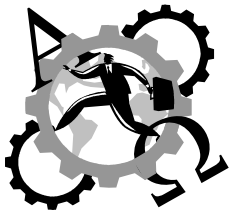
ISSN 2227-6890



ІНТЕГРАЦІЯ ФІЛОЛОГІЇ І ТЕХНІЧНИХ НАУК

УДК 004.912

С. В. ПЕТРАСОВА, М. О. КУЗЬМІНА

АВТОМАТИЧНЕ ВИДОБУВАННЯ КОЛОКАЦІЙ З КОРПУСУ ТЕКСТІВ

У статті розглядається метод автоматичного видобування колокацій з корпусів текстів української мови. Визначено поняття «колокація» з точки зору підходів до його аналізу у сучасній корпусній лінгвістиці. Проаналізовано статистичні методи та існуючі системи, що використовують статистичні міри для видобування колокацій. Описано структуру розробленого корпусу текстів, а також імплементацію статистичної міри MI для виявлення колокацій з україномовних текстів, що складаються з інструкцій технічної документації.

Ключові слова: колокація, корпус текстів, корпусна лінгвістика, статистичні методи, міра MI, технічна документація.

С. В. ПЕТРАСОВА, М. А. КУЗЬМИНА

АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ КОЛЛОКАЦИЙ ИЗ КОРПУСА ТЕКСТОВ

В статье рассматривается метод автоматического извлечения коллокаций из корпусов текстов украинского языка. Определено понятие «колокация» с точки зрения подходов к его анализу в современной корпусной лингвистике. Проанализированы статистические методы и существующие системы, использующие статистические меры для идентификации коллокаций. Описана структура разработанного корпуса текстов, а также имплементация статистической меры MI для выявления коллокаций в украиноязычных текстах, состоящих из инструкций технической документации.

Ключевые слова: коллокация, корпус текстов, корпусная лингвистика, статистические методы, мера MI, техническая документация.

S. V. PETRASOVA, M. O. KUZMINA

AUTOMATIC EXTRACTION OF COLLOCATIONS FROM A TEXT CORPUS

The article deals with the method for automatic extraction of collocations from the text corpus of the Ukrainian language. Definitions of the term "collocation" have been considered from the viewpoint of approaches to its analysis in modern corpus linguistics. Existing statistical methods and systems that use statistical measures for extraction of collocations have been analyzed, and their features have been described. The corpus of technical documentation has been developed and its structure has been described. To extract collocations from the texts of the Ukrainian language that consist of the instructions of technical documentation, the developed software implementation of MI measure has been described.

Keywords: collocation, text corpus, corpus linguistics, statistical methods, MI measure, technical documentation.

Вступ. Розвиток технологій обробки великих масивів слабкоструктурованої інформації, що налічують мільйони і навіть мільярди слів, зумовив стрімкий розвиток корпусної лінгвістики, важливим завданням якої є виявлення лінгвістично релевантної інформації, зокрема за рахунок використання статистичних методів. Корпусні дослідження дозволяють верифікувати лінгвістичні теорії та гіпотези, а також виявляти та інтерпретувати нові мовні факти [1].

Одним з напрямків дослідження корпусної лінгвістики є вирішення проблеми сполучуваності слів або видобування колокацій з масиву текстових даних.

Інтерес до вивчення колокацій пояснюється високою частотністю таких словосполучень в текстах різних функціональних стилів. Результати дослідження колокацій знаходять застосування при розробці нових пошукових систем, побудові систем машинного перекладу, розпізнавання та генерації текстів.

В межах напрямку контекстуалізму значення колокації розглядається як складне лінгвістичне явище, що потребує дослідження на всіх рівнях мовної структури. В аналізі значення найважливішу роль відіграє контекстуалізація, тобто прийом встановлення контексту стосовно кожного мовного рівня. На лексичному рівні колокації – це типове і постійне оточення слова, вказівка на його традиційну зустрічальність. Таким чином, під колокаціями розуміють характерні словосполучення, які часто зустрічаються та «поява яких поруч один з одним ґрунтується на регулярному характері взаємного очікування і задається не граматичними, а суто семантичними чинниками».

В межах семантико-синтаксичного підходу колокації розглядаються як семантико-синтаксичні одиниці або лексично визначені елементи граматичних структур. Вони характеризуються семантичною, синтаксичною та дистрибутивною регулярністю, внутрішньо притаманними властивостями словосполучень, а не їх появою у корпусах [2].

Характеристики високої частоти спільної зустрічальності недостатньо, щоб говорити про стійкість комбінацій слів. Тому вироблено цілий ряд статистичних мір (мір асоціації, або мір асоціативної зв'язаності), що обчислюють силу зв'язку між елементами в складі колокації. У загальному випадку, ці міри враховують як частоту спільної зустрічальності, так і інші параметри, насамперед частоту в даному корпусі кожного окремого елемента [3, 4].

В роботі пропонується розглядати колокації з точки зору статистичного підходу, спираючись на значення статистичних мір.

Аналіз останніх досліджень і публікацій. У зв'язку з постійним зростанням обсягів текстової інформації все більшого значення набувають спеціальні комп'ютерні програми (корпусні менеджери), що використовують статистичні міри для вилучення з корпусів парних лексичних відношень. Найбільш відомими універсальними корпусними менеджерами є SARA, XAIRA (BNC), CQP, які призначені для пошуку даних в корпусі, отримання статистичної інформації та надання результатів в зручній для користувача формі.

Прикладом статистичного підходу виявлення колокацій, в якому вперше реалізована методика автоматичного видобування всього діапазону колокацій, є система Xtract. Її новизна полягає у порушенні «канонічного порядку аналізу» сталих словосполучень: вихідними даними для ідентифікації колокацій стають не лінгвістичні ознаки, а статистичні характеристики. Програма видобування колокацій включає два компоненти: конкорданс опрацювання корпусу біржових звітів та їх статистичний аналіз. Для подальшого опрацювання лінгвістичними фільтрами зберігаються тільки статистично значущі пари слів.

Даний підхід до ідентифікації колокацій характеризується поетапним застосуванням статистичного аналізу, трьох лінгвістичних фільтрів (позиційного, синтаксичного, морфологічного) та оцінюванням точності експертом-лексикографом [2].

Прикладом корпусних менеджерів, здатних робити підрахунки частот слів або словоформ і частот спільної зустрічальності в україномовних текстах, є статистично-пошуковий апарат Корпусу української мови [5] та Українського Національного Лінгвістичного Корпусу [6].

В системах автоматичного визначення колокацій використовуються такі міри як міра асоціації MI, PMI, t-score, Log-Likelihood, Dice-міра та інші, які найчастіше застосовуються при обчисленні ступеня близькості між компонентами словосполучень в корпусі [7].

Міра MI [8] відноситься до точкових оцінок сили асоціації. В основі MI лежить поняття взаємної інформації (mutual information), запозичене з теорії інформації. Коефіцієнт взаємної інформації (1) порівнює залежні контекстно-зв'язані частоти з незалежними (при випадковій появі слів в контексті).

$$MI(n,c) = \log_2 \frac{f(n,c) \times N}{f(n) \times f(c)}, \qquad (1)$$

де n – ключове слово; c – колокат; $f(n, c)$ – частота зустрічальності ключового слова n в парі з колокатом c; $f(n)$, $f(c)$ – абсолютні (незалежні) частоти ключового слова n і колоката c в корпусі (тексті); N – загальне число словоформ в корпусі (тексті).

Міра дозволяє виділяти найбільш рідкісні і своєрідні колокації і підходить для виділення термінології, власних імен та інших конструкцій, в яких частота слів колокації мізерно мала.

Якщо значення MI більше 1, тоді дана комбінація слів вважається статистично значимою. У разі якщо MI приблизно дорівнює 0, комбінація слів вважається менш статистично значимою, слова з'являються в парі вкрай рідко. MI менше 0 означає, що n і c знаходяться у відношенні додаткової дистрибуції.

Значення міри MI залежить від розміру корпусу – чим більше досліджуваний корпус, тим вище в середньому одержувані за ним значення MI. Залежність міри MI від розміру корпусу ускладнює порівняння значень мір, отриманих на різних корпусах, або наприклад, на повній колекції та її частини. Один із способів вирішення цієї проблеми – це використання міри MI як засобу ранжування колокацій всередині одного корпусу за ступенем їх зв'язаності [9].

Міра t-score також враховує частоту спільної зустрічальності ключового слова і його колоката, відповідаючи на питання, наскільки невипадковою є сила асоціації (зв'язаності) між колокатами. Вона обчислюється за формулою (2).

$$t - score = \frac{f(n,c) - \frac{f(n) \times f(c)}{N}}{\sqrt{f(n,c)}}, \qquad (2)$$

де n – ключове слово; c – колокат; $f(n, c)$ – частота зустрічальності ключового слова n в парі з колокатом c; $f(n)$, $f(c)$ – абсолютні (незалежні) частоти ключового слова n і колоката c в корпусі (тексті); N – загальне число словоформ в корпусі (тексті).

Ця формула показує, наскільки розподіли ключового слова і колоката в корпусі (тексті) залежать один від одного. Однак можлива переоцінка деяких випадкових результатів, зокрема, поєднань високочастотного елемента з низькочастотним. Тому t-score зазвичай використовується в комбінації з іншими мірами, найчастіше з MI.

До недоліків використання цієї міри можна віднести те, що вона, в першу чергу, виділяє колокації з дуже частотними словами, зокрема, зі службовими словами. Тому для t-score необхідно задавати stop list, щоб відкинути найбільш частотні слова, поєднання з якими незмінно виявиться в самому верху таблиці: прийменники, займенники або сполучники.

Критерій t-score спрямований перш за все на виділення стійких конструкцій, кліше, і загальномовних стійких сполучень [9].

Широко застосовується міра Log-Likelihood (3) (логарифмічна функція правдоподібності):

$$\log - likelihood = 2\sum Oij \times \log \frac{Oij}{Eij}, \qquad (3)$$

де *Oij*, *Eij* – спостережувані і очікувані частоти.

Коефіцієнт Dice – міра, особливістю якої є те, що вона знаходить симетричні стійкі комбінації, що дозволяє виявляти в підкорпусі слова з обмеженою сполучуваністю, та, відповідно, комбінації, які з високим ступенем ймовірності претендують на статус колокацій [10].

Таким чином, жодна з найпоширеніших мір асоціацій не вирішує завдання виявлення колокацій безпомилково. Тому доцільно використовувати ці міри в комбінації, або створити новий алгоритм виявлення колокацій, який занижував би значення мір асоціації для неосмислених поєднань, або зовсім виключав би їх на проміжній стадії виділення.

Метою дослідження є вирішення завдання видобування колокацій з україномовних текстів із застосуванням статистичних методів. Визначення колокацій як статистично значущих одиниць дозволить автоматизувати опрацювання природномовної інформації, а також отримати дані про механізми утворення словосполучень для подальшого їхнього аналізу та використання корпусу технічної документації.

Матеріали і результати дослідження. Для вирішення поставленої задачі було розроблено корпус технічних інструкцій. Об’єм створеного корпусу гарантує типовість даних і забезпечує повноту уявлення всього спектра мовних явищ. Дані різного типу знаходяться в корпусі в своєму природньому оточенні, що дає можливість їх всебічного і об'єктивного вивчення.

Розроблений корпус містить текстові (txt) файли української мови, який має свою структуру (рис.1).

Корпус текстів

Назва підкорпусу	Назви файлів	Кількість слів	Довж
HTC	Desire_526G	16670	215
	Desire_601	16777	215
	Desire_610	16643	220
	Touch_Diamond	16981	229
Nokia	Nokia_5230	16120	226
	Nokia_6111	16325	230
	Nokia_C7-00	16895	232
	Nokia_N8-00	16463	224
Samsung	Note_8	16290	230
	Note_7	16285	226
	GT_I9100	16158	227
	GT_S7230E	16017	222

Рис. 1 – Метадані корпусу текстів технічної документації

До складу корпусу текстів технічних інструкцій входять три підкорпуси, які відібрані за принципом відомих компаній виробників мобільних пристроїв. Підкорпуси налічують однакову кількість файлів, а саме по 4 файли у кожному підкорпусі. Файли відібрані за існуючими моделями мобільних пристроїв обраних компаній та містять приблизно однакову кількість слів – 16000-17000. Загальний обсяг корпусу становить 197624 слів, що дозволяє віднести його до корпусів великого обсягу та дозволяє отримати достовірні дані про частоту тієї чи іншої комбінації в українській мові та в мові в цілому.

Для створення та аналізу лінгвістичного корпусу було обрано тексти науково-технічного стилю, які можуть бути представлені:

- науково-технічною літературою, тобто монографії, збірники та статті з різних проблем технічних наук;
- навчальною літературою з технічних наук (підручники, керівництва);
- технічною та товаросупровідною документацією (паспорти, технічні описи, інструкції з експлуатації та ремонту;
- проектною документацією: проекти, розрахунки, креслення [11].

Розроблений корпус текстів технічної документації характеризуються стислістю та однозначністю, логічністю і чіткою послідовністю викладання матеріалу, об'єктивністю інформації, великою кількістю спеціалізованих науково-технічних термінів, відсутністю емоційних оцінок та особливим стилем викладання матеріалу.

Інструкції з експлуатації є описами виробу та правил користування ним. Вони містять опис частин виробу, послідовність його складання, рекомендації з налаштування, користування і обслуговування. Особливу увагу в правилах з експлуатації приділяється правилам безпеки. Інструкції містять вступну частину, повний опис виробу та умови його функціонування, сервісне обслуговування і ремонт, можливі несправності та способи їх усунення, правила з транспортування, зберігання та утилізації. Інформація, представлена в інструкціях є стислою, містить велику кількість технічних термінів, які несуть основне смислове, інформаційне навантаження і є однозначними, тобто за ними закріплено тільки одне встановлене значення. В їхній структурі переважають іменники, прикметники, дієслова, слова з основним предметно-логічним значенням, безособові форми дієслова.

Більшість термінів обраної предметної області є не однослівними. Саме не однослівні терміни характеризуються терміном колокація.

На основі розглянутих статистичних методів роботи з колокаціями було обрано міру MI як засіб вирішення завдання автоматичного виявлення колокацій в україномовному корпусі. MI дозволяє виділяти ключові не однослівні терміни, які характеризують предметну область.

Для нормалізації значень міру MI було видозмінено шляхом використання метрики MI3 (зведення значення в куб).

Для визначення колокацій у корпусі української мови на матеріалі технічної документації було розроблено наступний алгоритм:

1. Підрахунок загального числа словоформ в корпусі.
2. Знаходження абсолютних частот усіх слів.
3. Знаходження частоти біграм.
4. Обчислення міри подібності для пар слів.

Встановлено поріг відбору колокацій: + 0,5 від мінімального значення та обмежено вивід результатів для виведення якомога більшого набору словосполучень, які на нашу думку, є найточнішими.

У результаті отримано колокації та їх числові частотні значення (рис. 2). Найкращий результат було отримано серед таких стійких словосполучень як захисна плівка, задня кришка та датчик наближення.

```
        Колокації
('очищено', 'вилучено'):57365.14
('вилучено', 'стерто'):57365.14
('допомого', 'стороннього'):36193.37
('чохол', 'захисну'):72386.75
('захисну', 'плівку'):57365.14
('закривайте', 'блокуйте'):57365.14
('блокуйте', 'датчик'):36193.37
('датчик', 'наближення'):36193.37
('наближення', 'придбайте'):57365.14
('сторонніх', 'гарнітур'):84038.41
('гарнітур', 'аксесуарів'):36193.37
('аксесуарів', 'металевими'):36193.37
('металевими', 'брелоками'):36193.37
('брелоками', 'висять'):36193.37
('вплинути', 'прийом'):93558.51
('задня', 'кришка'):36193.37
('кришка', 'зняття'):72386.75
('зняття', 'задньої'):84038.41
('задньої', 'кришки'):76566.26
('нижньою', 'стороною'):93558.51
('догори', 'задньою'):84038.41
('задньою', 'стороною'):93558.51
('невелику', 'щілину'):36193.37
('щілину', 'починаючи'):84038.41
('починаючи', 'отвору'):84038.41
```

Рис. 2 – Знайдені колокації

Для реалізації програмного забезпечення вирішення задачі визначення колокацій у створеному корпусі української мови була обрана високорівнева мова програмування Python та середовище Spyder. У середовищі встановлено бібліотеку nltk для роботи з природними мовами, зокрема українською.

Розроблена програма виділяє двослівні колокації декількох типів: термінологічні та загальномовного поєднання, імена власні, словосполучення, що характеризують тему тексту, а також деякі вільні сполучення.

Висновки. При застосуванні різних підходів до виділення колокацій, основні проблеми відбору складають: встановлення критеріїв ідентифікації, класифікації колокацій і оцінювання ефективності використовуваних прийомів і процедур. Проблема полягає в тому, що жодна система не отримує весь діапазон колокацій з аналізованого тексту. У зв'язку з цим виникає необхідність удосконалення методики розпізнавання колокацій в природно-мовному тексті на основі об'єктивних критеріїв [12].

В результаті проведеного аналізу методів автоматичного видобування колокацій було запропоновано алгоритм для визначення колокацій в україномовному корпусі текстів технічної документації. Програмна реалізація розробленого алгоритму виявлення двослівних колокацій базувалась на використанні метрики MI3.

Список літератури

1. Жуковська В.В. *Вступ до корпусної лінгвістики*. Житомир: Вид-во ЖДУ ім. І. Франка, 2013. 142 с.
2. Бобкова Т. В. Основні підходи до ідентифікації й вилучення колокацій із текстів. *Наукові праці. Філологія. Мовознавсво*. 2015. №241 (253). С. 10–16. URL: http://linguistics.chdu.edu.ua/article/viewFile/87653/83242/ (дата звернення: 14.03.2018).
3. Хохлова М.В. Экспериментальная проверка методов выделения коллокаций. *Slavica Helsingiensia. Корпусные подходы*. Под ред. А. Мустайоки, М. В. Копотева, Л. А. Бирюлина, Е.Ю. Протасовой. Хельсинки, 2008. С.343–357. URL: http://www.helsinki.fi/slavicahelsingiensia/preview/sh34/pdf/21.pdf (дата звернення: 14.03.2018).
4. Захаров В. П., Хохлова М. В. Выделение терминологических словосочетаний из специальных текстов на основе различных мер ассоциации. *Интернет и современное общество «IMS-2014»: сб. научн. статей XVII всеросс. объединенной конф*. СПб.: Университет ИТМО, 2014. С. 290–293.
5. *Корпус текстів української мови*. URL: http://www.mova.info/corpus.aspx?l1=209 (дата звернення: 14.03.2018).
6. *Український національний лінгвістичний корпус*. URL: http://unlc.icybcluster.org.ua/virt_unlc (дата звернення: 14.03.2018).
7. Петрасова С.В., Хайрова Н.Ф. Логико-лингвистическая модель идентификации семантически эквивалентных коллокаций. *Вісник НТУ «ХПІ»*. 2015. № 58 (1167). С. 14–17.
8. Evert S., Krenn B. Methods for the qualitative evaluation of lexical association measures. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France, 2001. P. 188–195.
9. Ягунова Е.В., Пивоварова Л.М. Извлечение и классификация коллокаций на материале научных текстов. Предварительные наблюдения. *V Межд. научно-практ. конференция "Прикладная лингвистика в науке и образовании"*. СПб, 2010. С. 356–364.
10. Захаров В. П., Богданова С. Ю. *Выделение коллокаций статистическими методами*. URL: https://www.docme.ru/doc/1337883/2082 (дата звернення: 14.03.2018).
11. Ягунова Е. В., Пивоварова Л. М. От коллокаций к конструкциям. *Acta linguistica petropolitana. Тр. Ин-та лингв. исслед. РАН*. 2011. URL: http://www.webground.su/data/lit/pivovarovayagunova/Otkollokatsiykkonstruktsiyam.pdf (дата звернення: 14.03.2018).
12. Бобкова Т.В. Корпус текстів: основні аспекти визначення. *Науковий вісник КНЛУ*. Київ: КНЛУ, 2014. № 29. С. 11–20.

References (transliterated)

1. Zhukovska V. V. *Vstup do korpusnoi lingvistyky* [Introduction to Corpus Linguistics]. Zhytomyr: I. Franko ZhDU Publ., 2013. 142 p.
2. Bobkova T. V. Osnovni pidhody do identyfikacii j vyluchennja kolokacij iz tekstiv [Basic approaches to the identification and extraction of collocations from texts]. *Scientific works. Philology. Linguistics.* 2015, no. 241 (253), pp. 10–16. Available at: http://linguistics.chdu.edu.ua/article/viewFile/87653/83242/ (accessed 14.03.2018).
3. Khokhlova M.V. Eksperimentalnaja proverka metodov vydelenija kollokacij [Experimental verification of methods for collocation extraction]. *Slavica Helsingiensia. Corpus approaches.* Helsinki, 2008. p. 343-357. Available at: http://www.helsinki.fi/slavicahelsingiensia/preview/sh34/pdf/21.pdf/ (accessed 14.03.2018).

4. Zakharov V.P., Khokhlova M.V. Vydelenie terminologicheskih slovosochetanij iz special'nyh tekstov na osnove razlichnyh mer associacii [Identification of terminological phrases from special texts based on various association measures]. *Internet i sovremennoe obshhestvo «IMS-2014»: sb. nauchn. statej XVII vseross. obedinennoj konf*. St. Petersburg: ITMO University, 2014, pp. 290-293.
5. *Korpus tekstiv ukrainskoi movy* [The text corpus of the Ukrainian language]. Available at: http://www.mova.info/corpus.aspx?l1=209/ (accessed 14.03.2018).
6. *Ukrainskyj nacionalnyj lingvistychnyj korpus* [Ukrainian National Linguistic Corpus]. Available at: http://unlc.icybcluster.org.ua/virt_unlc/ (accessed 14.03.2018).
7. Petrasova S.V., Khairova N.F. Logiko-lingvisticheskaja model identifikacii semanticheski jekvivalentnyh kollokacij [A logical and linguistic model for identification of collocation similarity]. *Visnyk NTU "KhPI"* [Bulletin of the National Technical University "KhPI"]. Kharkiv, NTU "KhPI" Publ., 2015, no. 58 (1167), pp. 14–17.
8. Evert S., Krenn B. Methods for the qualitative evaluation of lexical association measures. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France, 2001, pp. 188–195.
9. Jagunova E.V., Pivovarova L.M. Izvlechenie i klassifikacija kollokacij na materiale nauchnyh tekstov [Extraction and classification of collocations on the basis of scientific texts]. *V Int. Scientific and Practical Conf. "Applied Linguistics in Science and Education"*. St. Petersburg, 2010, pp. 356–364.
10. Zakharov V.P., Bogdanova S. Yu. *Vydelenie kollokacij statisticheskimi metodami* [Collocation extraction by statistical methods]. Available at: https://www.docme.ru/doc/1337883/2082/ (accessed 14.03.2018).
11. Jagunova E. V., Pivovarova L. M. Ot kollokacij k konstrukcijam [From collocations to structures]. *Acta linguistica petropolitana. Tr. In-ta lingv. issled. RAN.* 2011. Available at: http://www.webground.su/data/lit/pivovarovayagunova/Otkollokatsiykkonstruktsiyam.pdf (accessed 14.03.2018).
12. Bobkova T. V. Korpus tekstiv: osnovni aspekty vyznachennja [The corpus of texts: the main aspects of definition]. *Naukovyj visnyk KNLU*. Kyiv: KNLU, 2014, no. 29, pp. 11–20.

Надійшла (received) 19.03.2018

Відомості про авторів / Сведения об авторах / About the Authors

Петрасова Світлана Валентинівна (Петрасова Светлана Валентиновна, Petrasova Svitlana Valentynivna) – кандидат технічних наук, Національний технічний університет «Харківський політехнічний інститут», старший викладач кафедри інтелектуальних комп'ютерних систем; м. Харків, Україна; ORCID: https://orcid.org/ 0000-0001-6011-135X; e-mail: svetapetrasova@gmail.com

Кузьміна Марія Олександрівна (Кузьмина Мария Александровна, Kuzmina Maria Oleksandrivna) – Національний технічний університет «Харківський політехнічний інститут», магістр; м. Харків, Україна; e-mail: marika958034@gmail.com.

УДК 811.93

О. В. КАНІЩЕВА, М. В. ЧУХНЕНКО

АВТОМАТИЧНИЙ ПОШУК КЛЮЧОВИХ СЛІВ У КОРПУСІ МАСОВОЇ ЛІТЕРАТУРИ

Проаналізовані основні методи пошуку ключових слів у лінгвістичних корпусах, описані сфери їх застосування та їхні властивості. Розглянуто основні переваги і недоліки методів – кількісний метод аналізу, якісний метод аналізу та статистичний метод, які використовуються у сучасній корпусній лінгвістиці. Розроблено корпус масової літератури та описана його структура. Також було розроблено програмне забезпечення, яке реалізує автоматичний пошук ключових слів у створеному корпусі. Проведено аналіз отриманих результатів роботи програми.

Ключові слова: корпус, корпусна лінгвістика, масова література, статистичний аналіз, літературні жанри, ключові слова, Weirdness.

О. В. КАНИЩЕВА, М. В. ЧУХНЕНКО

АВТОМАТИЧЕСКИЙ ПОИСК КЛЮЧЕВЫХ СЛОВ В КОРПУСЕ МАССОВОЙ ЛИТЕРАТУРЫ

Проанализированы основные методы поиска ключевых слов в лингвистических корпусах, описаны сферы их применения и их свойства. Рассмотрены основные преимущества и недостатки методов – количественный метод анализа, качественный метод анализа и статистический метод, которые используются в современной корпусной лингвистике. Разработан корпус массовой литературы и описана его структура. Также было разработано программное обеспечение, которое реализует автоматический поиск ключевых слов в созданном корпусе. Проведено анализ полученных результатов работы программы.

Ключевые слова: корпус, корпусная лингвистика, массовая литература, статистический анализ, литературные жанры, ключевые слова, Weirdness.

O. V. KANISHCHEVA, M. V. CHUKHNENKO

AUTOMATIC SEARCH OF KEYWORDS IN THE BELLES-LETTRES CORPUS

The basic methods of search of key words in linguistic corpus have been analyzed, the spheres of their application and their properties have been described. The main advantages and disadvantages of the methods used in modern corpus linguistics – quantitative analysis method, qualitative analysis method and statistical method have been considered. The corpus of mass literature has been developed and its structure has been described. Besides formula of Weirdness which has been used in the software has been described. Also, software that implements the automatic search of keywords in the created corpus has been developed. The results of the program have been analyzed.

Keywords: corpus, corpus linguistics, mass literature, statistical analysis, literary genres, keywords, Weirdness.

Вступ. Розвиток корпусної лінгвістики, а також побудова корпусів є однією з актуальних проблем сучасного мовознавства. Корпус в лінгвістиці – це сукупність текстів, яка зібрана в єдине ціле за