15. Mungan G. Die semantischen Interaktionen zwischen dem präfigierenden Verbzusatz und dem Simplex bei deutschen Partikeln und Präfixverben. Frankfurt am Main: Verlag Peter Lang, 1986. 257 S.
16. Oguj O. D. Lexikologie der deutschen Sprache. Winnyts'a: Nowa knyha, 2003. 403 S.
17. Paul H. Deutsche Grammatik. Halle (Saale): VEB Max Niemeyer Verlag, 1959. 2. Aufl. 142 S.
18. Wilmanns W. Deutsche Grammatik. Gotisch, Alt-, Mittel- und Neuhochdeutsch. Straßburg: Verlag von Karl J. Trübner, 1899. S. 115-175.
19. Zifonun G. Zur Theorie der Wortbildung am Beispiel deutscher Präfixverben. München: Max Hueber Verlag, 1973. 192 S.

УДК 811.111'373.4
DOI https://doi.org/10.32782/tps2663-4880/2022.21.1.27

РОЛЬ ЧАСТОТНОСТІ У ЛІНГВІСТИЧНИХ ДОСЛІДЖЕННЯХ

THE ROLE OF FREQUENCY IN LINGUISTIC RESEARCH

Ділай І.П.,
orcid.org/0000-0001-9626-290X
кандидат філологічних наук, доцент,
доцент кафедри англійської філології
Львівського національного університету імені Івана Франка

Ділай М.П.,
orcid.org/0000-0001-5182-9220
кандидат філологічних наук, доцент,
доцент кафедри прикладної лінгвістики
Національного університету «Львівська політехніка»

У статті висвітлено переосмислення ролі частотності у сучасних лінгвістичних дослідженнях. Повторюваність мовних одиниць та їх сполук – вагомий чинник формування мовних категорій та стійких асоціацій, який визначає розвиток усіх рівнів мовної системи. Переосмислення ролі частотності пов'язане із стрімким розвитком корпусної лінгвістики та сучасних систем обробки природної мови. Аналіз значних масивів мовної (текстової) інформації вимагає застосування надійних статистичних величин.

Перші згадки про частотність у лінгвістиці стосуються списків частотності слів англійської мови, укладених з навчальною метою, та словників частотності. Особливої ваги набуває вивчення контекстуалізованої частотності не лише для тексту та дискурсу, але й для досліджень системних мовних зв'язків окремих одиниць, їх місця у ментальному лексиконі. Контекстуалізована частотність пов'язана з встановленням сполучуваності слів. У статті розглянуто асоціативні колокаційні міри, способи побудови колокаційних графів та мереж на їх основі.

Окрім поняття частотності та її видів у статті розглянуто суміжне поняття салієнтності, що пов'язано із закарбуванням у мовній свідомості певної нечастотної інформації, або, радше, відносно частотної інформації. Прямим експонентом салієнтності є ключові слова у тестах різного типу, які не завжди найбільш частотні. Здійснено огляд різних підходів до салієнтності та встановлення ключових слів за допомогою статистичних критеріїв.

Таким чином, у статті здійснено огляд наукової лінгвістичної літератури та розглянуто окремі сучасні підходи до частотності, виявлено переваги контекстуалізованої частотності, з'ясовано, як співвідносять частотність та салієнтність на основі аналізу ключових слів, наголошено на актуальності квантитативних методик для усіх напрямів сучасних лінгвістичних досліджень.

Ключові слова: частотність, салієнтність, корпус, контекстуалізована частотність, ключові слова, колокаційні міри.

This paper reconsiders the role of frequency in modern linguistic research. The recurrence of language units and their combinations is an important factor in the formation of language categories and stable associations, which determines the development of all levels of the language system. Taking another look at the role of frequency is associated with the rapid development of corpus linguistics and modern natural language processing systems. The analysis of large amounts of linguistic (textual) data requires the use of reliable statistical measures.

The early studies on frequency in linguistics address the frequency lists of English words compiled for educational purposes and frequency dictionaries. Of particular importance is the study of contextualized frequency not only for text and discourse, but also for the study of systemic linguistic relations of individual units, their place in the mental lexicon. Contextualized frequency is associated with the establishment of word collocability. The paper considers associative collocation measures, methods of constructing collocation graphs and networks based on them.

In addition to the concept of frequency and its types, this paper uncovers the related concept of salience, which is associated with the entrenchment in the language consciousness of certain non-frequent data, or, rather, relatively frequent

data. The direct exponent of salience is keywords in tests of various types, which are not always the most frequent words. An overview of different approaches to salience and keyword extraction using statistical criteria is presented.

Thus, the article reviews the linguistic literature and considers some modern approaches to frequency, specifies the benefits of contextualized frequency, clarifies how to relate frequency and salience based on keyness analysis, emphasizes the relevance of quantitative methods for all areas of modern linguistic research.

Key words: frequency, salience, corpus, contextualized frequency, keywords, collocation measures.

Постановка проблеми. Характерною особливістю мови є повторюваність її одиниць та їх сполук. Саме повторюваність веде до формування категорій та стійких асоціацій, а також до автоматичного повторення сполук [1, с. 50]. Теорії, які базовані на вживанні *(usage-based theories*), постулюють, що мовні структури формуються внаслідок повторення моделей мовного вживання. Таким чином, вивчення частотності вживання сприяє вивченню природи мовної організації. Частотність визначає розвиток лексичної, морфологічної та синтаксичної систем, хоч і не є єдиним чинником їх розвитку.

Аналіз останніх досліджень і публікацій. Роль частотності у розвитку мовної системи неодноразово привертала увагу сучасних дослідників [2; 3; 4; 5; 6]. З появою корпусної лінгвістики стало можливим досліджувати величезні масиви мовних даних, що призвело до того, що дослідження частотності набуло нової ваги та перспектив. Саме частотність вказує на те, як вживання впливає на систему мову [7, с. 3]. Почали укладатися списки та словники частотності, квантитативні методи вдосконалювались та вийшли на новий рівень з метою встановлення закономірностей дистрибуції слів у тексті. Такі напрямки досліджень націлені насамперед на вдосконалення навчальних методик, проте згодом все більше стають у пригоді для вдосконалення систем обробки природної мови.

Окрім поняття частотності у лінгвістиці, набуває поширення і ваги поняття **салієнтності** (salience). У когнітивній науковій традиції *салієнтний* означає "найменш очікуваний", "який не відповідає очікуванням", тому його можна визначати за допомогою стохастичних вимірів [6, с. 135]. Шмід (Schmid) розрізняє когнітивну салієнтність та онтологічну салієнтність [8]. Когнітивна салієнтність стосується стану тимчасової ментальної активації, онтологічна салієнтність позначає внутрішньо притаманну ознаку сутностей в реальному світі. Окремі сутності більше привертають увагу, аніж інші. Когнітивна активація може бути досягнута свідомим вибором. Салієнтний, таким чином, означає "завантажений у робочу пам'ять", "у центрі уваги".

Дещо відмінне трактування салієнтності знаходимо у Джиора (Giora, 2003): *салієнтний* розуміється як такий, який виділяється головним мозком та закодований у ментальному лексиконі людини [9, с. 15]. Існують різні ступені салієнтності, які прямо залежні від частотності вживання, конвенційності, знайомості та прототипності. З точки зору такого градуального підходу до салієнтоності, контекст не є визначальним чинником.

Д. Ґерертс (Geeraerts, 2017) запропонував своє трактування ономасіологічної салієнтності як відносної частотності, з якою *significant* асоціюється з конкретним *signifié* [10]. Таким чином, береться до уваги конкретний прагматичний контекст. Д. Дівяк (Divjak, 2019) вважає, що салієнтність стосується будь-якого (аспекту) стимулу, що робить його помітним для спостерігача. Те, що виділяється має більше шансів бути поміченим та закріпленим *(entrenched)* у пам'яті [6, с. 197].

В останні десятиліття актуальним стало виділення **ключових слів** у тексті за допомогою статистичних методик. Під ключовими дієсловами розуміють слова, які є значно більш чисельні або статистично більш салієнтні у певному корпусі у порівнянні з іншим корпусом [11]. В. Брезіна (Brezina) наголошує, що ключові слова є відносним терміном, який залежить від різниці лексичних частот у двох корпусах [12]. Відповідно, такого типу ключові слова отримали назву **корпусозіставні статистичні ключові слова** (corpus-comparative statistical keywords) [13, с. 566]. Визначення ключових слів є важливим для ідентифікації ключових концептів у дискурсах, типової лексики жанру чи варіанта мови, лексичного розвитку протягом часу тощо. П. Бейкер (Baker, 2011) виокремлює також так звані *слова-замки* (lockwords), які трапляються з однаковою частотою у корпусах, які порівнюємо [14].

Постановка завдання. Враховуючи активну наукову дискусію стосовно переосмислення ролі частотності в сучасних лінгвістичних дослідженнях, їх дедалі більшого і ґрунтовнішого опертя на емпіричні корпусні дані, застосуванні статистичних підрахунків, видається за доцільне систематизувати наявні підходи до заявленої проблематики та акцентувати на її актуальності з метою оптимізації методологічного апарату сучасних лінгвістичних досліджень, які проводяться у вітчизняному мовознавстві. Таким чином, **мета статті** – визначити роль частотності та її трак-

тування у сучасних лінгвістичних дослідженнях. Мета передбачає виконання таких **завдань:** 1) здійснити огляд наукової лінгвістичної літератури та виокремити сучасні підходи до частотності, 2) виявити переваги контекстуалізованої частотності, 3) з'ясувати, як співвідносять частотність та салієнтість на основі аналізу ключових слів.

Виклад основного матеріалу. Перші згадки про частотність у лінгвістиці стосуються т. з. списків частотності слів, які почали укладатися ще в XV столітті. Первісні такі списки – глосарії, в яких слова були згруповані за категоріями, наприклад, список Брайта (Bright, 1588), та для яких частотність не була основним критерієм. Проте у XVII столітті з'явилася праця Коменськи (Komenský, 1631) *Janua linguarum reserata*, яка базована на частотності. Списки частотності подавали кількість окремих випадків вживання (*tokens*) кожного слова (*type*) у вибраних текстах.

Найдавніший словник частотності англійської мови укладено Ейрес (Ayres, 1915) на основі 368000 корпусу слів з комерційного та приватного листування. Мета цього словника була виключно навчальна. Після того, як Елдрідж (Eldridge) (1911) опублікував працю про 6000 найпоширеніших англійських слів для вчителів, американський психолог, лексикограф та методист Торндайк (Thorndike) опублікував три праці зі списками частотності: *The Teacher's Word Book* (1921), яка містила 10000 слів, *A Teacher's Word Book of the Twenty Thousand Words Found Most Frequently and Widely in General Reading for Children and Young People* (1932) та *The Teacher's Word Book of 30,000 Words* (1944).

Перший корпусо-базований список частотності англійський слів *General Service List* належав М. Весту (West, 1953) та включав 2000 слів. З розвитком комп'ютерних технологій та появою електронних корпусів, зокрема першого Браунівського корпусу (Brown Corpus, 1961), з'явився список Кучера та Френсіса (Kučera, Francis, 1967). Після цього було укладено чимало списків, проте усі вони базувалися на різних критеріях: на різних одиницях підрахунку, вибірці, типі (писемний, усний) та розмірі корпусу.

Проте списки частотності ізольованих слів не видаються ефективними ані в навчальному процесі, ані в лексикографії, адже вони не беруть до уваги контекст. Слово може бути високочастотним в обмежених контекстах. Абсолютна частотність (*raw frequency*) не є вирішальною в сучасних корпусних дослідженнях. Корпусна лінгвістика завдяки програмі конкордації дозволяє виділяти та досліджувати колокації та колігації та основі статистичних критеріїв. Окрім статистичних мір сполучуваності, в межах корпусної лінгвістики розроблено формули контекстуальної різноманітності, **дисперсії,** яка вимірює однорідність дистрибуції слова у корпусі. Таким чином, контекстуалізовані частотності видаються кращими мірами, ніж ізольовані частотності [6, с. 30].

Вивчення значень слів у контексті є важливим для лінгвістичних та соціологічних досліджень. В. Брезіна (зазначає, що колокації, ключові слова та кодування вручну ліній конкордансу відіграють ключову роль у дослідженнях семантики та аналізу дискурсу [12, с. 66]. Значення найкраще вивчати за допомогою аналізу повторювальних лінгвістичних патернів у корпусі.

Колокації можуть базуватися або лише на частотності, як у більшості випадків, або на статистичній мірі, яку називають асоціативною мірою. Асоціативні міри (або колокаційні міри) – це статистичні міри, які обчислюють силу асоціацій між словами за різними аспектами зв'язків під час спільної появи (co-occurrence relationship). Різні асоціативні міри, які існують в корпусній лінгвістиці, дають дещо відмінні списки колокацій. Вибір асоціативної міри залежить від мети та завдань дослідження, від того, які аспекти колокаційних відношень хочемо виділити – частотність чи винятковість. В. Брезіна зазначає, що для дослідження виняткових колокатів варто застосовувати Dice, Log Dice, MI2; для встановлення виняткових нечастотних колокатів варто використовувати MU, MI; а для виняткових частотних – MI3. Для визначення частотних невиняткових колокатів використовують T-score, LL, MS, Frequency. Існують асоціативні міри, які враховують спрямованість (directionality) – Delta P, або дисперсію – Cohen's d. Розраховуються асоціативні міри за допомогою рівнянь, які містять значення з корпусу, а саме: кількість токенів у цілому корпусі, частотність вузла (node), тобто центру колокації у корпусі, частотність колоката у корпусі, частотність колокації (вузол + колокат) у колокаційному вікні (collocation window), тобто у певних межах зліва та справа від вузла, межі колокаційного вікна. Окрім спостережуваної частотності (observed frequencies), беруть до уваги очікувані частотності у корпусі (expected frequencies). Формули розрахунку цих мір подані зокрема у праці В. Брезіна [12, с. 72].

Сучасні корпусні інструменти автоматично розраховують колокаційні міри, а також дозволяють побудувати колокаційні графи та мережі. Колокаційний граф – це візуальна презентація

колокаційних відношень між вузлом та його колокатами. На рис. 1 зображено граф для вузла *girl* у корпусі творів Дж. Остін (777966 токенів, 20820 типів, 18888 лем) [15], застосовуючи log Dice. Використано корпусний інструмент #LancsBox: Lancaster University corpus toolbox [16].

Граф показує три виміри колокаційних відношень: силу асоціацій, частотність колокатів і позицію колоката у тексті. Сила асоціацій, яка обчислена тут за допомогою асоціативної міри log Dice, виражається як довжина зв'язків між вузлом та колокатом: що ближче колокат до вузла, то сильніші зв'язки. Відповідно, найсильніші асоціації вузол *girl* має у цьому корпусі зі словами *pretty, poor, sweet, fifteen, amiable, age, marry,*. Частотність показана відтінком кольору колоката: що інтенсивніший колір, то частотніший колокат. Відповідно, до найчастотніших належать *a, of, and, is, poor, pretty, who.* Позиція колоката в тексті (перед чи після вузла) відображена як позиція колоката на графі (зліва L: *pretty, poor, sweet, amiable, marry*; справа R: *fifteen, age, whom, who, world*).

Колокаційні мережі – це розширені колокаційні графи, які показують більші асоціативні патерни; це мережі пов'язаних колокацій, які починається з певного вузла, навколо якого визначаються колокати. Кожен з колокатів можна далі розглядати як новий вузол, для якого теж встановлюються колокати. На рис. 2 показано мережу, яка містить колокати вихідного вузла *girl,* колокати нового вузла *marry* (що є колокатом *girl*) та спільні для обох вузлів колокати (24: *a*, *all*, *and*, *as*, *be*, *could*, *do*, *her*, *i*, *is*, *know*, *me*, *miss*, *must*, *my*, *never*, *not*, *or*, *she*, *should*, *to*, *when*, *who* and *you*).

Таким чином можна будувати розширену мережу, яка показує, як окремі слова пов'язані прямими та перехресними асоціаціями. Навіть віддалені колокати можуть брати участь у формуванні значення певного слова, будучи у спільному концептуальному просторі з цим словом через асоціативні зв'язки з іншими словами, які прямо сполучаються з цим словом [12, с. 77]. Отже, колокаційні графи та мережі підсумовують складні значення слів у текстах та корпусах, надають корисну інформацію про структуру тематики текстів.

До важливих методик корпусної лінгвістики належить встановлення ключових слів. Методика ключових слів ґрунтується на виявлення слів, які є характерними для певних текстів чи корпусів, і які далі можна досліджувати використовуючи методи колокацій (аналізуючи випадки повторювального вживання слів разом) та конкордансу (аналізуючи приклади вживання слів у контексті). Процедура визначення ключових слів передбачає порівняння досліджуваного корпусу (або 'focus corpus', 'node corpus') з референтним корпусом, використовуючи статистичні міри, за допомогою яких визначають слова, що вживаються

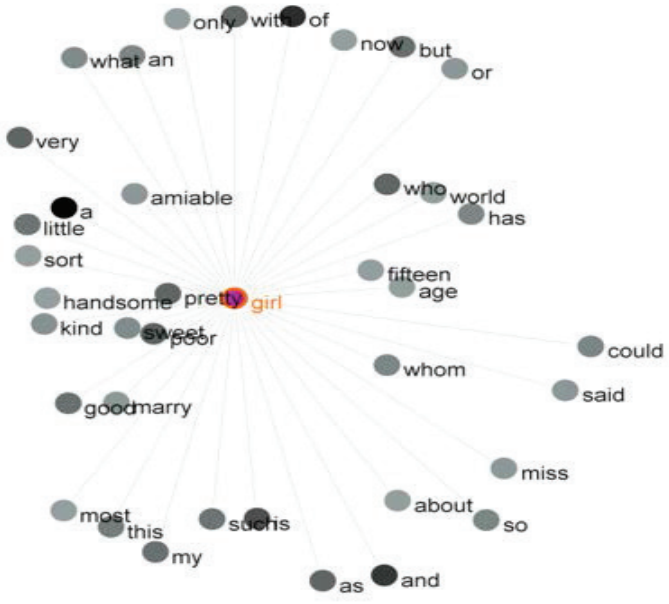

Рисунок 1. Колокати вузла *girl* у корпусі Austen (log Dice, L3–R3)

Рисунок 2. Колокаційна мережа вузла *girl* та його колоката *marry*

або частіше, або рідше. Якщо слово вживається частіше у досліджуваному корпусі, ніж у референтному, його називають позитивним ключовим словом (+), а якщо рідше – негативним (–). На практиці, ключові слова визначаються автоматично за допомогою корпусних програм, які генерують два списки слів, один – базуючись на досліджуваному корпусі, інший – відповідно до референтного корпусу, та порівнюють їх. Прийнято вважати, що більший за обсягом та подібний до досліджуваного референтний корпус забезпечує надійніше і більш фокусоване порівняння.

Традиційно для встановлення ключовості (keyness) використовується статистичний критерій log-likelihood (LL). У його підрахунку беруть до уваги спостережувані та очікувані частотності слова у досліджуваному та референтному корпусах. Кілґаррифф (Kilgarriff, 2009) пропонує для встановлення ключових слів обчислювати відношення між відносними частотами слів у корпусах [17]. Цю міру називають простим математичним параметром (SMP). Застосовують також такі статистичні міри як %DIFF, log ratio, Cohen's d. Формули обчислення детально описано у В. Брезіна [12, с. 84].

Варто також згадати, що розглядають два типи аналізу ключовості: експланаторний та фокусований [18]. Експланаторний аналіз ключовості ґрунтується на аналізі конкордансу, який показує контекст лексичних одиниць. Фокусований аналіз ключовості передбачає порівняння нормалізованої частотності певних одиниць у двох корпусах задля вирішення певної дослідницької мети.

Процедура встановлення ключових слів може також застосовуватися до категорій вищого рівня абстракції, зокрема лем та ключових концептів [19]. Таким чином можна встановити концепти та семантичні сфери, типові для певного типу дискурсу. Дослідники наголошують, що список ключових слів – це результат певних рішень, починаючи з вибору референтного корпусу та закінчуючи вибором статистичної міри.

Висновки. На основі здійсненого огляду наукової лінгвістичної літератури виокремлено сучасні підходи до частотності. Встановлено перевагу контекстуалізованої частотності та застосування складних статистичних критеріїв для дослідження системномовних явищ, зокрема колокацій. З'ясовано процедуру встановлення ключових слів у текстах на основі виявлення салієнтих одиниць.

Сучасне застосування частотності в лінгвістичних дослідженнях виходить за межі об'єктивації та верифікації їх результатів та стає невід'ємним інструментом наукового пошуку, вивчення мовної структури, створення мовних моделей та програм обробки природної мови. Таке переосмислення ролі частотності в сучасних лінгвістичних дослідженнях стало можливим з розвитком корпусної лінгвістики та сучасних комп'ютерних технологій, здатних обробляти значні масиви інформації.

Перспективи подальших розвідок у цьому напрямку вбачаємо у розробці та застосуванні найбільш оптимальних статистичних критеріїв для конкретних лінгвістичних завдань із застосуванням новітніх надбань корпусної лінгвістики.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ:

1. Bybee J. L. Usage-Based Theory and Exemplar Representation. *The Oxford Handbook of Construction Grammar /* eds. T. Hoffman, G. Trousdale. Oxford University Press : Oxford, 2013.

2. Frequency Effects in Language Acquisition. Defining the Limits of Frequency as an Explanatory Concept / eds. I. Gülzow, N. Gagarina. De Gruyter Mouton : Berlin, Boston, 2007.

3. Ambridge B., Lieven E. V. M. Child Language Acquisition: Contrasting Theoretical Approaches. Cambridge University Press : Cambridge, 2011.

4. Ambridge B., Kidd E., Rowland C. F., Theakston A. L. The ubiquity offrequency effects in first language acquisition. *Journal of Child Language*. 2015. 42(2). P. 239-73.

5. Experience Counts: Frequency Effects in Language / eds. H. Behrens, S. Pfänder. De Gruyter : Berlin, Boston, 2016.

6. Divjak D. Frequency in Language: Memory, Attention and Learning. Cambridge : Cambridge University Press, 2019.

7. Bybee J. L., Hopper P. Frequency and the Emergence of Linguistic Structure. John Benjamins : Amsterdam, Philadelphia, 2001.

8. Schmid H. The Notion of Entrenchment and Salience in Cognitive Linguistics. *The Oxford Handbook of Cognitive Linguistics /* eds. D. Geeraerts, H. Cuyckens. Oxford : Oxford University Press, 2007. P. 117–138.

9. Giora R. On our mind: Salience, context, and figurative language. Oxford University Press, 2003.

10. Geeraerts D. Entrenchment as Onomasiological Salience. *Entrenchment and the Psychology of Language Learning: How We Reorganize and Adapt Linguistic Knowledge* / ed. H.-J. Schmid. Berlin : De Gruyter Mouton and APA, 2017.

11. Scott M. PC analysis of key words – and key key words. *System*. 25(2). 1997. P. 233–45.

12. Brezina V. Statistics in Corpus Linguistics: A Practical Guide. Cambridge : Cambridge University Press, 2018. 316 p.

13. O'Halloran K. How to use corpus linguistics in the study of media discourse. *The Routledge Handbook of Corpus Linguistics* / eds. A. O'Keeffe, M. McCarthy. London and New York : Routledge, 2010. P. 563–577.

14. Baker P. Times may change, but we will always have money: diachronic variation in recent British English. *Journal of English Linguistics*.2011.39(1).P.65-88.

15. Austen's Corpus. *#LancsBox: Lancaster University corpus toolbox*. URL: http://corpora.lancs.ac.uk/lancsbox/ (дата звернення: 25.01.2022).

16. #LancsBox: Lancaster University corpus toolbox. URL: http://corpora.lancs.ac.uk/lancsbox/ (дата звернення: 25.01.2022).

17. Kilgarriff A. Simple maths for keywords. *Proceedings of the Corpus Linguistics Conference*. Liverpool, 2009.

18. Gabrielatos C. Keyness Analysis: nature, metrics and techniques. *Corpus Approaches to Discourse: A critical review* / eds. C. Taylor, A. Marchi. Oxford : Routledge. 2018. P. 225–258.

19. Rayson P. From key words to key semantic domains. *International Journal of Corpus Linguistics.* 200. 13(4). P. 519–49.