

рмації будуть оказувати вирішальне впливання на організацію комунікативного пространства, якість контенту, видвигаючи нові вимоги до професіоналам в сфері комунікації.

Ключеві слова: віртуальна комунікація, Інтернет речей, інформаційні та комунікаційні технології, гіперконнективність, персоналізація Інтернет, конфіденціальна інформація

A. Belova, Dr. Philol. Sci., Full Professor
Institute of Philology, Taras Shevchenko National University of Kyiv

Virtualization of modern communication

The article highlights prospects and trends of virtual communication and cyberspace evolution in the direction of the Internet of Things, hyperconnectivity with the focus on information and communication technology, political and economic factors, behavior models, age preferences. Visualization of communication, Internet personalization, deglobalized future, privacy will make crucial impact on communication, content quality putting forward new requirements for professionals in communication-related fields.

Key words: virtual communication, the Internet of Things, information and communication technology, hyperconnectivity, Internet personalization, privacy

УДК 811.81'324-112=161.2

К. А. Білик, асп.
Інститут філології КНУ імені Тараса Шевченка, м. Київ

ОСОБЛИВОСТІ ЗАСТОСУВАННЯ ФУНКЦІЇ ЛОГАРИФМІЧНОЇ ПРАВДОПОДІБНОСТІ ДЛЯ ПОРІВНЯННЯ СПЕЦІАЛІЗОВАНИХ КОРПУСІВ

Стаття присвячена дослідженню проблем корпусної лінгвістики. Розглядаються способи застосування функції логарифмічної правдоподібності для порівняння корпусів, зокрема корпусу поетичних текстів Р. М. Рільке з референтним корпусом, який знаходиться в обмеженому доступі. Доводиться доцільність застосування функції логарифмічної правдоподібності для пошуку слів, асоційованих із ключовим словом, і розробляється відповідна методологія.

Ключові слова: корпусна лінгвістика, Р. М. Рільке, референтний корпус, нормативний корпус, асоціація.

Корпусна лінгвістика – це метод дослідження великих масивів текстів, добраних з метою всебічного відображення певного

мовного явища. Вона є логічним продовженням методу експерименту та опитування, але, на відміну від них, вивчає мовні явища на автентичних зразках ужитку: матеріал корпусу створюється в реальних умовах, він не обмежений місцем проведення і має значно більший обсяг.

Корпусна лінгвістика – відносно нова методологія, яка викристалізувалася в мовознавстві останньої третини ХХ ст. Є всі підстави вважати її ефективною для дослідження художніх текстів, оскільки вона уможлиблює роботу з загальним масивом праць одного автора й порівняння їх із загальномовними стандартами. Інтерес до корпусної лінгвістики спостерігається в європейському та американському мовознавстві вже протягом кількох десятиліть. Особливо багато уваги її розвитку приділяли американські дослідники (В. Френсіс, Г. Кучера, Д. Байбер, С. Ганстон), однак на сьогодні ця методологія стала загальносвітовою, оскільки вона придатна для роботи з будь-якою мовою. Багато галузей лінгвістики успішно послуговуються здобутками корпусної лінгвістики: лексикографія (Дж. Сінклер, В. Морріс, Г. Кйеллмер, У. Гайд, А. П. Гілі, Г.-М. де Шрювер, О. Норлінг-Крістенсен), перекладознавство й контрастивні студії (С. Йоганссон, Д. Кенні), методика викладання (В. Хенг, Е. Брілл, М. Голлідей, С. Гренджер, Г. Беннетт, С. Конрад, А. Гольдберг), синтезування мовлення (В. Л. Джонсон, С. Нараянан, Р. Вітні, Р. Дас) і, що особливо важливо, когнітивна лінгвістика (Ч. Філлмор, Ш. Гріс, Р. Ленекер, Д. Глінн, К. Фішер, Ш. Вульф, А. Стефанович, Х. Туммерс, Р. Баасен, А. Дейнан, М. Делорж) й семантика (Д. Байбер, Дж. Ліч, Б. Лоув, Дж. Сінклер, Ш. Гріс, А. Стефанович, М. Делорж, Д. Див'як).

Витоки сучасної корпусної лінгвістики зазвичай убачають у конкордансах сакральних текстів (найперші датовано ще ХVІІ століттям), де кожне вжите у Біблії слово наведено з його контекстом. Схожі за структурою публікації існували до 80-х років, зокрема кілька конкордансів текстів Р. М. Рільке було складено саме в ті роки: Р. С. Брауна [3], Дж. Р. Бартлетта [2], Ю. К. Гольдсмита [6]. Із появою та розвитком комп'ютерних технологій стоси паперів було оцифровано, і корпусна лінгвістика тепер асоціюється виключно із цифровими масивами даних.

Першим корпусом у сучасному розумінні вважається Brown corpus of Standard American English, який налічує 1 млн токенів, узятих із різножанрових текстів 1961-го року. Довгий час цей корпус був чи не єдиним (поряд із корпусом Lancaster-Oslo-Bergen), тож велику кількість досліджень засновано саме на ньому. Серед них і The American Heritage Dictionary of the English Language, перша лексикографічна праця, яка використовує корпусно-статистичні дані (частотність, дані про вжиток) [8]. Вважається, що методологія, використана у цьому словнику, знаменує перехід від прескриптивної лінгвістики до дескриптивної.

Стрімкий розвиток комп'ютерної техніки неминуче результував у підвищенні вимог до корпусів, і корпусу обсягом 1 млн слів виявилось замало для повноцінної репрезентації мови. До того ж, з'ясувалося, що корпусна лінгвістика може звертатися до специфічних проблем, а це вимагало складення нових, спеціалізованих (на противагу до загальних) корпусів. Наприклад, з'явилися акустичні (на противагу до письмових) корпуси, корпуси СМС, корпуси текстів, продукованих учнями або студентами (Learner Corpora), паралельні корпуси тощо. Нерідко такі корпуси створювалися з доволі неочікуваною метою: так, було складено корпус розмовної військової мови для того, щоб озвучити анімованого персонажа [9], корпус описів героїв популярних комп'ютерних ігор для вивчення їх наративів (І. Карілло-Масо [5]) або для оцінки різноманітності мови персонажів для подальшого використання гри у якості навчального матеріалу (наприклад Р. Пурушотма [9], Й. А. Ренкін [10], С. Л. Торне [12]); корпус фільмових діалогів став базою як для навчально-методичного матеріалу (М. А. Вокер) [13], так і для створення моделей пліткування у комп'ютерній грі [4] (Дж. Бруск).

Основною перевагою корпусної лінгвістики для вивчення поетичних текстів Р. М. Рільке є те, що він дозволяє одночасно квантитативне і квалітативне дослідження, адже за допомогою корпусного аналізу можна встановити не лише частоту вживання певного слова, але і його безпосередні та більш віддалені коокуренти, їх характер та семантичне наповнення.

Дослідження проводиться на матеріалі укладеного нами спеціального корпусу художніх творів Р. М. Рільке, який охоплює всі основні опубліковані художні твори Р. М. Рільке : "Офіри Ларам" ("Larenopfer", 1895), "11 візій 5 Христа" ("Christus elf Visionen"), "Увінчаний снами" ("Traumgekrönt"), "Мені до свята" ("Mir zur Feier", 1899), "Свят-вечір" ("Advent", 1897), "Книга годин" ("Das Stunden-Buch", 1905), "Книга образів" ("Das Buch der Bilder, 1902, 1906), "Нові поезії" ("Neue Gedichte", 1907, 1908), "Реквієм" ("Requiem", 1909), "Дуїнянські елегії" ("Duineser Elegien", 1923), "Сонети до Орфея" ("Die Sonette an Orpheus", 1923), "Дві празькі оповідки" ("Zwei Prager Geschichten", 1899), "Оповіді про любого Бога" ("Geschichten vom lieben Gott", 1904), "Пісня кохання і смерті корнета Крістофа Рільке" ("Die Weise von Liebe und Tod des Cornets Christoph Rilke", 1906), "Нотатки Мальте Лявридса Бригге" ("Die Aufzeichnungen des Malte Laurids Brigge", 1910). Тексти впорядковано у строгому діяхронічному порядку окремими файлами. Корпус налічує майже 296 344 токенів і є репрезентативним, оскільки він охоплює всю творчість Р.М. Рільке.

Більшість дослідників погоджуються, що пошук ключових слів – вдалий вихідний пункт для дослідження спеціалізованого корпусу. Ключове слово – це слово (тип), що зустрічається у корпусі набагато частіше за інші. Однак аналіз спеціалізованого корпусу вимагає порівняння його з нормативним (таким, що представляє мовну норму), а це передбачає нормалізацію статистичних даних про частотність. Програми, що використовуються для аналізу корпусів, такі як AntConc [1], найчастіше звертаються до функції логарифмічної правдоподібності (а також до критерію χ^2 -квадрату і точного тесту Фішера [14]) для порівняння частотностей. За допомогою обраного статистичного тесту вираховується нормалізована очікувана кількість шуканих слів у аналізованому корпусі, яка порівнюється з його реальним значенням. Такий підхід дозволяє уникнути непоказових результатів, тобто таких, що є типовими як для аналізованого тексту, так і для нормативного.

За умови обмеженого доступу до референтного корпусу (тобто такого, що не дає можливості отримати файл його вмісту для підключення до програми, пропонуючи лише текстовий пошук) можна скористатися формулою функції логарифмічної правдоподібності (Log-Likelihood, LL), виведеної у статті П. Рейсона і Р. Гарсайда [11]. Змінними, необхідним для побудови цієї функції, є чотири наступні: кількість токенів у досліджуваному й референтному корпусі й кількість шуканих лексем у кожному корпусі.

Прикладом референтного корпусу, придатного для порівняння з корпусом Р. М. Рільке, є корпус DWDS, обмежений хронологічними рамками 1900–1926 років (який містить 27 244 тис. токенів). За допомогою суми результатів пошуку за шаблонами $\$l=\text{Лексема}^*$ и $\$l=*\text{лексема}^*$ можна отримати кількість усіх словоформ і композитів шуканої лексеми. Підставивши результат пошуку слова *Бог* до формули функції логарифмічної правдоподібності, з'ясуємо, що слово *Бог* є дійсно значимим і ключовим для корпусу Рільке, оскільки результат $LL=534,43$ у більш ніж 100 разів перевищує один ступінь вільності, який є достатнім для засвідчення позитивного результату (докладніше див. [15]).

Ми вважаємо, що таку саму формулу можна використовувати для пошуку релевантних коокурентів ключового слова у корпусі, які, у свою чергу, можуть свідчити про асоційованість слів. Референтним і нормативним корпусом у такому випадку стане загальний корпус текстів Рільке, а спеціальними – збережені конкорданси із 20 (Корпус20), 50 (Корпус50), 80 (Корпус80), 100 (Корпус100) і 150 (Корпус150) знаків справа і зліва відносно досліджуваного слова. Нами з'ясовано, що при перевірці частотних слів корпусу за таким методом можна простежити дві наступні тенденції:

– позитивна, за якої слова можуть бути нерелевантними у Корпус20, однак демонструвати зростаючий показник LL зі збільшенням вікна конкордансу, найчастіше досягаючи апогею в Корпус80 та Корпус100. Таку тенденцію ми вважаємо випадком консоційованості й підтвердженням асоційованості слів.

– негативна, за якої слова можуть демонструвати релевантність у Корпус20 або Корпус50 (насамперед, через невисоку нормалізовану очікувану кількість знайдених слів і незначні розміри Корпус20 і Корпус50), однак вона знижуватиметься зі збільшен-

ням вікна конкордансу. Такий графік ми вважаємо підтвердженням самостійності досліджуваного слова відносно ключового слова в його конкордансі.

Окремим випадком першої тенденції можна вважати слова, які демонструють різке високі результати показника логарифмічної правдоподібності (до кількох сотень). Нами було з'ясовано, що така поведінка є характерною для колокованих слів, які зустрічаються не на відстані безпосереднього сусідства. Таку поведінку ми вважаємо доказом асоційованості колокатів та показником загальнономовної асоціації (на противагу індивідуальній асоціації, котра демонструє менш високі результати).

Таким чином, функція логарифмічної правдоподібності може бути використана для пошуку ключових слів у корпусі з референтним корпусом у напіввідкритому доступі, а також для відстеження асоціатів ключового слова. В результаті дослідження часткових корпусів, представлених конкордансами ключового слова Бог у корпусі Р. М. Рільке, регулярними коокурентами цього слова є: історія (Geschichte), кінець (Ende), великий (groß), рука (Hand), природа (Natur), дім (Haus), корінь (Wurzel), земля (Erde), серце (Herz), віра (Glaube). До частотних у корпусі, однак не консоційованих зі словом Бог відносимо: дитина (Kind), очі (Augen), жінка (Frau), смерть (Tod), ніч (Nacht), страх (Angst). Отже, маємо підтвердження домінування сакральних художніх концептів у концептосфері текстів Р.М. Рільке та їх вплив на не-сакральні концепти, що є представленими не-сакральною лексикою (РУКА, СЕРЦЕ), у тому числі просторовою (ЗЕМЛЯ, ГОРА, БУДИНОК).

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. *Anthony L.* AntConc : A learner and classroom friendly, multi-platform corpus analysis toolkit / Laurence Anthony // proceedings of IWLeL. – 2004. – С. 7–13.
2. *Bartlett J. R., Rilke R. M.* Word index to Rainer Maria Rilke's German lyric poetry / James Ronald Bartlett and Rainer Maria Rilke. – Ann Arbor, Mich. : Univ. Microfilms Internat. – 1969.
3. *Brown R. E.* Index zu Rainer Maria Rilke: "Die Aufzeichnungen des Malte Laurids Brigge" / ed. by Russell E. Brown. – Frankfurt am Main : Athenäum-Verl., 1971. – 302 с.
4. *Brusk J.* Using Screenplays as Corpus for Modeling Gossip in Game Dialogues / Jenny Brusk // DiaHolmia. – 2009. – С. 139–144.

5. *Carrillo Masso I.* Developing a methodology for corpus-based computer game studies / Isamar Carrillo Masso // *Journal of Gaming & Virtual Worlds*. – 2009. – Т. 1. – № 2. – С. 143–169.
6. *Goldsmith U. K.* Rainer Maria Rilke: A verse concordance to his complete lyrical poetry. – Leeds : Maney, 1980. – XIX, 1593 S.
7. *Johnson W. L. et al.* Limited domain synthesis of expressive military speech for animated characters / W. Lewis Johnson [et al.] // *Speech Synthesis*, 2002. Proceedings of 2002 IEEE Workshop on. – IEEE, 2002. – С. 163–166.
8. *The American heritage dictionary of the English language* / ed. by William Morris. – Boston : American Heritage, 1969. – 1550 c.
9. *Purushotma R.* Commentary: You're not studying, you're just. / Ravi Purushotma // *Language Learning & Technology*. – 2005. – Т. 9. – № 1. – С. 80–96.
10. *Rankin Y. A., Gold R., Gooch B.* Playing for keeps: gaming as a language learning tool / Yolanda A. Rankin, Rachel Gold, and Bruce Gooch // *ACM SIGGRAPH 2006 Educators program*. – ACM, 2006. – С. 44–50.
11. *Rayson P., Berridge D., Francis B.* Extending the Cochran rule for the comparison of word frequencies between corpora / Paul Rayson, Damon Berridge, and Brian Francis // *7th International Conference on Statistical analysis of textual data (JADT 2004)*. – 2004. – С. 926–936.
12. *Thorne S. L., Fischer I., Lu X.* The semiotic ecology and linguistic complexity of an online game world / Steven L. Thorne, Ingrid Fischer, and Xiaofei Lu // *ReCALL*. – 2012. – Т. 24. – № 3. – С. 279–301.
13. *Walker M. A., Lin G. I., Sawyer J.* An Annotated Corpus of Film Dialogue for Learning and Characterizing Character Style / Marilyn A. Walker, Grace I. Lin, and Jennifer Sawyer // *LREC*. – 2012. – С. 1373–1378.
14. *Weeber M., Baayen R. H., Vos R.* Extracting the lowest-frequency words: Pitfalls and possibilities / Marc Weeber, R. Harald Baayen, and Rein Vos // *Computational Linguistics*. – 2000. – Т. 26. – № 3. – С. 301–317.
15. *Билык Е. А.* Поиск ключевых слов в корпусе Р. М. Рильке в условиях ограниченного доступа к референтному корпусу / Е. А. Билык // *Международный научно-исследовательский журнал: Сборник по результатам International Research Journal Conference XXXVII*. – Екатеринбург: Полиграфисті, 2015. – № 2 (33) Часть 2. – С. 47–49.

Стаття надійшла до редакції 29.04.15

Е. А. Билык, асп.
Институт филологии КНУ имени Тараса Шевченко, г. Киев

Особенности использования функции логарифмического правдоподобия для сравнения специализированных корпусов

Статья посвящена решению проблем корпусной лингвистики. Рассмотрены способы применения функции логарифмического правдоподобия для срав-

нения корпусов, в частности корпуса поэтических текстов Р. М. Рильке с референтным корпусом, находящимся в ограниченном доступе. Доказывается целесообразность применения её для поиска ассоциированных с ключевым словом и разрабатывается соответствующая методология.

Ключевые слова: корпусная лингвистика, Р. М. Рильке, референтный корпус, нормативный корпус, ассоциация.

К. Bilyk, PhD student

Institute of Philology, Taras Shevchenko National University of Kyiv

Application of log-likelihood function for comparison of two specialized corpora

The article discusses the challenges of corpus linguistics. It describes the usage of Log-Likelihood function as a means of comparison of two corpora, one allowing only limited access to its data. The article also suggests a method of searching for the associates of the keyword in the corpus of poetic texts by R. M. Rilke.

Keywords: corpus linguistics, R. M. Rilke, reference corpus, normative corpus, association.

УДК 81'42: 811.133.1

В. Б. Бурбело, д-р філол. наук, проф.

Інститут філології КНУ імені Тараса Шевченка, м. Київ

ПРО МОДЕЛЮВАЛЬНІ ЗАСАДИ ДИСКУРСУ

Статтю присвячено виокремленню основних типів дискурсивних моделей за напрямком конструювання відношень дискурсивного актора з позначуваним світом. Основну увагу приділено аналізу доцентрового екзистенційного франкомовного дискурсу.

Ключові слова: дискурс, модель, конструювання, телеологія, семіозис.

Сучасні мовознавчі науки приділяють значну увагу дослідженню мовленнєвої діяльності у різноманітних її виявах, зокрема дискурсивному. За допомогою мови і, ширше, різних семіотичних систем людина орієнтується у світі та адаптується до нього. Освоєння світу в людських спільнотах відбувається через його семіотизацію та надання йому певного смислу у різних видах спільній діяльності, взаємодії, спрямованій на досягнення визначеної