DOI 10.36074/logos-06.02.2026.035

ЛІНГВОМЕТРІЯ В СИСТЕМІ СУЧАСНОЇ КОМП'ЮТЕРНОЇ ТА МАТЕМАТИЧНОЇ ЛІНГВІСТИКИ: МЕТОДОЛОГІЧНІ ЗАСАДИ ТА НАПРЯМИ РОЗВИТКУ

Нагорна Світлана Сергіївна[1]

1. кандидат філологічних наук
Доцент кафедри англійської філології
Київський національний лінгвістичний університет, УКРАЇНА
ORCID ID: 0009-0009-0407-2590

Останні десятиліття розвитку мовознавства характеризуються активним становленням інженерно-лінгвістичної методології дослідження мови. Цей процес зумовлений, з одного боку, прагненням створювати формалізовані моделі, здатні адекватно відтворювати реальні мовні факти, а з іншого - об'єктивною потребою в розробленні ефективних систем автоматичної обробки текстової інформації та їх упровадженні в практику. У цьому контексті комп'ютерна лінгвістика набуває двоєдиного характеру: вона охоплює як загальнометодологічні питання побудови інженерно-лінгвістичного експерименту для перевірки лінгвістичних гіпотез, так і розроблення конкретних алгоритмів автоматичної обробки мовних даних.

Лінгвометрія як кількісний напрям мовознавства відіграє ключову роль у цьому процесі, оскільки забезпечує інструментарій для виявлення статистичних закономірностей мовної системи та мовлення. Її розвиток безпосередньо пов'язаний із практичними завданнями корпусної лінгвістики, комп'ютерної лексикографії, машинного перекладу, стилометрії, інформаційного пошуку та автоматизованого аналізу текстів.

Лінгвометрія сформувалася на перетині математичної лінгвістики, статистики та інформаційних наук. **Методологічні засади** кількісного підходу до мови були закладені у працях Дж. Ціпфа, який сформулював закон частотності лексем і продемонстрував підпорядкованість мовних одиниць стабільним статистичним закономірностям. Подальший розвиток цих ідей у

роботах Г. Мандельброта привів до використання фрактальних моделей для опису мовних процесів та до усвідомлення самоорганізаційної природи мовних даних.

У 1960–1970-х роках статистична лінгвістика зосередилася на вимірюваності мовних явищ, частотних розподілах морфологічних і лексичних одиниць, що згодом стало основою корпусного підходу (Х. Баайен, Б. Кендел). Наприкінці XX — на початку XXI століття розвиток національних і багатомовних корпусів сприяв стандартизації кількісних методів та появі нових індексів лексичної різноманітності (TTR, MTLD, HD-D), які широко застосовуються в сучасних дослідженнях.

В українському мовознавстві лінгвометричні студії представлені працями Л. Масенко, О. Стишова, А. Загнітка, І. Кочан, де аналізуються частотні параметри слововживання, мовна норма, стилістична маркованість і корпусна репрезентативність. Водночас залишаються недостатньо опрацьованими питання інтерпретації статистичних результатів у контексті мовної норми та інтеграції кількісних і якісних методів аналізу.

Метою статті є систематизація методологічних засад лінгвометрії як міждисциплінарного напряму та визначення її ролі у сучасній комп'ютерній і математичній лінгвістиці, а також обґрунтування необхідності застосування статистичних методів для виявлення закономірностей функціонування мовних одиниць.

Дослідження має **теоретико-методологічний характер** і базується на аналізі класичних та сучасних лінгвометричних праць, частотних словників, корпусних досліджень і моделей статистичної лінгвістики. Застосовано методи теоретичного узагальнення, порівняльного аналізу, систематизації наукових підходів, а також елементи статистичної інтерпретації мовних даних, описаних у попередніх емпіричних дослідженнях.

Особливу увагу приділено проблемі визначення одиниць підрахунку, формалізації критеріїв їх виділення та перевірюваності отриманих результатів. Методологічною основою виступає принцип нульової гіпотези, що використовується для зіставлення вибірок у стилістичних, корпусних і психолінгвістичних дослідженнях.

Мова як складна система дискретних одиниць характеризується ймовірнісною природою та підпорядковується дії статистичних законів. Це проявляється на всіх мовних рівнях — від фонемного до синтаксичного. Повторюваність одиниць, масовість мовленнєвих реалізацій і багатофакторність впливів зумовлюють неможливість детермінованого опису мовних процесів без використання статистичних методів.

Аналіз частотних співвідношень дозволяє виявляти закономірності типу «закону переваги», відповідно до якого невелика кількість високочастотних одиниць становить ядро мовної системи, тоді як більшість одиниць є низькочастотними. Подібні закономірності мають істотне значення для порівняльно-історичного мовознавства, типології мов, стилометрії та автоматичної обробки текстів.

Водночас ефективність лінгвометричного аналізу залежить від чіткого визначення одиниць підрахунку. Недостатня формалізація понять призводить до непорівнюваності результатів і дискредитації статистичних методів. Саме тому ключовим принципом є прозорість методики, відтворюваність і можливість перевірки результатів іншими дослідниками

Попри стрімкий розвиток лінгвометричних підходів, залишається низка проблемних аспектів. Недостатньо дослідженими є питання інтерпретації статистичних закономірностей у контексті мовної норми та варіативності. Потребують додаткової уваги методи інтеграції кількісних моделей із якісним лінгвістичним аналізом, зокрема при описі семантичної структури текстів. В українському контексті актуальною є також проблема недостатньої репрезентативності корпусів, що впливає на надійність отриманих лінгвометричних показників.

Таким чином, сучасний стан дослідження лінгвометрії характеризується міждисциплінарністю, широким застосуванням цифрових технологій та прагненням до створення універсальних моделей мовних процесів. Подальші розробки у цій сфері передбачають інтеграцію когнітивних, комп'ютерних і статистичних методів, що дасть змогу поглибити розуміння структурної організації мови та механізмів її функціонування у реальному мовленні.

Ще у 80-і роки намітилися два принципових підходи до побудови лінгвістичного забезпечення інформаційних систем: (а) створення порівняно простих систем, що використовують мінімально необхідні дані про мову; (б) створення складних систем, що використовують максимально можливі дані. Нам здається марною суперечка щодо монополії одного з цих підходів; конструктивними представляються вияв класів лінгвістичних завдань, для рішення яких той чи інший підхід є оптимальним, і розумний компроміс між ними в тих випадках, коли він є необходіним. Не менш важливо, з нашої точки зору, виявити коло завдань, рішення яких вимагає людського інтелекту й принципово погано піддається алгоритмізації.

Лінгвометрія - галузь науки на межі математики й лінгвістики, що вивчає найзагальніші закони будови символьних послідовностей, або знакових систем, до яких належать деякі абстрактні математичні структури, штучні та природні мови. Інколи розмежовують математичну лінгвістику як галузь

математики і математичну лінгвістику як розділ мовознавства, підкреслюючи при цьому, що між ними існує тісна взаємодія, бо вони використовують той самий понятєвий апарат. Отже, можна вважати математичну лінгвістику єдиною семіотичною дисципліною, яка досліджує форм. методами - алгебричними, теоретико-множинними, логіко-математичними - деякі властивості символьних послідовностей, напр., властивість бути членом певного класу послідовностей, здатність перетворюватися на послідовності іншої будови, властивість взаємозаміни між деякими ланцюжками символів тощо. Основними поняттями, що використовуються в математичній лінгвістиці, вважають:

1) множинність вихідних символів (алфавіт);

2) відношення між елементами алфавіту, що сприймаються як аксіоми (постулюються);

3) правила виводу, тобто обчислення всіх можливих множин символьних ланцюжків;

4) ізоморфізм, тобто одно-однозначні відношення між елементами послідовності, при яких кожному елементові однієї послідовності ставиться у відповідність елемент ін. послідовності;

5) гомоморфізм, одно-багатозначні відношення, коли одному елементу першої послідовності відповідає кілька елементів другої і навпаки;

6) відмічений (маркований) ланцюжок, тобто такий, що відповідає правилам виводу (граматично правильний, допустимий);

7) входження символу в послідовність, тобто поява його на заданому місці в ланцюжку;

8) поділ вихідної множини класу ланцюжків за певними правилами на підкласи.

Використання операцій, що базуються на цих поняттях, дає можливість одержати аналоги граматичних класів і підкласів, категорій, парадигм, синтаксичних одиниць та відношень. Властивості відношення одиниць досліджуваної знакової системи виявляють і вивчають шляхом побудови синтезувальних (породжувальних, дедуктивних) й аналітичних (індуктивних) математичних моделей. Важливим етапом використання математичної моделі та її елементів і операцій є інтерпретація їх у термінах певної мови. Інтерпретувати модель значить поставити у відповідність кожному елементу, правилу, відношенню, поняттю, використовуваному в моделі, клас одиниць, правило, категорію, поняття природної мови. При інтерпретації моделі між нею і мовою можуть бути як ізоморфні, так і гомоморфні відношення. Методи і положення математичної лінгвістики є теоретичною базою для створення алгоритмічних мов, для побудови систем автоматичного опрацювання

мовного матеріалу: машинного перекладу, інформаційного пошуку, автоматизації видавничих процесів, реферування й анотування наук, літератури, створення термінологічних банків, машинних фондів різних мов, автоматичного укладання словників, машинного розпізнавання і синтезу усного мовлення та ін.

Сфера дії статистичних законів надзвичайно широка: усі складні системи підпорядковуються, перш за все, законам статистичним. А такими системами є і живі організми, і їхня поведінка, й економіка, і результати діяльності великих колективів, і розвиток науки, і мова.

Важко знайти лінгвістичну роботу, в якій узагалі не було б елементарних підрахунків. Кількісні методи у лінгвістиці допомагають правильно організувати лінгвістичні спостереження, забезпечити надійність, точність, достовірність висновків у науці про мову. Ці методи ввійшли до мовознавчої практики, а лінгвостатистика як наука нараховує 30-річний досвід.

Лінгвометрія розглядається і як техніка обробки лінгвістичних даних, і як метод дослідження мови та мовлення, і як концепція, система ідей та уявлень про об'єкт лінгвістичної науки. І якщо техніка квантитативної лінгвістики використовується зараз у багатьох дослідженнях, то методика реалізується тільки тоді, коли дослідник відчуває, що лише за допомогою кількісного підходу можна одержати нові дані або перевірити набуті знання про лінгвістичний об'єкт, коли дослідник переконаний в імовірнісній природі лінгвістичного об'єкта і ставить перед собою завдання - описати його у кількісних показниках.

Можливість використання кількісних методів у мовознавстві базується на особливостях будови мови та мовлення. Мова - це система, яка складається з дискретних одиниць, що мають кількісні характеристики. Ці характеристики притаманні одиницям усіх рівнів. Мова має ймовірнісний характер: це код з імовірнісними обмеженнями. У загальному розумінні код - це засіб подання інформації у формі, здатній для передавання інформації каналами зв'язку. Будь-який код - це певна множина фізично різних знаків, кожен з яких може співвідноситися з тим чи іншим об'єктом з множності об'єктів, на які розповсюджується дія даного коду.

Мовлення є реалізацією системи мови, її елементів. Можна вказати на декілька факторів, що дозволяють застосовувати кількісні методи при дослідженні мовних та мовленнєвих даних:

1) дискретність одиниць;

2) масовість мовних одиниць (а у мовленні практично нескінченний ланцюжок);

3) повторюваність їх у висловлюваннях;

4) можливість вибору певного елемента з ряду однорідних. На мовлення впливають закони мови (закономірності будови одиниць мови, використання їх у мовленні), закони сполучуваності одиниць у мовленні, закони жанру, теми висловлювання, смаки автора, його психофізіологічний стан та ін. Дія цих факторів так переплітається, що інколи неможливо визначити результати їхнього впливу. Але якщо сукупність цих факторів відносно постійна, то будова мовлення буде характеризуватися такими рисами, які можуть розкриватися кількісними методами.

Основним завданням лінгвометрії є застосування кількісних методів для розкриття закономірностей функціонування одиниць мови у мовленні, а також установлення закономірностей будови тексту.

Прості кількісні співвідношення між словами, складами і фонемами дозволяють дати класифікацію мов, яку можна використати і при вивченні їх історії. Так, у випадку, якщо слова в мові односкладні, середня довжина слова занадто мала для того, щоб було можливим членування слова на частини; тому в мовах з односкладовими словами слово не ділиться на морфеми. Разом з тим необхідність розрізнення кількох тисяч морфем (слів) при обмеженому інвентарі складів робить необхідним розрізнення складів за допомогою музичного наголосу. Тому мови з односкладовими словами-морфемами завжди є мовами з музичним наголосом (в'єтнамська, класична китайська, деякі центрально-африканські мови і т. п.).

Висновки та перспективи подальших досліджень

Лінгвометрія є важливим теоретичним і прикладним напрямом сучасного мовознавства, що забезпечує інструментарій для виявлення глибинних закономірностей мовної структури та мовлення. Її розвиток підтверджує статистичну природу мовних процесів і відкриває нові можливості для інтеграції лінгвістики з комп'ютерними та когнітивними науками.

Перспективи подальших досліджень пов'язані з удосконаленням корпусної бази української мови, розробленням методів інтерпретації кількісних показників у межах мовної норми та варіативності, а також із поєднанням лінгвометрії з методами машинного навчання й когнітивного моделювання.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ:

[1] Zipf G. K. Human Behavior and the Principle of Least Effort. Cambridge, MA : Addison-Wesley, 1949.
[2] Mandelbrot B. The Fractal Geometry of Nature. New York : W. H. Freeman, 1982.
[3] Baayen R. H. Word Frequency Distributions. Dordrecht : Kluwer Academic Publishers, 2001.
[4] Baayen R. H. Analyzing Linguistic Data: A Practical Introduction to Statistics using R. Cambridge : Cambridge University Press, 2008.

[5] Biber D., Conrad S., Reppen R. Corpus Linguistics: Investigating Language Structure and Use. Cambridge : Cambridge University Press, 1998.
[6] Manning C. D., Schütze H. Foundations of Statistical Natural Language Processing. Cambridge, MA : MIT Press, 1999.
[7] Gries S. Th. Quantitative Corpus Linguistics with R: A Practical Introduction. New York : Routledge, 2009.
[8] Jurafsky D., Martin J. H. Speech and Language Processing. 3rd ed. Draft. Stanford University, 2023.
[9] Köhler R., Altmann G., Piotrowski R. (eds.) Quantitative Linguistics: An International Handbook. Berlin ; New York : Walter de Gruyter, 2005.
[10] Altmann G. Aspects of Language Quantification. Bochum : Brockmeyer, 1988.
[11] McCarthy P. M., Jarvis S. MTLD, Vocd-D, and HD-D: A Validation Study of Sophisticated Approaches to Lexical Diversity Assessment // Behavior Research Methods. 2010. Vol. 42. P. 381–392.
[12] Sinclair J. Corpus, Concordance, Collocation. Oxford : Oxford University Press, 1991.
[13] Stubbs M. Text and Corpus Analysis. Oxford : Blackwell, 1996.
[14] Масенко Л. Т. Мова і суспільство. Київ : Либідь, 2004.
[15] Стишов О. А. Українська лексика кінця XX століття. Київ : КМ Academia, 2003.
[16] Загнітко А. П. Теоретична граматика української мови. Донецьк : ДонНУ, 2011.
[17] Кочан І. М. Лінгвістичний аналіз тексту. Львів : ЛНУ імені Івана Франка, 2014.
[18] Дарчук Н. П. Корпусна лінгвістика. Київ : Видавничий дім Дмитра Бураго, 2010.
[19] Широков В. А., Дарчук Н. П. Комп'ютерна лінгвістика. Київ : Наукова думка, 2011.