УДК 801:001.89

Бойко О. О.,
аспірант кафедри прикладної лінгвістики
Одеського національного університету імені І. І. Мечникова

ВИКОРИСТАННЯ ПРОГРАМИ MICROSOFT OFFICE EXCEL ДЛЯ ЛІНГВІСТИЧНОГО АНАЛІЗУ ТЕКСТУ

Анотація. У статті викладається методика користування програмою Microsoft Office Excel для оперативного лінгвістичного опрацювання великого масиву текстів. Досліджено функції сучасного програмного забезпечення та розкрито можливості використання тегсету для сегментації матеріалу та швидкого пошуку необхідних фрагментів тексту за допомогою хештегової індексації. Показано актуальність використання програмного забезпечення як альтернативи традиційному картографуванню.

Ключові слова: комп'ютерна лінгвістика, корпусна лінгвістика, індексація, хештеги, програмне забезпечення, Microsoft Office Excel.

Постановка проблеми у загальному вигляді. Лінгвістика сьогодні перебуває на тому етапі розвитку, коли класичні методи обробки текстів поступаються сучасним, технологічним. Про це свідчить активний розвиток корпусної лінгвістики та її підрозділу – комп'ютерної лінгвістики. Велика кількість текстів – сучасних, класичних, усних та письмових, завантажуються в національні корпуси, їх розмічають за допомогою спеціальних програм – тегерів та парсерів, що дає змогу віднаходити словоформи за різними критеріями, а також коркондансерів, за допомогою яких слова подаються разом із контекстом. Однак для роботи над таким корпусом потрібно мати певну кваліфікацію та завантажувати специфічне програмне забезпечення. **Актуальність** розвідки полягає в тому, що на сучасному етапі розвитку комп'ютерної лінгвістики існує дуже мало методичних розробок, які б впроваджували використання комп'ютерних засобів в обробці масивів текстів для індивідуальних досліджень.

Аналіз останніх досліджень та публікацій. Корпусна та комп'ютерна лінгвістика як дисципліни викладаються в працях українських (В. Жуковська, Є. Захаров, В. Карпіловська, О. Селіванова та інших) та закордонних (В. Плунгян, Є. Калініна, В. Бочаров, M. Baroni та інших) дослідників. Зокрема, В. Жуковська розрізняє корпусну та прикладну лінгвістику, характеризуючи комп'ютерну лінгвістику як таку, що відрізняється «обов'язковістю використання комп'ютерних засобів до оброблення лінгвальних даних<...> Комп'ютерна лінгвістика займається розвязанням таких проблем, як автоматичний переклад, автоматизоване добування інформації з природних текстів, *конструювання зручних інтерфейсів між людиною та машиною* [курсив наш], кількісний опис спілкування на природних мовах» [1, с. 19]. Саме конструювання зручного інтерфейсу між людиною та машиною маємо на меті в наведеному дослідженні.

В. Жуковська, описуючи процедуру корпусного аналізу, акцентує увагу на трьох етапах: 1) ідентифікація мовних даних за допомогою категоріального аналізу; 2) співвідношення мовних даних за допомогою статистичних методів; 3) інтелектуальна інтерпретація результатів. Якщо перші два кроки повинні бути найбільшою мірою автоматизованими, то останній вимагає людської розумової сутності, адже будь-яка інтерпретація є актом залучення розумових здібностей, а тому не може бути переведена в алгоритмічну процедуру [1, с. 20]. Відзначимо, що під час індивідуального опрацювання текстів дослідник може інтерпретувати та анотувати всі фрагменти власними силами без залучення Інтернет-спільноти.

Як відзначає дослідниця «на разі у складі Національної словникової бази Українського мовно-інформаційного фонду Національна академія наук (далі – НАН) України функціонує та постійно розвивається Український національний лінгвістичний корпус (далі – УНЛК), що розробляється під керівництвом академіка НАН України В. А. Широкова» [1, с. 35]. Крім того, існує лінгвістичний портал http://www.mova.info (Інституту філології Київського університету імені Тараса Шевченка), на якому представлено Дослідницький корпус сучасної української мови обсягом понад 3 млн словоформ, на матеріалі якого можна впроваджувати нові лінгвістичні розвідки та використовувати його як інформаційно-довідкову систему.

Метою дослідження є пояснення методу використання офісної програми Microsoft Excel як альтернативи паперовому картографуванню, а також впровадження розмітки інформації за допомогою хештегів для подальшого полегшення пошуку та індексації фрагментів тексту. **Об'єктом** дослідження є використання програмного забезпечення під час лінгвістичного аналізу текстів, **предметом** – створення тегсету в програмі Microsoft Office Excel для оперативної індексації необхідного матеріалу. **Базою даних** слугують сучасні фентезійні твори.

Виклад основного матеріалу дослідження. Класичним методом збору та обробки матеріалу є картографування. Сьогодні картографування зберігає свою актуальність, але в трансформованому вигляді. Оформлення паперових карток, їх індексація, подальше переведення інформації в електронний варіант є на сьогодні неефективним засобом з тієї причини, що гуманітарні дослідження потребують опрацювання великої кількості текстів та більш точного його індексування. Більшість текстів наявні в електронному варіанті у вільному доступі, а, отже, їх можна опрацьовувати без роздруковування на паперових носіях. Проблемним питанням у багатьох дослідників є структурованість різноскерованого матеріалу, зібраного з різних джерел.

Якщо проводити аналогію між індивідуальним масивом текстів, що обробляється одним дослідником, та національним корпусом текстів, який є результатом роботи багатьох користувачів, побачимо, що в національному корпусі використовуються анотації, або теги (tags, annotation) – «спеціальні мітки, що приписуються словам у текстах корпусу та позначають різноманітні лінгвістичні категорії, наприклад, граматичні, синтаксичні та інші. <...>деякі лінгвісти взагалі несхвально відносяться до процесу анотування корпусу, особливо до внесення анота-

ції в корпус вручну, та це надає вказаному критерію особливої ваги<...> Здійснена анотація певним чином нав'язує користувачеві корпусу готовий лінгвістичний аналіз даних, здійснений на основі певних наукових позицій укладачів» [1, с. 23]. Схоже анотування ми будемо використовувати під час роботи над лінгвістичною темою, тому що у власній роботі дослідник може будь-яким чином інтерпретувати та анотувати фрагменти текстів, дотримуючись обраної наукової позиції.

Анотування відбувається за допомогою спеціальних міток – тегів, які можуть бути *лінгвістичними* (граматичними), *структурними* (речення, абзац тощо) або *екстралінгвістичними (**метарозмітка**)* – відомості про автора, його вік, стать тощо). Процес розмітки, як відзначає У. Шандрук, передбачає низку таких процедур:

1) сегментизація тексту;
2) формалізація параметрів анотування;
3) створення тегсету чи набору формальних кодів з відповідною семантикою;
4) визначення анотаційної схеми та її принципів [7].

Для дослідника в індивідуальній роботі необхідними є два пункти: сегментація тексту по картках та створення тегсету. Ми покажемо, як саме можна використовувати офісну програму Microsoft Excel для виконання цих завдань.

Переважна більшість користувачів ПК працюють з операційною системою Windows 7 або 10. Зважаючи на це, найактивніше використовуються офісні програми типу Microsoft Office Word (зберігає текстовий документ у розширенні .doc, .docx), LibraOffice (із розширенням .odt) і, набагато рідше (через обмежений функціонал), текстовий редактор типу WordPad (розширення .txt) та інші. Наведені текстові редактори дають змогу форматувати текст, обирати тип та розмір кеглю, встановлювати відповідний до вимог міжрядковий інтервал тощо.

Проте ширші можливості для зручного опрацювання текстового матеріалу надає офісна програма Microsoft Excel. Найчастіше нею користуються для роботи з математичними формулами, розрахунками, обліком тощо. Функції Excel можна активно використовувати й в гуманітарній галузі.

Починати роботу потрібно зі створення нового документу. Новий документ Excel має за замовчуванням три ***аркуші***, сукупність аркушів є ***книгою***. За необхідністю аркуші можна додавати, змінювати їх назву та колір – це полегшує візуальне сприйняття інформації користувачем. Для виконання цієї операції потрібно натиснути правою кнопкою миші на один з аркушів та вибрати з меню пункт «Додати» – «Аркуш». У тому ж меню можна обрати функцію «Перейменувати» або «Колір аркуша». Також користувач має можливість переміщувати аркуші в довільному порядку, затиснувши ліву кнопку миші на аркуші. Кількість створюваних аркушів – необмежена, отже, на одному аркуші можна картографувати одне або більше джерел. Зручним є використання одного аркуша для одного джерела. Для пересування між аркушами (коли їх вже велика кількість) можна використовувати стрілки, розміщені в лівому нижньому куті екрану. Стрілки ◀▶ пересувають на один аркуш вліво/вправо, стрілки |◀▶| – в кінець або початок списку аркушів.

На всіх аркушах є стовпці та рядки, тобто кожен аркуш – це гіпотетично безмежна таблиця, а будь-яка клітинка, утворена на перетині стовпця та рядка – готова картка для збору матеріалу, що полегшує **сегментацію тексту**. Для більшої зручності перед початком роботи з аркушем необхідно відформатувати клітинку для уніфікації формату зібраної інформації. Для цього необхідно виділити все поле через поєднання клавіш Ctrl+A, після чого зсунути курсором межу першого стовпчика на необхідну ширину на правий бік. Для виконання цієї операції потрібно навести курсор в поле, яке позначене латинськими літерами.

Натиснувши праву кнопку миші викликаємо меню та обираємо «Формат комірок», в якому слід обрати вкладку «Вирів-

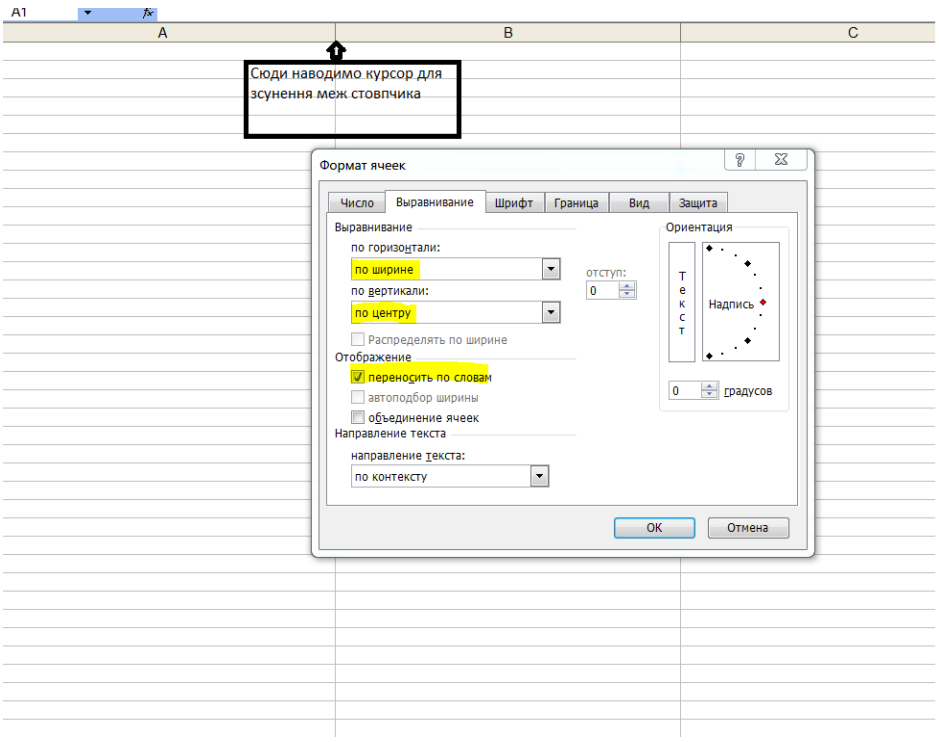

Рис. 1.

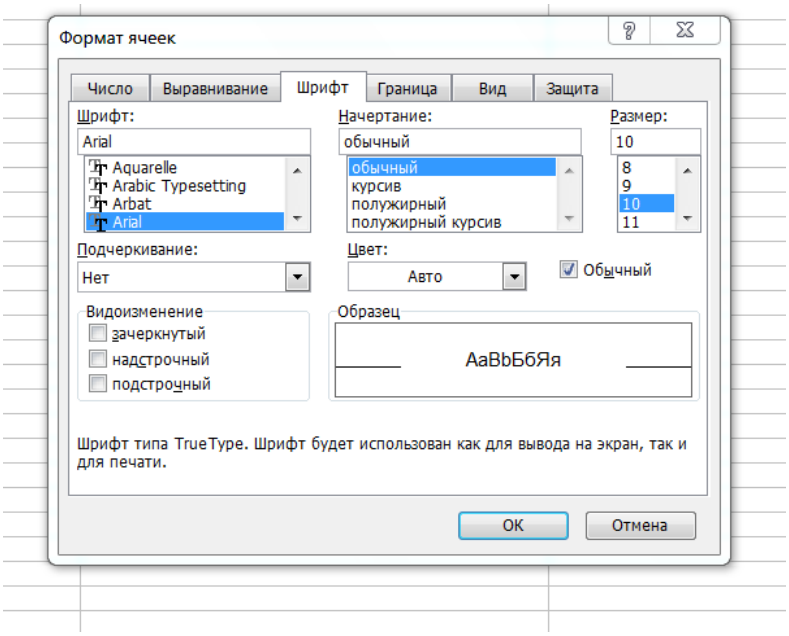

Рис. 2.

нювання». Для більшої зручності читання тексту обираємо вирівнювання за горизонталлю: «За шириною», за вертикаллю – «По центру», та ставимо позначку в полі «Переносити по словах» (див. рис.1).

Тепер будь-який обсяг тексту, розміщений в одній клітинці, буде показуватися повністю та коректно. У вкладці «Шрифт» обираємо Arial або Times New Roman як такі, що є найбільш розбірливими для читання. Кегль – 10, в такому випадку в одну клітинку можна вмістити більше інформації, а клітинка буде меншою за розміром (див. рис. 2).

Процес перенесення інформації в клітинки може відбуватися двома методами. Перший – автоматичний, його можна використовувати лише за наявності електронного тексту. В цьому випадку необхідна інформація виділяється курсором, копіюється через праве меню миші або через гарячі клавіші Ctrl+C (копіювати) та Ctrl+V (вставити). У автоматичного копіювання є лише один плюс – швидкість процесу, але є й низка недоліків: 1) знижений рівень інтелектуального опрацювання інформації користувачем; 2) некоректність копіювання з документів у форматі PDF або з відкритої інтернет-сторінки, що призводить до спотворення тексту; 3) зміна форматування (кеглю, переносів, міжрядкового інтервалу), що порушує уніфікацію подання інформації на картках та ускладнює її візуальне сприйняття користувачем. Ми вважаємо, що більш зручним та коректним є ручне введення інформації в кожну клітинку – як з електронного, так та з паперового носія. Звісно, для швидкого введення бажано опанувати «сліпий» десятипальцевий набір тексту, що значно прискорить роботу будь-якого науковця.

Відзначимо також, що в клітинці Excel не можна починати рядок з тире (якщо потрібно ввести репліку діалогу): програма зчитує це як помилку в формулі. З огляду на це починати рядок потрібно з першого слова або з трикрапки, якщо цитата взята з певного контексту, а не з початку речення.

Під час опрацювання великого обсягу текстів за наявності чітко поставлених теми, мети та завдань бажано аналізувати повністю весь твір, одразу сегментуючи необхідні фрагменти. Для полегшення подальшого пошуку необхідної інформації ми будемо використовувати **хештеги** по типу тих, якими користуються в соціальних мережах для пошуку публікацій за певною темою.

Особливість хештегів полягає в тому, що вони є абсолютно своєрідними, довільними та семантично зрозумілими. Написання хештегів латинкою відрізняє їх від основного тексту та надає можливість шукати тільки потрібну інформацію. «Хештег» відрізняється від звичайного тегу, який використовується в інтернет-корпусах, тим, що перед ним ставимо знак #, наприклад: #fantasy. Хештег може складатися з одного або декількох слів. Два слова можуть писатися разом, через тире або через нижнє підкреслювання (#prectitle, #prec-title, #prec_title). Між # та хештегом, а також між двома словами в одному тезі не повинно бути розривів, інакше програма некоректно ідентифікує запит. З іншого боку, між основним текстом та тегом, між декількома тегами повинен бути розрив – теж для коректності індексації. Як видно з прикладів нижче, слова, що входять до тегів, не обов'язково повинні бути написані з урахуванням граматики англійської або будь-якої іншої мови. Головний орієнтир створення – зручність та зрозумілість для користувача.

Проілюструємо використання хештегів на прикладі роботи над темою «Інтертекстуальність у сучасному фентезі». Під час опрацювання великого корпусу сучасного російського та українського фентезі нам потрібно картографувати інтертекстуальні елементи, тому ми створюємо власну «колекцію» тегів – **тегсет**; з технічних причин, які ми опишемо нижче, їх повинно бути не більше 24.

Тепер під час внесення фрагменту тексту до картки ми аналізуємо які наявні в ньому, та згідно з цим атрибутуємо текст одним або декількома хештегами. Наведемо приклад з книги Д. Корній «Зворотній бік сутіні» [4, с. 32]: *Як це де? У три-*

Ми створили 17 тегів-шифрів:

#citate	Цитати	#oneiric	Оніричні елементи
#allusion	Алюзії	#religion	Релігійні алюзії
#epigraph	Епіграфи	#intediscourse	Інтердискурсивні елементи
#reminisce	Ремінісценції	#phraseology	Фразеологічні одиниці
#referention	Референції	#vstav_zhanr	Вставні жанри
#precname	Прецедентні імена	#prec_hronotop	Прецедентний хронотоп
#precsituation	Прецедентні ситуації	#miphology	Міфологеми
#prectitle	Прецедентні заголовки	#folk	Фольклорні одиниці
#fairy	Казкові алюзії		

дев'ятому царстві, у тридесятій державі<...> Чи як у ваших казках це місце величають? Якщо по-справжньому – це Порубіжжя. – ***#phraseology #fairy #folk #allusion***. Отже, ми зможемо знайти цей фрагмент, якщо нам потрібно буде аналізувати фразеологічні одиниці, казкові та фольклорні алюзії, всі алюзії.

Процес додавання хештегів полегшується автоматизацією. Тегсет зручніше зберігати в окремому текстовому файлі або на окремому листі в Excel. На початку роботи відкриваємо водночас файл із тегами та Excel. Активізувавши вікно Excel, двічі відправляємо команду Ctrl+C: відкриваємо буфер обміну. В буфері обміну водночас може зберігатися до 24 одиниць; якщо додавати більше, нові одиниці будуть заміщувати собою ті, що були додані першими. Тепер починаємо копіювати теги – виділяємо їх мишею та натискаємо Ctrl+C, після чого тег одразу з'явиться в буфері обміну. Перший обраний тег стане останнім в списку, отже, керуємося у виборі релевантністю тегів: рідко використовувані будуть внизу списку (див. рис. 3). На зображенні видно, що поруч відкриті документ Microsoft Word і Microsoft Excel.

Для коректного додавання хештегу після текстового фрагмента необхідно двічі натиснути на клітинку з текстом, потім – один раз натиснути на обраний хештег в буфері обміну, та він додається в клітинку. Якщо потрібно поставити два теги, ставимо розрив та тиснемо на інший хештег. Звернемо увагу, що на «картку» слід тиснути двічі, тому що інакше хештег замінить собою весь раніше набраний текст. Втім, будь-яку некоректну операцію можна скасувати, використовуючи «гарячу» комбінацію Ctrl+Z.

Розмітка за допомогою хештегів дає можливість швидко відшукувати весь обсяг тематичних «карток»-клітинок – як на одному листі, так та у всій книзі документу. Пошук викликається сполученням клавіш Ctrl+F. За замовчуванням пошук ведеться по одному листу (див. Рис. 4), а для пошуку по всій книзі потрібно перейти у «Параметри» та обрати пошук по книзі (див. Рис. 5).

Під час роботи з матеріалом може виникнути необхідність замінити хештег через його неактуальність – значення тегу може розширитися або змінитися. Наприклад, спочатку ми використовували хештег #intermediate на позначення інтермедіальних елементів, але згодом розширили його значення до інтердискурсивного – #interdiscourse. Щоб замінити всі

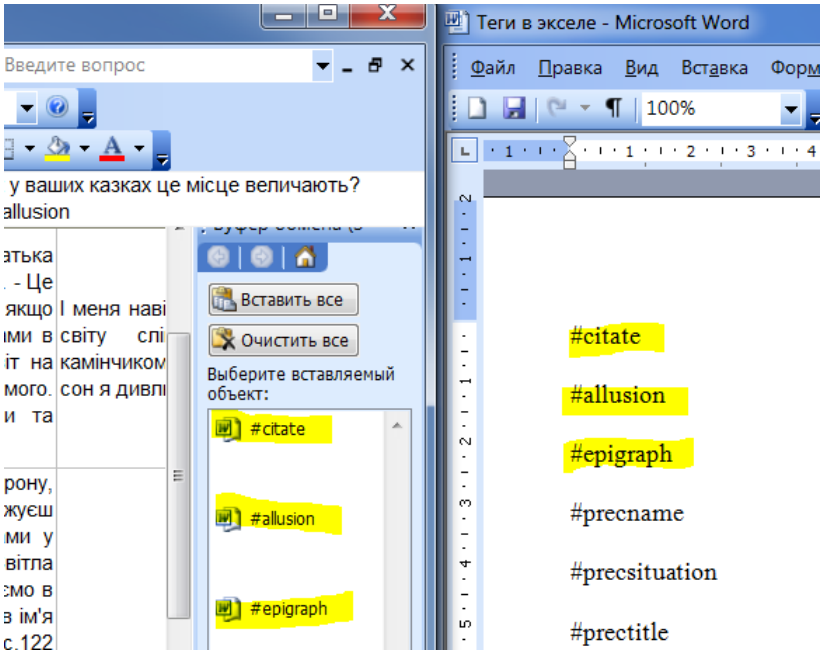

Рис. 3

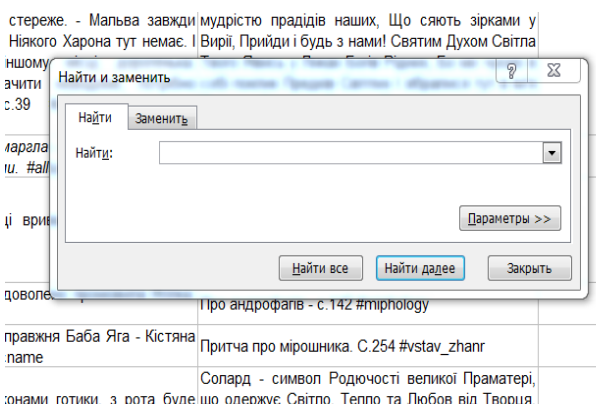

Рис. 4

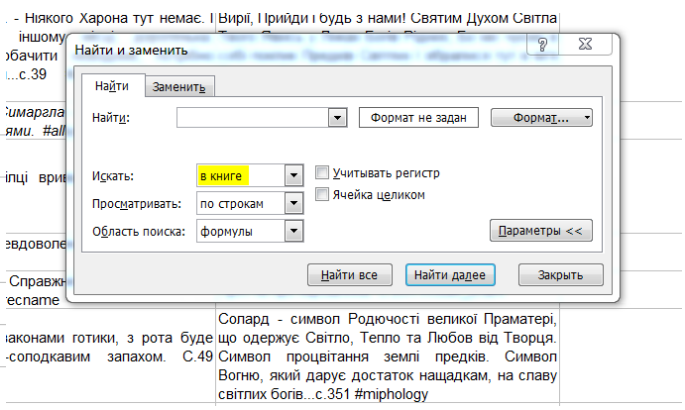

Рис. 5

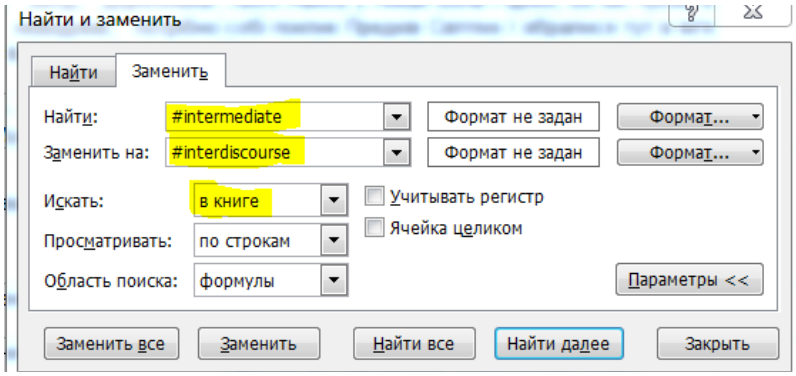

Рис. 6

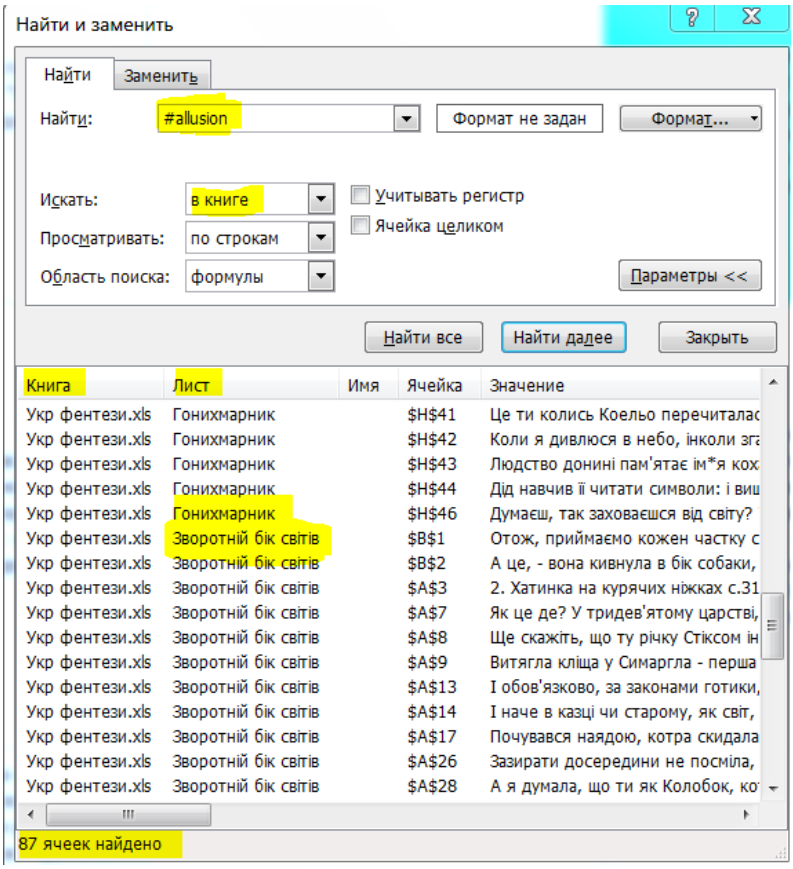

Рис. 7

проставлені хештеги одночасно, потрібно перейти у вкладку «Замінити» та обрати пункт «У книзі» – для заміни по всьому документу або «На аркуші», відповідно, для заміни на одному листі (див. Рис. 6). Після чого обираємо «Замінити все».

Ще одна перевага – після введення запиту у полі пошуку показується кількість знайдених елементів, що допоможе при використанні кількісних методів. Наприклад, у книзі «Українське фентезі» знайдено 87 клітинок за запитом #allusion (див. Рис. 7). З правого боку видно, який текст знаходиться на клітинці. Шукати можна одразу за декількома тегами – так ми знайдемо фрагменти тексту, в якому є, припустимо, та алюзії, та цитати, та прецедентні імена (в тому випадку, якщо ми розмітили клітинку трьома хештегами).

Висновки та перспективи дослідження. Використання програмних засобів персонального комп'ютера та розуміння принципів індексації інформації та її пошуку стануть в пригоді дослідникам у будь-якій сфері, але особливо – в гуманітарних науках, які вимагають опрацювання великих масивів текстів. Розмітка інформації за допомогою хештегів активізує інтелектуальний процес інтерпретатора, який одразу анотує обраний фрагмент, а також розмітка полегшує пошук необхідних фрагментів тексту під час написання тексту статті, курсової, дипломної роботи або дисертації. Програму Microsoft Excel можна використовувати як альтернативу «аналоговому» картографуванню на паперових носіях, тому що в цьому випадку вся інформація зберігається в одному документі, легко індексується та піддається статистичній обробці. Ми вважаємо перспективними пошуки нових функцій офісних програм, які встановлені на будь-якому персональному комп'ютері, а також пошук можливості впроваджувати принципи подібної індексації в Інтернет-просторі, та у перспективі – створення масивів інтертекстуальних елементів, міфологем або інших дискурсивних одиниць в українському національному корпусі текстів.

Література:

1. Жуковська В.В. Вступ до корпусної лінгвістики : навчальний посібник. Житомир : Вид-во ЖДУ ім. І. Франка, 2013. 142 с.
2. Захаров В.П., Богданова С.Ю. Корпусная лингвистика : учебник для студентов гуманитарных вузов. Иркутск : ИГЛУ, 2011. 161 с.
3. Карпіловська Є.А. Вступ до прикладної лінгвістики : комп'ютерна лінгвістика. Донецьк : ТОВ «Юго-Восток, Лтд», 2006. 188 с.
4. Корній Д. Зворотній бік сутіні. Харків : Книжковий Клуб Сімейного Дозвілля, 2016. 288 с.
5. Селіванова О.О. Корпусна лінгвістика. *Сучасна лінгвістика : напрями та проблеми :* підручник. Полтава : Довкілля-К, 2008. С. 667–669.
6. Сребрянская Н.А. Новые возможности лингвистического исследования художественного текста с помощью компьютера. URL: https://cyberleninka.ru/article/v/novye-vozmozhnosti-lingvisticheskogo-issledovaniya-hudozhestvennogo-teksta-s-pomoschyu-kompyutera
7. Шандрук У. Дослідження особливостей проектування корпусу на основі текстової моделі. URL: http://ena.lp.edu.ua:8080/ bitstream/ntb/40815/2/2017_Shandruk_UDoslidzhennia_osoblyvostei_45-47.pdf
8. Шведова М., Січінава Д. Корпусна лінгвістика та лексико-граматична типологія. *Українське мовознавство.* Київський національний університет імені Тараса Шевченка. № 43. 2013. С. 95–103.

Бойко О. А. Использование программы Microsoft Office Excel для лингвистического анализа текста

Аннотация. В статье излагается методика использования программы Microsoft Office Excel для оперативной лингвистической обработки большого массива текстов. Исследуются функции современного программного обеспечения и раскрываются возможности использования тегсета для сегментации материала и быстрого поиска необходимых фрагментов текста с помощью хештеговой индексации. Показывается актуальность использования программного обеспечения как альтернативы традиционному картографированию.

Ключевые слова: компьютерная лингвистика, корпусная лингвистика, индексация, хештеги, программное обеспечение, Microsoft Office Excel.

Boiko O. Using Microsoft Office Excel for linguistic text analysis

Summary. The article describes the method of using the program Microsoft Office Excel for the operational linguistic processing of a large array of texts. The functions of modern software are explored and the possibility of using a tagset for material segmentation and a quick search for the necessary text fragments using hashtag indexing is revealed. The urgency of using software as an alternative to traditional mapping is shown.

Key words: computational linguistics, corpus linguistics, indexing, hashtags, software, Microsoft Office Excel.