінгвістиці основи порівняння словникового складу текстів (В. Левицький, І. Носенко, Р. Фрумкіна), квантитативного дослідження стилів (В. Перебийніс, Г. Мартиненко, В. Горєва, М. Култхард, М. Лоуверс), а також принципи побудови частотних словників (П. Алексєєв, Р. Фрумкіна, А. Скляревич, Т. Якубайтіс, Ю. Тулдава) та лексико-семантичних полів (Г. Міллер, К Фельбаум, З. Вердієва, Е. Кузнєцова). Комплексне застосування квантитативних методів дало змогу О. Павлишенко вперше: а) дослідити квантитативну структуру лексико-семантичних полів дієслова в об'ємній електронній базі авторських текстів англомовної літератури; б) виявити константи поділу ієрархічної структури лексико-семантичних полів на ядро, близьку та віддалену периферію; в) увести поняття семантичної відстані між вибіркою авторських текстів та лінгвостилістичною нормою; г) запропону-

[2]Монографія ґрунтована на результатах НДР «Дослідження єдності Європи за даними лексики», яку виконано у Воронезькому державному університеті в межах державного завдання в 2012–2013 рр.

вати методику виявлення маркерів авторського ідіолекту в частотній структурі лексико-семантичних полів; ґ) інвентаризувати та проаналізувати субмодальні психо-семантичні поля дієслова в авторському лексиконі; д) проаналізувати стилеметричну значущість частотного розподілу семантичних полів дієслова; е) виявити ядро та периферію семантичної подібності текстів в авторському стилі; є) дослідити синонімічний обсяг лексем у частотному розподілі семантичних полів дієслова в авторських текстах.

Активне використання квантитативних методів у лінгвістиці актуалізує опрацювання вишівського курсу «Квантитативна лінгвістика» (пор. також: навчальні посібники з математичної лінгвістики: для студентів, що навчаються за напрямами галузей знань «Інформатика та обчислювальна техніка» (напрям «Комп'ютерні науки»), «Системні науки та кібернетика» (напрям «Системний аналіз») і споріднених галузей знань, пов'язаних із вивченням прикладної лінгвістики та інформаційних технологій [9]; для відділень структурно-математичної, прикладної та комп'ютерної лінгвістики [10]). Курс «Квантитативна лінгвістика» (галузь знань 03 «Гуманітарні науки», спеціальність 035 «Філологія») належить до циклу математичної та природничо-наукової підготовки навчального плану бакалаврів філології (прикладної лінгвістики) (освітня програма «Прикладна лінгвістика» / «Applied Linguistics»), передбачає розширення та систематизацію теоретико-прикладних знань слухачів курсу із застосування квантитативних методів у лінгвістичних дослідженнях і містить два змістових модулі:

1) змістовий модуль «Квантитативна лінгвістика: кваліфікаційні основи» (теми «Квантитативна лінгвістика: теоретичні засади», «Квантитативна лінгвістика: етапи становлення та розвитку», «Квантитативні методи», «Організація вибірки», «Індекси та коефіцієнти у квантитативній лінгвістиці», «Поняття частоти у квантитативній лінгвістиці», «Репрезентація квантитативних досліджень»);

2) змістовий модуль «Досвід застосування квантитативних методів у лінгвістиці» (теми «Квантитативні дослідження у фонології та фонетиці», «Квантитативні дослідження у морфеміці, словотворі та морфології», «Квантитативні дослідження у лексиці та фразеології», «Квантитативні методи у лексикографії», «Квантитативні дослідження у синтаксисі», «Квантитативні дослідження у стилістиці», «Квантитативні дослідження в сучасній лінгвопарадигмі»).

У результаті вивчення навчальної дисципліни «Квантитативна лінгвістика» студент повинен: 1) *набути* таких результатів навчання: а) систематизувати знання із загальної теорії квантитативної лінгвістики; б) опанувати навички формування різних типів виборок; в) опанувати технологію застосовування квантитативних методів під час дослідження одиниць різних мовних рівнів; 2) *знати:* а) основні ідеї та концепції вітчизняних і закордонних учених у галузі квантитативної лінгвістики; б) базові терміни квантитативної лінгвістики; в) алгоритми реалізації квантитативних методів у лінгвістиці; г) рівні квантитативної лінгвістики; ґ) широкомасштабні лінгвоквантитативні дослідження в Україні; д) особливості організації різних типів виборок; е) основні різновиди індексів і коефіцієнтів та правила їхнього обчислення; є) специфіку укладання частотних словників; ж) можливості застосування квантитативних методів під час досліджень із фонетики, фонології, морфеміки, словотвору, морфології, лексики, фразеології, синтаксису, стилістики, а також у дослідженнях з психолінгвістики, соціолінгвістики, лінгвоперсонології, лексико-семантичної макротипології мов, конструкційної граматики тощо; 3) *уміти*: а) визначити генеральну сукупність для різних типів квантитативних досліджень мови і мовлення залежно від поставленої мети і задач; б) формувати випадкові, механічні та зональні вибірки; в) встановлювати різноманітні індекси та коефіцієнти (часткові індекси, індекси перебігу певної якості, коефіцієнт варіації, коефіцієнт полісемантичності слів, ранговий коефіцієнт, коефіцієнт різноманітності тощо).

Висновки та перспективи подальших досліджень у цьому напрямі. Отже, використання квантитативних методів у вітчизняному та закордонному мовознавстві має власні традиції та сучасні напрацювання в рамах лінгвоперсонології, лексико-семантичної макротипології мов, принципу іконічності в мові, конструкційної граматики, стилеметрії з елементами психолінгвістики, що можуть ефективно застосовуватися в освітньому процесі вищої школи, у наукових темах і проектах.

Перспективу вбачаємо у визначенні особливостей сучасної системної квантифікації фонетики та фонології, лексики та фразеології, граматики тощо.

ЛІТЕРАТУРА

1. Данилюк І. Кольорова мапа поетичного спадку Василя Стуса у MATHEMATICA / Ілля Данилюк // Вісник Донецького національного університету. Сер. Б : Гуманітарні науки. – Донецьк : ДонНУ, 2014. – № 1–2. – С. 78–83.

2. Данилюк І. Кольорова мапа роману «Криничар» Мирослава Дочинця у Mathematica / Ілля Данилюк // Граматичні студії : [зб. наук. праць]. – Вінниця : ДонНУ, 2016. – Вип. 2. – С. 84–88.

3. Данилюк І. Г. Кольорова мапа трилогії «Волинь» Уласа Самчука у Mathematica / І. Г. Данилюк // Лінгвокомп'ютерні дослідження : [зб. наук. праць]. – Вінниця : ДонНУ, 2016. – Вип. 9. – С. 111–121.

4. Загнітко А. Принцип іконічності в порівняльно-історичній лінгвопарадигмі / Анатолій Загнітко // Лінгвокомп'ютерні дослідження : [зб. наук. праць]. – Вінниця : ДонНУ, 2018. – Вип. 11. – С. 121–134.

5. Козлова Т. О. Іконічність у лексиці індоєвропейської прамови : [монографія] / Т. О. Козлова. – Запоріжжя : Кругозір, 2015. – 640 с.

6. Козлова Т. О. Принцип іконічності та його реалізація в лексиці індоєвропейської прамови : дис. … д-ра філол. наук : спец. 10.02.15 «Загальне мовознавство» / Тетяна Олегівна Козлова. – К., 2017. – 501 с.

7. Кретов А. Единство Европы по данным лексики : [монография] / Алексей Кретов, Оксана Воевудская, Инна Меркулова, Владимир Титов. – Воронеж : Издательский дом ВГУ, 2016. – 412 с.

8. Павлишенко О. А. Квантитативні характеристики лексико-семантичних полів дієслова в авторських текстах англомовної художньої літератури : автореф. дис. ... канд. філол. наук : спец. 10.02.04 «Германські мови» / О. А. Павлишенко. – Львів, 2017. – 25 с.

9. Пасічник В. Математична лінгвістика : Кн.1. Квантитативна лінгвістика : [навч. посіб.] / Володимир Пасічник, Юрій Щербина, Вікторія Висоцька, Тетяна Шестакевич. – Львів : Новий Світ-2000, 2012. – 359 с.

10. Перебийніс В. Математична лінгвістика : [навч. посіб.] / Валентина Перебийніс. – К. : Вид. центр КНЛУ, 2014. – 125 с.

11. Ситар Г. Статистичний аналіз фразеологізованих речень : показник асоціації mutual information / Ганна Ситар // Українське мовознавство. – Вип. 1 (46). – К. : КНУ імені Тараса Шевченка, 2016. – С. 103–125.

12. Ситар Г. Обчислення показника асоціації MI log Freq як статистичний метод дослідження синтаксичних фразеологізмів / Ганна Ситар // Гуманітарна освіта в технічних вищих навчальних закладах : [зб. наук. праць]. – К. : Університет «Україна», 2016. – Вип. 34. – С. 114–125.

13. Ситар Г. В. Синтаксичні фразеологізми в розрізі конструкційної граматики : [монографія] / Г. В. Ситар. – Вінниця : ТОВ «Нілан-ЛТД», 2017. – 458 с.

14. Danyliuk I. Color Map of Taras Shevchenko's «Kobzar» with Mathematica / Illja Danyliuk // Лінгвістичні студії / Linguistic Studies : [зб. наук. праць]. – Вінниця : ДонНУ, 2014. – Вип. 29. – С. 218–223.