УДК 811.111.

СТВОРЕННЯ ВЛАСНОГО КОРПУСУ АМЕРИКАНСЬКИХ КІНОСЦЕНАРІЇВ

Скобнікова О. В.
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

У статті досліджено проблему створення власного корпусу текстів на прикладі корпусу американських кіносценаріїв фільмів, призначених для сімейного перегляду. З'ясовано методику та критерії конструювання лінгвістичних корпусів. Розглянуто типологію та основні характеристики розробленого корпусу.

Ключові слова: *корпус, кіносценарій, розмітка, метарозмітка, лексична насиченість, лексична щільність, формальність, ключові слова.*

Skobnikova O. V. Creating the own corpus of American film scripts. *The article deals with the problem of creating the own corpus of texts on the example of the corpus of American family film scripts. The methodology and criteria for constructing linguistic corpus are considered. The typology and main characteristics of the created corpus are determined. Special attention is drawn to the technological process of creating a corpus which included: finding the sources of linguistic material; entry of the data in the form of the texts of the film scripts presented in plain text format (*.txt), annotation, part-of-speech (POS) tagging, converting tagged texts into a specialized linguistic information retrieval system or corpus manager which provides rapid multi-dimensional search and statistical processing. In this research we used the AntConc manager. We focused on analysis of the created corpus which included: defining the total number of tokens and total number of types in the corpus, finding the type-token ratio (TTR) and standard type-token ratio (STTR), making a list of the most frequent word forms, clarification of the hapax legomena (words used in corpus only once), detection of frequency of distribution of different parts of speech, finding the index of the lexical density, defining the average length of the sentence of the corpus, determination of the index of formality, making a keyword list. We found that most key words are lexically neutral, belong to the core vocabulary and relate to everyday family life. Two words from the list belong to the colloquial style. There are also words that occur to be technical and directorial remarks.*

Key words: *corpus, film script, tagging, annotation, type-token ratio, lexical density, formality, key words.*

Постановка проблеми та обґрунтування актуальності її розгляду. Лінгвістичний корпус – це зібрання текстів, в основі якого лежить логічний задум, логічна ідея, яка об'єднує тексти. Тип корпусу та його структура залежать від його призначення та визначаються широким спектром дослідницьких і прикладних завдань, для розв'язання яких вони створюються, та особливостями мовного матеріалу, покладеного в їх основу.

Відповідно до цих чинників, корпусна лінгвістика оперує двома різними типами корпусів текстів. Корпуси першого типу універсальні, вони відображають в собі все різноманіття мовної діяльності. До цього типу належать масштабні корпуси Brown University Standard Corpus of Present-Day American English та Corpus of Contemporary American English.

Корпуси другого типу відображають об'єктивацію певного лінгвістичного або культурного феномену в суспільній мовній практиці, вони побудовані ad hoc (для спеціальної мети) [5]. До укладення таких корпусів вдаються, коли виникає необхідність вивчити певні тексти, які досі не увійшли до відомих корпусів. У цьому випадку лінгвіст може скласти свій власний корпус зі своїх джерел і досліджувати вже його.

Аналіз останніх досліджень і публікацій. На сьогодні як вітчизняні (Демська-Кульчицька [1], Жуковська [2]), так і зарубіжні (Gries [8], McKee [10], Nini [11]) науковці порушують проблему створення власних корпусів. Прикладами таких спеціальних корпусів можуть слугувати Eearly Modern English Medical Texts та Middle English Medical Texts від видавництва John Benjamins.

Формулювання мети і завдань статті. Для нашого дослідження ми створили власний корпус, що складається з текстів англомовних кіносценаріїв як основного складника кінотекстів – Corpus of American Family Movies (COAFM), що і визначено **метою** дослідження. Задля досягнення мети слід виконати такі **завдання**:

- визначити джерела лінгвістичного матеріалу;
- ввести отримані тексти в корпус даних;
- виконати попереднє опрацювання текстів;
- розмітити тексти корпусу;
- конвертувати розмічені тексти в корпусний менеджер;
- розглянути основні характеристики створеного корпусу.

Виклад основного матеріалу дослідження. Однією із вагомих проблем сучасної корпусної лінгвістики є визначення обсягу корпусу, достатнього для отримання надійних висновків. Так, згідно з найбільш загальноприйнятим підходом, мінімальний обсяг корпусу першого типу повинен становити не менше 1 мільйона слововживань (word tokens). Щодо обсягу корпусу другого типу немає одностайності, оскільки корпусний аналіз можливо здійснювати навіть на дуже малих за обсягом текстах. Прикладом може слугувати студія американського

мовознавця Майкла Стаббса, у якій він досліджував лінгвістичні особливості лише двох листів обсягом у декілька сотень слів [12, 81−100]. Але в основному дослідники погоджуються, що для повноцінного лінгвістичного аналізу такий корпус повинен містити не менше 10 тисяч слововживань. Корпус COAFM налічує більше 700 тисяч слововживань, а це означає, що він достатньо великий, щоб надати об'єктивну та репрезентативну інформацію про особливості вживання досліджуваних лінгвістичних явищ, одержати можливість визначити, що є типовим, а що рідкісним явищем.

Методологія конструювання корпусу залежить від його типу. Побудова корпусів першого типу ґрунтується на принципі дедукції – реалізації проблеми коректності руху від загального (об'єктивно реальної мовної практики носіїв мови) до конкретного корпусу текстів, що відображає це загальне. Методологія побудови корпусів другого типу повинна коректно відображати конкретні лінгвістичні феномени в корпусі текстів, спеціально створеному для їх відображення [3]. Зазначимо, що ці обидва підходи часто застосовуються комбіновано.

Дослідники виділили основні критерії для конструювання лінгвістичних корпусів:

1) корпус має бути достатньо великого обсягу;

2) корпус має бути структурованим або розміченим;

3) тексти-складники певного корпусу мають бути в електронному варіанті;

4) в поняття «електронний корпус» входить, як правило, спеціальне програмне забезпечення для роботи із цим корпусом. При конструюванні корпусу COAFM всі ці критерії були враховані та будуть детально описані нижче.

При створенні корпусу COAFM ми послуговувались технологічним процесом створення власного корпусу, запропонованим В. В. Жуковською [2, 85–87], що передбачає поступове виконання таких кроків:

1. Визначення джерел лінгвістичного матеріалу. Ми використовували публічно доступні джерела, а саме інтернет-сайти www.imsdb.com, http://www.script-orama.com, http://scripts-onscreen.com, www.simplyscripts.com, www.moviescriptsandscreenplays.com, з яких ми обрали 50 кіносценаріїв, перелічених нижче.

2. Введення даних. Тексти кіносценаріїв в корпусі представлені в простому текстовому форматі (plain text, *.txt). Plain text – це проста послідовність літер, пропусків (пробілів) і знаків пунктуації, – отже, цей формат розпізнає більшість корпусних менеджерів.

3. Попереднє опрацювання тексту. На цьому етапі всі тексти, отримані з різних джерел, пройшли перевірку й коректування. Також ми здійснили зовнішню розмітку, або метарозмітку (annotation), яка містить відомості про авторів та відомості про тексти: автор, назва, рік видання, тематика.

4. Розмітка тексту. Розмітка (tagging) була зроблена за допомогою інтернет ресурсу CLAWS (the Constituent Likelihood Automatic Word-tagging System), що знаходиться за посиланням http://ucrel.lancs.ac.uk/claws/. Цей сайт пропонує тегування за частинами мови (Part-of-speech (POS) tagging), або морфологічне тегування, яке є найпоширенішою формою анотації корпусу.

5. Заключний етап. Цей етап передбачає конвертування розмічених текстів у структуру спеціалізованої лінгвістичної інформаційно-пошукової системи, або корпусного менеджеру, що забезпечує швидкий багатоаспектний пошук і статистичну обробку. В нашому дослідженні використовуємо менеджер AntConc, який пропонує опції потужного конкордансера, генератора частотного списку, аналізатора сполучуваності, візуалізує входження досліджуваних слів у певному масиві та має багато інших функцій.

Розглянувши класифікації корпусів, запропоновані Е. Харді, Т. МакЕнері, В. В. Риковим, В. П. Захаровим та О. Демською-Кульчицькою [1; 3; 5; 9], ми дійшли висновків, що залежно від певних критеріїв корпус COAFM можна віднести до таких типів:

• за типом мовних даних це корпус писемного мовлення;

• за мовою текстів – одномовний, англомовний;

• за критерієм літературності – змішаний;

• за доступністю – закритий, оскільки він створений з вузькоспецифічною метою та не призначений для публічного використання;

• за метою створення – спеціалізований, оскільки обмежується одним жанром;

• за хронологічністю – синхронний, адже містить тексти конкретного часового проміжку, а саме 1964–2016 років написання;

• за способом існування, або динамічністю, цей корпус належить до статичних, бо відображає певний часовий стан мовної системи;

• за призначенням – ілюстративний, оскільки створений не тільки задля того, щоб виявити нові факти, але щоб підтвердити і обґрунтувати результати, вже отримані нами під час дослідження корпусів BC та COCA;

• за обсягом текстів – повнотекстовий;

• за спільністю авторства – загальний, оскільки містить тексти, написані різними авторами;

• за розміткою та її характером цей корпус є розміченим, тобто таким, у якому словам та реченням присвоєні певні теги (tags), зокрема зроблено синтактико-морфологічну розмітку.

Розглянемо загальні характеристики корпусу Corpus of American Family Movies.

Корпус складається з текстів п'ятдесяти кіносценаріїв, що належать до жанру сімейного кіно та були зняті в 1964–2016 роках: It's a Wonderful Life (1964), Mary Poppins (1964), E.T. the Extra-Terrestrial (1982), The Flinstones (1987), Big (1988), Field of Dreams (1989), Father of the Bride (1991), The Addams Family (1991), My Girl (1991), Mrs. Doubtfire (1993), The Secret Garden (1993), My Girl 2 (1994), Liar! Liar! (1996), Parent Trap (1997), The Family Man (2000), The Royal Tenenbaums (2001), Stuart Little (2001), My Big Fat Greek Wedding (2002), Elf (2003), Cheaper by the Dozen (2003), Cheaper by the Dozen 2 (2005), The Pacifier (2005), Nanny McPhee (2005), Charlotte's Web (2006), The Nanny Diaries (2007), Old Dogs (2009), Tooth Fairy (2010), A Nanny for Christmas (2010), Letters to God (2010), Nanny McPhee Returns (2010), Ramona and Beezus (2010), Diary Of A Wimpy Kid (2010), The Tree of Life (2011), We Bought a Zoo (2011), The Descendants (2011), Judy Moody and

the Not Bummer Summer (2011), The Odd Life of Timothy Green (2012), What Maisie Knew (2012), Moonrise Kingdom (2012), Louder Than Words (2013), Annie (2014), Alexander and the Terrible, Horrible, No Good, Very Bad Day (2014), Boyhood (2014), Captain Fantastic (2014), Little Men (2016), The Edge of Seventeen (2016), Bridget Jones's Baby (2016), The Hollars (2016), The Real O'Neils (2016), My Big Fat Greek Wedding 2 (2016), The Great Gilly Hopkins (2016). Вибірка була виконана на основі даних інтернет-сайтів, що пропонують перелік найкращих фільмів для сімейного перегляду: 50 Best Family movies (https://www.imdb.com/list), Top 50 Kids & Family Movies (https://www.rottentomatoes.com/), 50 Best Kids Movies to Watch Together on Family Movie Night (https://www.timeout.com),The 40 Best Family & Kids Movies (https://www.pastemagazine.com). Ці фільми можна оцінити як репрезентативні, оскільки вони відповідають усім характеристикам сімейних кінофільмів. Отже, вони можуть бути розглянуті як модель фільмів для сімейного перегляду.

Одиницею зберігання в корпусі є окремий текст кіносценарію, до того ж назва файлу складається з назви твору та року його написання.

Загальна кількість слововживань (tokens) в корпусі COAFM становить 710064 одиниці, загальна кількість словоформ (types) – 23004 одиниці. Зазначимо, що словоформа це повторювана одиниця мови, однакова послідовність звуків або букв, а слововживання – одиниця мовленнєвої діяльності, будь-який ланцюжок букв або звуків між двома пробілами [6; 7]. Згідно з лінгвостатистичним законом Д. Ципфа, сутність якого полягає в тому, що відношення рангу слова в частотному словнику до частотності слова в мові становить постійну величину (константу), у будь-якому масиві текстів невелике число словоформ утворює більшу частину реальних слововживань [4].

Лексична насиченість корпусу (type-token ratio, або TTR), що складає 3,2%, була обчислена за формулою, запропонованою Е. Харді та Т. МакЕнері [9]:

TTR = Vt / Nt x 100, де

Vt – number of types

Nt – number of tokens.

Проте зазначимо, що індекс TTR не є інформативним для нашого дослідження, позаяк він залежить від обсягу тексту: що більший текст, то менший відсоток. У такому випадку більш доречним було обчислити ***стандартизовану лексичну насиченість*** (standard type-token ratio, або STTR), яка вираховується за тією ж формулою [10]. В основі обчислення цього індексу лежить метод випадкового відбору з тексту фрагментів довжиною 1000 слів і обчислення для них TTR з подальшим усередненням одержаних даних. Скориставшись цим методом, ми обчислили TTR у п'яти фрагментах і з'ясували, що STTR корпусу COAFM дорівнює 3,3%. Це дещо менше ніж, наприклад, в корпусі Brown Corpus, індекс STTR якого становить 4%.

Список найбільш частотних словоформ у будь-якому англомовному корпусі не має значних відмінностей та складається в основному з артиклів, сполучників, прийменників, займенників та деяких дієслівних форм. Корпус COAFM загалом відповідає середнім значенням для англійської мови, що засвідчує перелік 50 ***найчастотніших словоформ***:

the, you, I, a, to, and, it, of, in, is, that, he, do, on, we, what, this, for, me, with, at, his, my, her, are, she, your, up, have, not, out, no, am, just, be, all, know, they, but, so, was, like, as, oh, go, him, get, there, here, right.

На противагу найуживанішим, в корпусі є багато слів, які вживаються тільки один раз. Такі слова називають ***hapax legomena*** (від грецького «щось, сказане один раз»). Помічено, що в середньому майже 40% слів у корпусі подибуємо лише один раз. С. Гріс, американський експерт у сфері корпусної лінгвістики, вважає, що такі нечастотні, поодинокі вживання так само необхідно враховувати при корпуснолінгвістичному аналізі, як і найчастотніші слововживання. Цю так звану «не-зустрічальність» (non-occurrence) С. Гріс вважає релевантною для корпуснолінгвістичного аналізу [8].

Слів, що трапляються в корпусі COAFM лише 1 раз, виявлено 9319 одиниць, що становить 40,5%, і це підтверджує загальну теорію, зазначену вище. Серед таких слів є всі повнозначні частини мови, окрім числівників та займенників: прикметники (*harmful, ecological, claustrophobic*), іменники (*skyscraper, jumpsuit, womanizer*), дієслова (*breathed, rotated, tremble*), прислівники (*gladly, brilliantly, wonderfully*) та вигуки (*hahaha, eeh, uhm*).

Попереднє проведення морфологічної розмітки дало нам змогу з'ясувати, що ***розподіл частотності різних частин мови*** в корпусі виглядає таким чином:

• прикметники (AJ) 36372 од.;

• іменники (NN) 119731 од., серед них власні назви (NP) 53944 од.;

• дієслова (V) 154673 од., серед них смислові дієслова (VV) 97879 од., модальні дієслова (VM) 9944 од.;

• прислівники (AV) 56588 од.;

• сполучники (CJ) 25459 од.;

• займенники (PN) 77731 од.;

• артиклі (AT) 42657 од.;

• вигуки (IT) 12219 од.;

• числівники (RD) 5425 од., серед них кількісні (CRD) 4055 од. та порядкові (ORD) 1370 од.;

• прийменники (PR) 50264 од.

Зазначимо, що прийменник *of* (PRF) є найчастотнішим та налічує 9583 одиниці, тоді як всі інші прийменники (PRP) налічують 40663 одиниці, тобто його частка становить 19%. До того ж цей прийменник функціонує не зовсім так, як інші прийменники. Більшість прийменників входить в колокації зі словами, що за ними слідують (*through the sky, in trouble, for a man*), а *of* превалює в словосполученнях зі словами, що йому передують (*group of, top of, glimpse of*).

Найчастіше в корпусі натрапляємо на дієслова, які становлять 27% від усіх слів, та іменники, які складають 21%, що свідчить про те, що тип тексту кіносценаріїв здебільшого наближений до розповіді. Цікаво, що прикметники в кіносценаріях становлять лише 6%, так як частка описів в цьому жанрі невелика. Найрідше подибуємо числівники (2%) та вигуки (1%).

Завдяки наявності морфологічної розмітки корпусу ми вирахували індекс ***лексичної***, або ***функціональної щільності*** (the index of the lexical density), який є співвідношенням кількості службових слів до кількості повнозначних слів [13]. Більш лексично щільними,

таким чином, є тексти, у яких використовується менше службової лексики. В корпусі COAFM цей індекс становить 0,485% та обчислений за формулою

$$L_{dn} = N_f / N_c$$ де

L_{dn} – коефіцієнт лексичної щільності,

N_f – кількість службових слів (function words),

N_c – кількість повнозначних слів (content words)

За допомогою інтернет-сайту http://textalyser.net/ нами була визначена ***середня довжина речення*** корпусу COAFM, що дорівнює 5,5 слів. Найкоротше речення містить всього одне слово, наприклад: (*Parents… Dad! No!*). Найдовше речення є режисерською ремаркою та складається з 25 слів*:*

He looks into the catcher's mitt, shakes off the first signal, takes the turn, wipes the sweat off his brow, leans back and fires.

Також ми вирахували ***індекс формальності*** (the index of formality**)**, який становить 0,32% та визначений за формулою, запропонованою А. Ніні [11]

$$F = (NN + AJ + PR – PN – VV – CJ – AV – IT + 100) / 2$$

Зіставивши корпус COAFM з референтним корпусом Brown Corpus, ми склали ***список ключових слів*** (key words) досліджуваного корпусу.

Ключові слова обчислюються шляхом порівняння частотності слова в досліджуваному корпусі з частотністю слова в корпусі довідковому (референтному, або опорному). У результаті порівняння частоти слововживань система привласнює одиницям, що відрізняються несподівано високою частотністю, статистичну назву ключового слова. Зазначимо, що ключові слова визначаються лише порівняно з референтним корпусом, який, як правило, більшого розміру, та не тотожні словам з найвищою частотністю. Наприклад, найчастотніше слово в досліджуваному корпусі, означений артикль *the*, не є ключовим.

У досліджуваному тексті (корпусі) обробляються всі слова, за винятком тих, які сам дослідник може помістити в так званий StopList. На даному етапі ми визначили список заборонених слів (stop-words) з метою відсіяти службову лексику (прийменники, сполучники та ін.) та власні назви.

Задля підрахунку величини «ключового характеру» (keyness) програма обробляє чотири параметри: частотність слова в досліджуваному корпусі, частотність слова в опорному корпусі, кількість усіх слів у досліджуваному корпусі і кількість усіх слів в опорному корпусі.

Наведемо перші 50 ключових слів для корпусу COAFM:

little, think, want, mom, really, room, house, night, continued, say, something, door, kids, tell, never, way, mean, sorry, buddy, love, please, thank, marry, home, thing, again, sure, head, old, new, still, even, great, only, should, because, family, hand, people, turns, cut, face, later, car, life, mother, give, first, next, help.

Очевидним є те, що більшість слів є лексично нейтральними, належать до базового вокабуляру (core vocabulary**)** та стосуються повсякденного сімейного життя (*room, house, mother, love, please, thank, home*). Два слова зі списку представляють розмовний стиль (*mom, buddy*). Також наявні слова, що становлять технічні та режисерські ремарки (*next, cut, turns, face*).

Висновки та перспективи подальших досліджень у цьому напрямі. Підсумовуючи все викладене, робимо висновок, що корпус COAFM можна оцінити як репрезентативний з огляду на параметри та характеристики, які були визначені при відборі текстів, та може бути розглянутий як модель функціонування мови сучасних кінотекстів.

Отже, створений нами електронний текстовий корпус мовних даних дає можливість його результативного використання. Це відкриває широкі перспективи для подальших лінгвістичних досліджень, завдання яких передбачають використання мовного матеріалу англомовних американських кіносценаріїв сімейних фільмів.

ЛІТЕРАТУРА

1. Демська-Кульчицька О. М. Базові поняття корпусної лінгвістики / О. М. Демська-Кульчицька // Українська мова. – 2003. – № 1. – С. 42–47.

2. Жуковська В. В. Вступ до корпусної лінгвістики : [навч. посіб.] / В. В. Жуковська. – Житомир : Вид-во ЖДУ ім. І. Франка, 2013. – 142 с.

3. Захаров В. П. Корпусная лингвистика : [учеб. для студентов гуманитарных вузов] / В. П. Захаров, С. Ю. Богданова. – Иркутск : ИГЛУ, 2011. – 161 с.

4. Кутузов А. Б. Корпусная лингвистика. Лекция 2 [Электронный ресурс] / А. Б. Кутузов. – Режим доступа : http://tc.utmn.ru/files/corpus_2.pdf.

5. Рыков В. В. Корпус текстов как реализация объектно-ориентированной парадигмы / В. В. Рыков // Труды Международного семинара «Диалог-2002». – М. : Наука, 2002. – С. 34–35.

6. Гируцкий А. А. Общее языкознание : [учеб. пособие] / А. А. Гируцкий. – Минск : ТетраСистемс, 2003. – 303 с.

7. Щерба Л. В. Языковая система и речевая деятельность / Л. В. Щерба. – М. : КомКнига, 2007. – 427 с.

8. Gries S. Th. Language and Linguistics / S. Th. Gries // Language and Linguistics Compass. – 2009. – № 5. – Vol. 3. – P. 17.

9. Hardie A. Statistics / Andrew Hardie, Tony McEnery // BROWN K. (ed.). Encyclopedia of Language and Linguistics, 2nd edition. – Amsterdam : Elsevier, 2006. – P.138–146.

10. McKee G. Measuring Vocabulary Diversity Using Dedicated Software / Gerard McKee, David Malvern, Brian Richards // Literary and Linguistic Computing. – 2000. – № 15 (3). – P. 323–337.

11. Nini A. Authorship Profiling in a Forensic Context. PhD thesis. Aston Uni. [Electronic resourse] / Andrea Nini. – Mode of access : http://publications.aston.ac.uk/25337/1/Nini_Andrea_2015.pdf.

12. Stubbs M. Text and corpus analysis: computer-assisted studies of language and culture / Michael Stubbs. – Oxford : Blackwell, 1996. – 288 p.

13. Ure J. Lexical density and register differentiation / Jean Ure // G. Perren and J. L. M. Trim (eds). Applications of Linguistics. – London : Cambridge University Press, 1971. – P. 443–452.