Міністерство освіти і науки України

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»

Кафедра мультимедійних систем факультету інформатики

Ukrainian plays' analysis: examining dialogues by gender

Текстова частина до курсової роботи
за спеціальністю «Комп'ютерні науки» 122

Керівник курсової роботи

ас. Смиш О.Р.

_____

(підпис)

"____" _____ 2023 р.

Виконав студент КН-3

Черніков А.А.

"____" _____ 2023 р.

Київ – 2023

Міністерство освіти і науки України
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»
Кафедра мультимедійних систем факультету інформатики

ЗАТВЕРДЖУЮ
Зав.кафедри мультимедійних систем,
Доцент., к. ф.-м. н. О.П. Жежерун
_____(підпис)
“____”_____2022 р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ
на курсову роботу

студенту 3-го року БП КН факультету інформатики

Чернікову Андрію Антоновичу

Тема: Ukrainian plays' analysis: examining dialogues by gender

Зміст ТЧ курсової роботи:
Анотація
Вступ
1. Natural language processing (NLP)
2. Огляд дослідження The Pudding
3. Збір та обробка корпусу п’єс
4. Методи видобування інформації
5. Аналіз отриманих результатів
Висновки
Список літератури

Дата видачі “__” _____ 2022 р. Керівник _____
(підпис)

Завдання отримав _____
(підпис)

Календарний план виконання курсової роботи:

№ п/п	Назва етапу курсової роботи	Термін виконання етапу	Примітка
1.	Узгодження теми для курсової роботи	08.11.2022	
2.	Пошук та аналіз джерел за темою роботи	15.11.2022 – 23.11.2022	
3.	Збір п’єс для обробки	08.12.2022 – 03.01.2023	
3.	Перед обробка корпусу п’єс	12.01.2023 – 26.01.2023	
4.	Розробка методів для вилучення необхідної інформації з файлів	03.02.2023 – 05.03.2023	
5.	Аналіз отриманих даних	16.03.2023 – 02.04.2023	
6.	Складання плану для текстової частини роботи	05.04.2023 – 08.04.2023	
7.	Написання тексту роботи	12.04.2023 – 10.05.2023	
8.	Завершення програмної та текстової частини роботи	10.05.2023	

Зміст

Анотація

У цій роботі використано сучасні методи обробки природної мови для аналізу стереотипів та упереджень, які пов'язані зі статтю персонажів українських п'єс. Висновки, отримані в ході дослідження, можуть знайти застосування у вивченні українськомовного літературного простору в мережі Інтернет.

Вступ

Література є відображенням тієї громади, у якій вона створювалась. Репрезентація статі персонажів у літературі є важливою темою, оскільки вона відіграє роль у формуванні суспільного сприйняття та ставлення до чоловічих та жіночих ролей. Українські п’єси, як і будь-які інші різновиди літературних творів, не є винятком у цьому.

Метою цього наукового дослідження є перевірка наявності гендерних стереотипів у зібраних українських п’єсах та отримання результатів, на основі яких видно, як і яким чином українські п’єси відображають персонажів жіночої та чоловічої статі.

Завданням дослідження є використання інструментів обробки природної мови для створення методів, що вилучають необхідну інформацію для проведення аналізу.

Джерелами для ідеї проведення цього аналізу є дві статті з платформи The Pudding. У них описано, як американські науковці з галузі дослідження даних опрацювали сценарії кінофільмів для пошуку нерівноцінності за статтю персонажів. [2, 3]

Отримані результати вказують на наявність домінування чоловічих персонажів у творах над жіночими та візуалізують конкретну ступінь цієї нерівноцінності. Окрім цього дають змогу переконатися в можливостях функціоналу відповідних аналізаторів та їхньої успішності для проведення подібних досліджень.

Наукова новизна роботи полягає у тому, що подібного аналізу в українськомовному просторі для дослідження залежності між статтю та відображенням чоловічих та жіночих образів в п’єсах – не проводилося.

У першому розділі розглянуто поняття природної мови та сучасні інструменти для її обробки.

Другий розділ присвячено огляду джерел із проведеним дослідженням по сценаріях кінофільмів.

У третьому розділі описано процес збору та обробки природної мови з використанням аналізатору UDPipe.

У четвертому розділі розглянуто методи для виокремлення інформації для аналізу.

П’ятий розділ присвячено візуалізації вилучених даних та знаходженню тенденцій за статтю персонажів.

1. Natural language processing (NLP)

Мова – є невіддільним компонентом будь-якої громади, адже уможливлює комунікацію та взаємодію між людьми. Мова, котрою люди спілкуються в повсякденному житті, називається природною, оскільки вона з'явилася, через необхідність в обміні думками – природно. Людська мова є складною для сприйняття машинами через велику кількість неочевидних конструкцій і правил. Виникла потреба у створенні мов, котрі мали б загальноприйняту структуру, котру могли б розуміти не тільки люди з різних куточків світу, а й комп'ютери. Сучасними штучними мовами є мови програмування та мови символів, що застосовуються в математиці. Природна мова досі не підтримуються жодною мовою програмування через свою перенасиченість та неоднозначність. Відмінності в граматиці та семантиці ускладнюють її пряме використання в мовах програмування, де потрібні чіткі та однозначні правила та синтаксис.

Обробка природної мови (Natural language processing) – це одна з галузей комп'ютерних наук, котра поєднує елементи лінгвістики та штучного інтелекту. Основна мета NLP у тому, щоб зводити людську мову до єдиного формату, для однозначного сприйняття та обробки комп'ютером. Досягання цієї мети відбувається через поєднання загальних граматичних правил, обчислювальної лінгвістики й методів машинного та глибинного навчання (від англ. machine and deep learning).

Граматичні правила відіграють роль правил у мовах програмування. Вони надають обчислювальній машині «розуміти» граматичні структури мови, розподіл слів та формування речень. Обчислювальна лінгвістика є галуззю штучного інтелекту, яка відповідає за аналіз мовленнєвих особливостей, до яких належать фонетика, морфологія, синтаксис та семантика. Фонетика відповідає за звукову структуру мови, морфологія за

побудову форм слів, синтаксис за утворення словосполучень, а семантика за значення та семантичні зв’язки між словами.

Поєднання цих технологій дає змогу електронним обчислюваним машинам виконувати обробку людської мови, використовуючи навчені моделі. Моделі – це набори правил, що було навчено на обширних текстових даних. Великий обсяг забезпечує тренування з різноманітними варіаціями конструкцій мови, що збільшує точність у розпізнаванні та аналізі тексту.

Успішність тренування моделі залежить не тільки від кількості даних, а й від їхньої якості. Чим більше різноманітних лінгвістичних структур наявно в тренувальному корпусі, тим кращі будуть результати опрацювання реальних даних.

Застосування моделей обробки мови уможливлює сприйняття комп’ютером контексту та семантики тексту для визначення змісту та намірів мовця.

У контексті NLP, розуміння семантичного значення намірів мовця відіграє важливу роль, оскільки правильна інтерпретація вислову не завжди збігається з буквальним значенням поодиноких слів.

Кожна людська мова має відповідні моделі в засобах оброки тексту. Проте складність полягає в тому, що підтримка та оновлення цих моделей залежить від потреби в використанні конкретної природної мови. Українська мова на відміну від англійської не є міжнародною, як наслідок не має широкої підтримки на всіх інструментах обробки мови.

Далі розглянуто методи NLP, котрі використовуються інструментами обробки мови, які застосовано до українських п’єс.

Токенізація (сегментація) – це розбиття природного тексту, що надходить на вхід програми на найменші одиниці в реченні (токени). Токенами можна вважати не тільки слова, а й знаки пунктуації. Метою

процесу застосування токенізації є полегшення зчитування вхідних даних програмою. [5]

Стемінг – це процес відокремлення суфіксів та префіксів для отримання основи (кореня) слова. [6]

Лематизація – це процес повернення слова до первинної форми, наявної в словнику. Процес лематизації є складнішим та потребує більше часу, через пошук правильної базової форми слова, з урахуванням морфологічних правил. [6]

Видобування інформації (information extraction) – це процес автоматичного виокремлення необхідної для аналізу структурованої інформації з неструктурованого джерела. У роботі цей метод застосовано для отримання дієслів, котрі мають конструкцію з очевидною статтю. [7]

Розмічування частин мови (Part-of-Speech tagging, POS) – це процес категоризації слів у корпусі за визначеними граматичними правилами до відповідної частини мови. В українській мові це: іменник, прикметник, сполучник тощо. Позначення використовується для визначення граматичної ролі кожного слова в реченні та його морфологічних властивостей. [8]

2. Огляд дослідження The Pudding

Стаття «Film Dialogue» [2] на The Pudding є аналітичним дослідженням гендерних стереотипів у репліках кінофільмів.

Для дослідження зібрано інформацію з приблизно 2000 сценаріїв до кінофільмів різних жанрів та років випуску. Вилучено всі діалоги із цих сценаріїв та проаналізовано за різними критеріями. Одним із них був аналіз за відсотком реплік, промовлених чоловічими та жіночими персонажами.

Ключовим методом є розподіл діалогів за статтю. Співставлено імена персонажів із кожного сценарію до їхніх відповідників на платформі IMDB. Виявлено, який актор чи акторка грали відповідного персонажа. Проведення співставлень необхідне через особливості англійської мови. Вона містить імена з неоднозначною статтю, окрім цього, у тексті трапляються псевдоніми, у такий спосіб уникнено чинник некоректного розподілу. Після проведення аналізу, статистично виявлено, що в багатьох фільмах чоловічі персонажі говорять частіше, ніж жіночі. Це відображає загальну тенденцію до нерівності статей у кіноіндустрії.

Прикладом надано статистичні дані, що базуються на порівняні чоловічих та жіночих персонажів у мультфільмах Disney. У січні 2016-го року виявлено, що чоловіки говорять частіше, ніж жінки («men speak more often than women»)[2] в мультфільмах про принцес. Взято вибірку з 30 мультфільмів. Результатом є, що у 22 з них чоловічі фрази мають перевагу над жіночими.

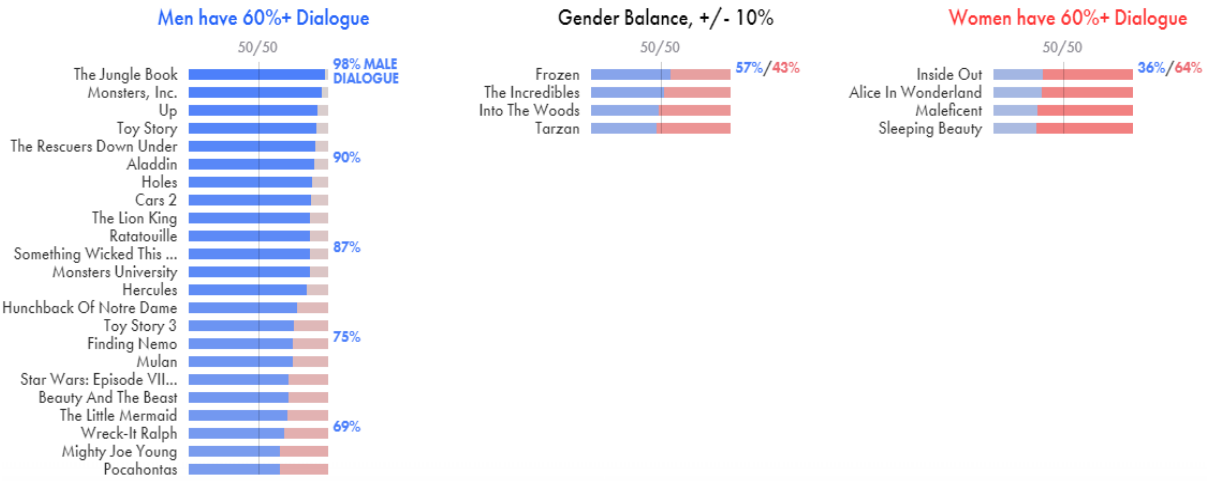

Рисунок 1. – Візуалізація співвідношення персонажів за статтю у мультфільмах Disney.

Окрім цього, виявлено, що лише у 22 % від усіх сценаріїв жіночі репліки переважають. Загальна тенденція – за кількістю реплік у конкретному фільмі, жіночі персонажі частіше перебувають на другому місці, а саме у 34 % випадків. Поміж усіх фільмів, варіації коли 2 з 3х найголовніших ролей займають жіночі персонажі – 18 %. У випадку із чоловічими персонажами, аналогічна ситуація трапляється приблизно у 82 %.

Також проведено аналіз за віком акторів. Продемонстровано, що кількість ролей у жінок, які старші за 40 років, значно зменшується. У чоловіків навпаки – більше вільних ролей для акторів немолодого віку.

У досліджені присутні недоліки. Першим є встановлене обмеження – персонажі в сценаріях співставляються з відповідниками в базі даних IMDB, лише за умови, що сумарна кількість їхніх слів не менше ста. За умови використання такого підходу, втрачається певна кількість персонажів, що впливає на кінцевий результат аналізу.

Окрім цього, для кожного фільму обрано певну версію сценарію. Як наслідок – не враховано потенційне використання версії, яка відрізняється від фінальної. Це також впливає на результат дослідження, оскільки кінцеві версії могли бути зміненими, і кількість реплік відповідно.

Стаття «She Giggle, he Gallops» [3] The Pudding є аналізом чоловічих та жіночих тропів у сценаріях.

Попередньо проведений аналіз за кількістю реплік не є вичерпним для формування висновків щодо зображення чоловічих та жіночих ролей на екрані. Це стало мотивом для проведення дослідження на основі тропів.

Гендерні тропи, як-от «жінка гарна, або чоловіки діють, чоловіки не плачуть» («women are pretty/men act, men don’t cry»)[3] важливі для аналізу не менше за репліки.

Основною методологією в цьому дослідженні є вилучення біграм та частотний аналіз. Біграмами називають пари послідовних слів, у яких першим словом є «він» («he») або «вона» («she»). Застосування біграм обумовлено тим, що в англійській мові, на відміну від української, немає можливості визначити стать персонажа, лише за дієсловом.

Проведено частотний аналіз із найвживанішими біграмами. Результатом першої візуалізації є діаграма вірогідностей. Певні слова частіше належать до «неї», інші до «нього»

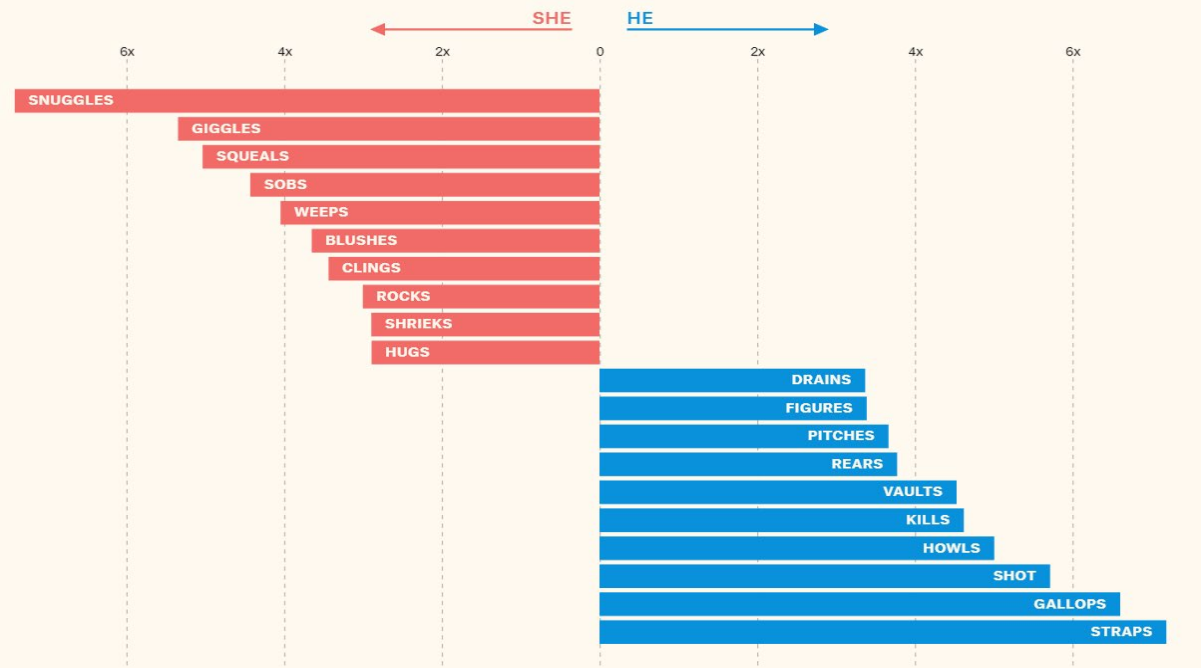

Рисунок 2. – Діграма вірогідностей належності слів до «неї» або до «нього»

Результатом другої біграми є 800 найвживаніших пар «він»/«вона» з дієсловами. Червоним кольором позначено ті, що частіше вживано в парі з жінками, сірим – слова, що вживаються рівноцінно, а синім – відповідні пари з чоловіками.

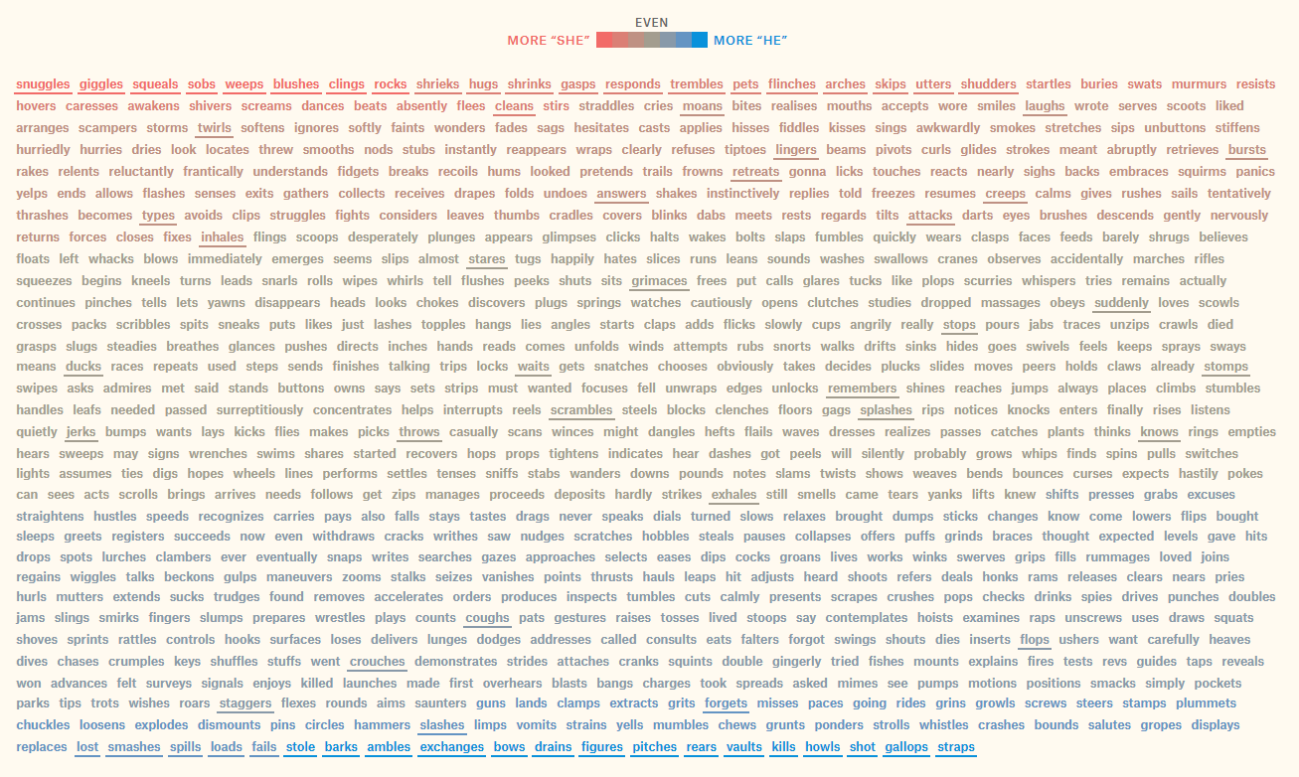

Рисунок 3. – 800 найвживаніших пар «він»/«вона» з дієсловами

Далі проведено аналіз впливу статі автора сценарію на вислови відповідних героїв. Значної різниці між висловами персонажів у фільмах, де автором є чоловік чи жінка, не виявлено. Автори, незалежно від статі, рівноцінно наділяли жіночих персонажів більш романтичними та ніжними висловами, тоді, як чоловіків – мужніми та агресивними. Дослідники припускають, що на результат дослідження вплинуто незначною кількістю жіночих авторів – 15 %, чоловічих, відповідно – 85 %. Висунуто припущення, що за умови однакової кількості авторів для обох статей, жіночі персонажі кінофільмів, набудуть менше стереотипних висловів.

3. Збір та обробка корпусу п’єс

3.1 Збір корпусу п’єс

Збір інформації для корпусу п’єс є важливим етапом дослідження, оскільки від якості даних залежить правильність та об’єктивність отриманих результатів.

Корпус – це структурована збірка текстів або інших видів інформації, організованих у набори даних. [9] У роботі корпус представлено збіркою творів у єдиному форматі.

У контексті обробки природної мови, використання тренувальних корпусів уможливлює створення та вдосконалення моделей обробки людської мови.

Збір здійснено з декількох джерел: назви класичних п’єс [10], пошук яких здійснено через портал бібліотеки української літератури [14]; сучасні п’єси з бібліотек української драматургії [11, 12, 13]. Метою поєднання класичних та сучасних творів є збільшення різноманітності корпусу.

П’єси вивантажено та переведено до формату ’.txt’, оскільки процес застосування інструментів обробки вимагає єдиного формату збірки даних.

Вебсторінки сучасних п’єс містять однакові примірники в різних форматах та мають різні назви файлів. Для вирішення цієї проблеми, розроблено метод для ідентифікації п’єс з однаковим вмістом.

Ідея методу полягає в тому, щоб визначати подібність п’єс на основі вилучених символів із середини тексту кожної з них. Ініціалізовано змінну ‘num_chars_to_read’ та присвоєно їй значення 5000. Це кількість порівнюваних символів, які зчитуються з кожної п’єси та циклічно порівнюються з усіма іншими. Для відстеження послідовності символів залучено ‘SequenceMatcher’ з модуля ‘difflib’. Задано порогове значення подібності ‘similarity_threshold’ – 0.6. Якщо подібність послідовності символів з однієї п’єси перевищує 60 % з іншою, то велика вірогідність

того, що вони мають однаковий контент. Усі знайдені в такий спосіб дублікати – видалено.

3.2 UDPipe REST API

Для проведення аналізу українських п’єс, здійснено попередню обробку корпусу, застосувавши функціональність UDPipe.

UDPipe є інструментом для обробки природної мови, який надає можливість розпізнавання синтаксичних структур української мови через роботу з універсальними залежностями (universal dependencies). Універсальні залежності (УЗ) є стандартною схемою анотації залежностей слів. Залежності вказують на зв’язки між словами в реченні. Кожне слово, може мати одну або більше залежностей до інших слів. УЗ показують, які слова є головними та, які є залежними. Використання залежностей дає можливість встановлювати зв’язок між різними частинами мови: підметами, присудками, додатками, сполучниками тощо.

UDPipe навчений визначати універсальні залежності на основі «золотого морфосинтаксового стандарту». Золотий стандарт – це корпус, що, містить документи українською мовою, де повністю вручну розмічені залежності між словами в декілька шарів. [16] Основним призначенням створення стандарту є тренування автоматичних аналізаторів, як UDPipe для обробки текстів українською мовою.

Процес обробки з використанням UDPipe відбувається в декілька етапів. Процес виконання цих етапів називається пайплайном.

Пайплайн для української моделі охоплює раніше згадані токенізацію, лематизацію, стемінг, позначення частин мови та встановлення залежностей між словами на основі золотого стандарту.

Результатом обробки корпусу текстових документів є збірка файлів CONLL-U формату. Це стандартний формат, котрий анотує дані за

відповідними реченнями та словами (токенами). Проанотовані рядки слів, містяться в полях, розділених табуляціями. [17] Поля сформовано з: унікальних ідентифікаторів токенів у тексті; форми токенів; леми токенів; морфологічної мітки частини мови; морфологічні ознаки токена, які свідчать про рід, відмінок тощо; залежності між токенами та інші.

Обробку проведено з використанням запитів через наявний API в UDPipe. Обробка відбувається в декілька етапів.

Встановлено модуль ‘punkt’ з бібліотеки ‘nltk’, для виявлення меж речень, процес полегшує обробку запитів від UDPipe, запити містять окремо розділені речення.

Задано URL API UDPipe та параметри запиту, що, містять українську модель обробки мови останньої версії – ukrainian-iu-ud-2.10–220711

Ітеративна обробка файлів у корпусі через надсилання POST-запитів до API UDPipe. Відповідь від вебсерверу записується у вихідний файл із розширенням ‘.conllu.txt’.

Використання API від UDPipe не є обов’язковою вимогою для обробки файлів, оскільки ця технологія може бути встановлено на клієнтський комп’ютер. У кожного з цих підходів є свої переваги та недоліки.

Переваги виконання обробки через API запити в тому, що немає необхідності завантажувати інструмент на комп’ютер, цей підхід є доцільним у випадку обмежених ресурсів комп’ютера або за небажанням встановлювати додатковий програмний засіб. Окрім цього, використання API є універсальним рішенням для виконання запитів на обробку з будь-якого пристрою, незалежно від операційної системи. До недоліків підходу можна віднести: залежність від стабільного з’єднання до Інтернет мережі; обмеження API на кількість одночасних запитів, деякі сайти встановлюють його для уникнення перевантаження мережі через велику кількість запитів;

онлайн-підхід є повільнішим за локальне встановлення через затримки в роботі мережі.

Встановлення UDPipe на клієнтський комп’ютер має переваги над недоліками підходу через запити до API, однак є необхідність у слідкуванні за оновленнями програмного засобу для отримання найактуальніших даних.

4. Методи видобування інформації

4.1 Regular expressions for character extraction

Для виконання основної мети роботи – аналізу п’єс на наявність гендерних упереджень, необхідно вилучити персонажів, їхні репліки та слова.

Автоматичні аналізатори мають у своєму інструментарії метод під назвою Named entity recognition (NER). Він використовується для виявлення та класифікації іменованих сутностей у тексті, як-от особи, місця, організації тощо. Цей метод поширений серед багатьох аналізаторів, охопно з UDPipe. Однак його використання для виокремлення персонажів п’єс є складною задачею. По-перше, з’являється необхідність у фільтруванні іменованих сутностей, оскільки метод вилучає не тільки імена а й власні назви. По-друге потрібно реалізувати відслідковування поточної іменованої сутності персонажа, для підрахування відповідних слів і реплік Задача ускладнена тим, що оброблений CONNL-U файл містить не тільки сирий текст, слова якого потрібно рахувати, а й інші ідентифікатори аналізатору UDPipe.

Вирішенням у подоланні цієї незручності є застосування регулярних виразів (regular expression, regex) на сирому тексті формату ‘.txt’. Регулярні вирази – це шаблони, котрі містять правила, за якими, вони знаходять потрібні слова або токени, які відповідають критеріям шаблону.

Кожне певне відображення персонажів у творі потребує відповідного шаблону. Наприклад, для персонажів позначених у тексті на початку рядка єдиною комбінацією великих букв та двокрапкою (‘АНДРІЙ:’) потрібно створити відповідний регулярний вираз (regex). Цей regex представлено так: ‘^(?:[А-ЯЄЇІ]+:)’. У цьому прикладі ‘^’ виконує перевірку чи слово розташоване на початку рядка; ‘?’ вказує на те, що послідовність букв із двокрапкою може бути присутньою в рядку нуль або один разів; ‘[А-ЯЄЇІ]’

позначає поодинокі букви імені в межах українського алфавіту та у верхньому регістрі; ‘+’ виконує операцію знаходження послідовності з одного або з багатьох токенів, що стоять перед ним. За відсутності цього символу, буде збіг лише по першій букві у верхньому регістрі.

Створення та перевірка шаблонів здійснюється через онлайн-сервіс для регулярних виразів [18]. На цьому сервісі надано зручний інтерфейс, де можна прописати regex і текст для тестування та отримати відповідь щодо кількості успішних збігів за шаблоном.

Для кожного поодинокого набору представлення персонажів створено відповідний regex. Утворено відповідну колекцію з регулярних виразів для послідовного добирання шаблону під кожну п’єсу.

Метод виокремлення героїв циклічно проходить по корпусу п’єс у форматі ‘.txt’, адже саме в такому форматі є можливість визначати слова на початку рядка, на відміну від обробленого формату CONLL-U, і по черзі використовує регулярні вирази для кожної п’єси, поки їх не буде оброблено.

Для отримання реплік кожного персонажа, необхідно порахувати кількість збігів для кожного героя за регулярними виразами. Створено словник ‘defaultdict’ з модуля ‘collections’ зі значеннями за замовчуванням. За допомогою ‘defaultdict’ створено словник ‘character_count’, який містить лічильник для кількості реплік. Якщо під час перебирання файлів виявлено збіг із регулярним виразом, разом із вилученням імені персонажа оновлюється лічильник для його реплік. Словник створює для кожного персонажа лічильник лише один раз і тому зі збігами дублікати не створюватимуться.

Підрахунок слів виконано схожим чином, як із репліками, тобто лічильник оновлюється в словнику. Логіка визначення слів певного персонажа полягає у розподілені всіх слів у рядку, котрі перебувають після регулярного виразу. Коли відбувається збіг за regex, послідовність слів після регулярного виразу розбивається на поодинокі токени і за кожен

відбувається збільшення відповідного лічильника для підрахунку слів у словнику ‘character_count’.

4.2 Визначення статі персонажів

Одним зі способів визначення статі персонажів є виокремлення раніше згаданих іменованих сутностей з іменами, в поєднанні з ідентифікатором статі, проставленим UDPipe. Однак, UDPipe схильний до некоректної ідентифікації статі персонажів. Наприклад, для імені «Настя», встановлюється ідентифікатор зі значенням нейтральний (‘Gender=Neut’). У цьому випадку, гарантовані помилки не тільки для наявних в тексті псевдонімів, а й для популярних імен. Окрім цього, багато текстів містять імена сформовані послідовністю букв із пробілами або складені з імен та прізвищ. UDPipe обробляє їх як поодинокі букви та два поодинокі слова відповідно.

З огляду на те, що імена попередньо вилучені за допомогою регулярних виразів, варто застосувати модель спеціально натреновану на корпусі українських імен для визначення статі. Модель має працювати таким чином, щоб зчитавши ім’я – присвоїти йому відповідну стать із деяким відсотком впевненості.

Такі моделі є в багатьох аналізаторах, а також через API-сервер, з надсиланням імен в POST-запитах. Однак, усі вони підтримають лише англійську мову, бо натреновані на корпусах з англійськими іменами. Українського відповідника наразі немає. Відповідно – проведено розподіл ідентифікаторів між чоловічою та жіночою статтю вручну.

4.3 Виокремлення дієслів за статтю

Для виокремлення дієслів за статтю, розроблено метод, який циклічно проходить по обробленим CONLL-U файлам та вилучає відповідні поля, що мають зв’язки з дієсловами. Такими полями є: дієслово позначене тегом ‘VERB’, базова форма дієслова (лема) та відповідне поле із значенням статі ‘Gender=Fem’ або ‘Gender=Masc’. Цей метод вилучає дієслова, котрі однозначно трактуються за статтю. Наприклад, слова «зробив», «сказав», «приніс», однозначно вказують на особу чоловічої статі, а слова: «зробила», «сказала», «принесла», відповідно – жіночої.

4.4 Виокремлення авторів п’єс

Визначення авторів п’єс є важливим етапом дослідження, оскільки за результатами цього процесу зроблено відповідні висновки щодо розподілу реплік та слів персонажів за статтю.

У більшості випадків, текст п’єси містить контактну інформацію про автора разом з його іменем та прізвищем.

Метод базується на використанні інструменту визначення іменованих сутностей (NER).

Здійснено циклічний обхід п’єси з вилученням перших двох іменованих сутностей для кожного тексту. Реалізація обумовлена тим, що тексти п’єс містять контактну інформацію про автора на початку файлів. Проаналізовано вміст отриманих результатів для перевірки наявності коректної інформації. Випадки, коли поміж вилучених іменованих сутностей присутня інформація, яка не стосується автора – виправлено. У разі відсутності автора – присвоєно значення «no author». Ідентифікатори статі авторів встановлено вручну.

4.5 Вилучення нецензурної лексики у п’єсах.

Аналіз наявності нецензурної лексики є доповненням до основного дослідження, яке вказує на додаткові деталі, котрі використано для отримання фінальних результатів.

Реалізація цього методу першочергово полягає у визначенні слів, які ідентифікуються як нецензурні. Використано корпус з українськими ненормативними словами різних варіацій, сумарна кількість яких – 7129 слів.[19]

Метод циклічно проходить по корпусу сирих п’єс, розділяє текст на окремі рядки та починає перевірку на збіги з корпусом із образливою лексикою. Для того, щоб збіги віднаходилися правильно тільки для поодиноких цілих слів, уникаючи випадків коли нецензурне слово є підстрічкою основного слова, використано відповідний regex:

```
pattern = r'\b(' + '|'.join(obscene_words) + r')\b'
```

5. Аналіз отриманих результатів

5.1 Аналіз персонажів різної статі за кількістю реплік та слів

Сумарна кількість українських п’єс, які взято для аналізу – 550. Значна кількість з них має погано структурований текст. Як наслідок – успішне виокремлення персонажів та їхніх реплік із використанням регулярних виразів, наявне лише у 457 п’єсах.

Результат виокремлених персонажів та відповідних реплік із словами:

- сумарна кількість чоловічих персонажів – 2611;
- сумарна кількість жіночих персонажів – 1682;
- сумарна кількість чоловічих реплік – 107990;
- сумарна кількість жіночих реплік – 83336
- сумарна кількість чоловічих слів у репліках – 1466742;
- сумарна кількість жіночих слів у репліках – 1088997;

Видно, що чоловічих персонажів набагато більше ніж жіночих, а саме – 60,82% чоловіків, проти 39,18% жінок. Згідно з цим, відповідні показники реплік та слів мають значну різницю. Слід зазначити, що пряме порівняння показників реплік та слів не є інформативним, оскільки співвідношення персонажів не рівне. Для отримання точнішого обчислення, підраховано середні показники для цих параметрів:

- середня кількість реплік на чоловічого персонажа – 41.36
- середня кількість реплік на жіночого персонажа – 49.55
- середня кількість слів на чоловічого персонажа – 561.75
- середня кількість слів на жіночого персонажа – 647.44

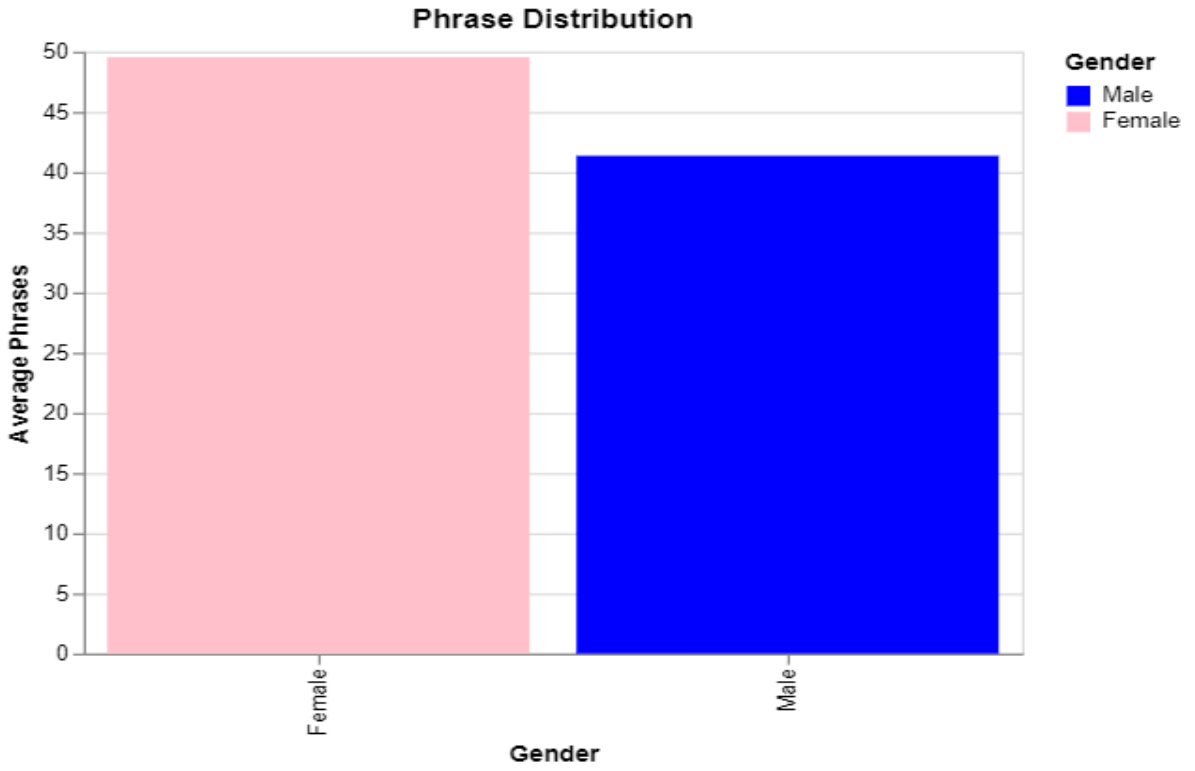

Рисунок 4. – Середній розподіл реплік за статтю

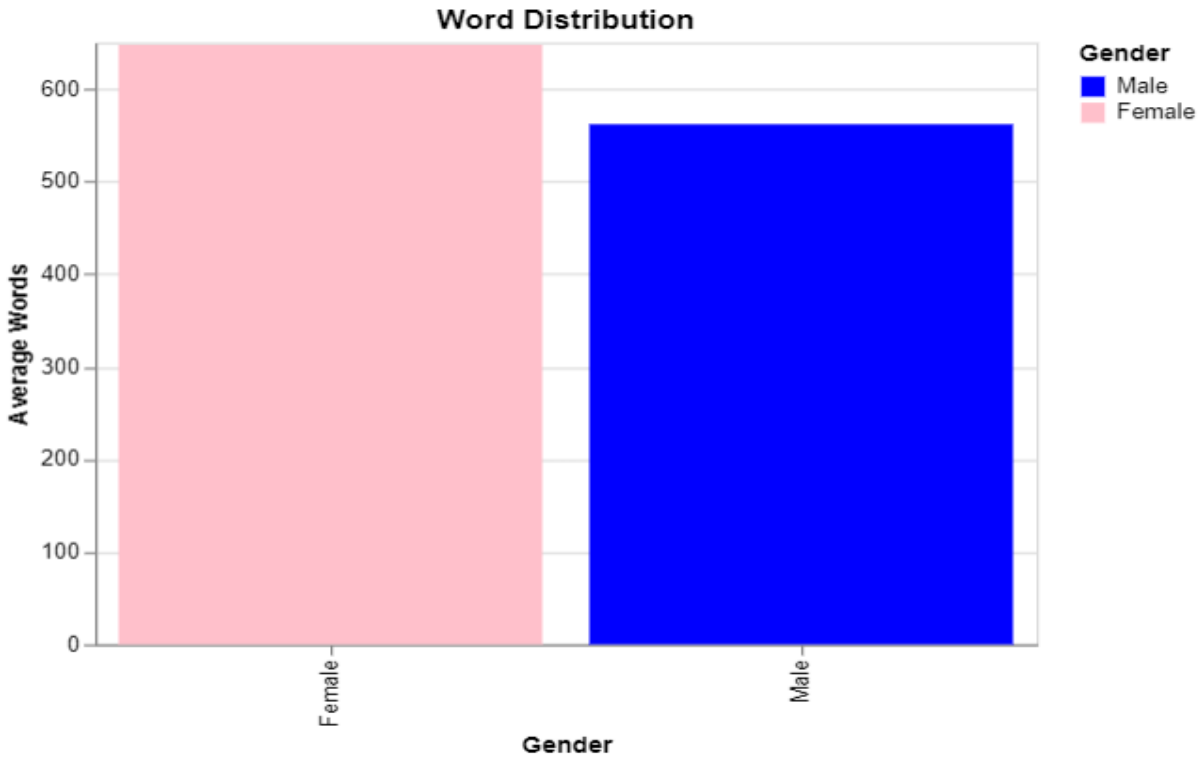

Рисунок 5. – Середній розподіл слів за статтю

За цими результатами можна стверджувати, попри те, що чоловічі персонажі перевищують кількісно жіночих, у середньому жіночі персонажі промовляють на 19,79% більше реплік, а кількість слів у цих репліках у середньому на 15,26% більше, тобто репліки довші.

Зважаючи на великий розрив у кількості персонажів, вирішено провести аналіз за співвідношенням авторів чоловічої та жіночої статі. Успішно виокремлено авторів з 421 п’єси, інші позначено – «no author»:

- кількість п’єс, написаних чоловіками – 216
- кількість п’єс, написаних жінками – 205
- кількість п’єс з невідомим автором – 36

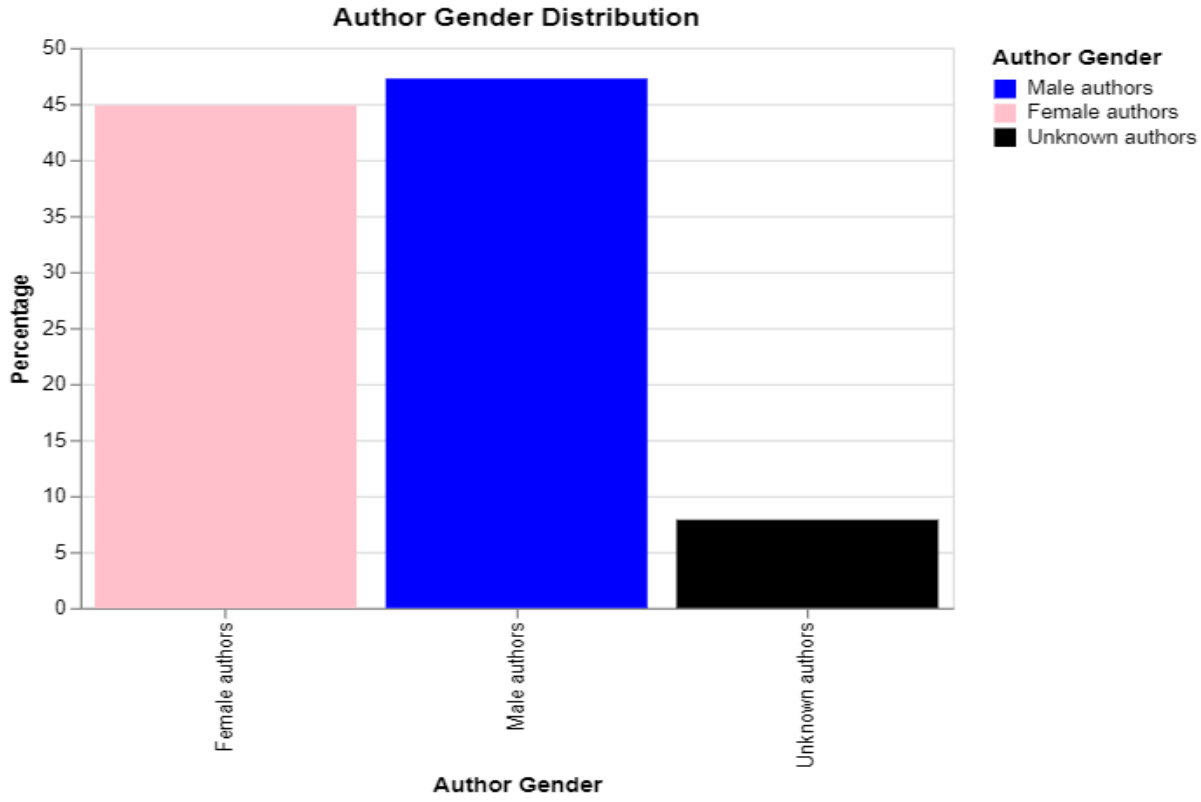

Рисунок 6. – Розподіл авторів за статтю

З діаграми видно, що різниця між статтю авторів п’єс незначна (п’єс, написаних чоловіками на 2,4% більше). З цього можна зробити висновок,

що: і чоловіки і жінки схильні до використання більшої кількості чоловічих персонажів у своїх творах.

Окрім цього, проведено аналіз за відсотковим співвідношенням між чоловічими та жіночими репліками й словами для кожної поодинокої п’єси.

Підраховано сумарну кількість чоловічих і жіночих реплік та слів, і визначено відсоткове співвідношення для кожної п’єси. Відсортовано за спаданням відсоткового показника для чоловіків та за зростанням цього показника для жінок. Округлено результати для групування по стовпцях.

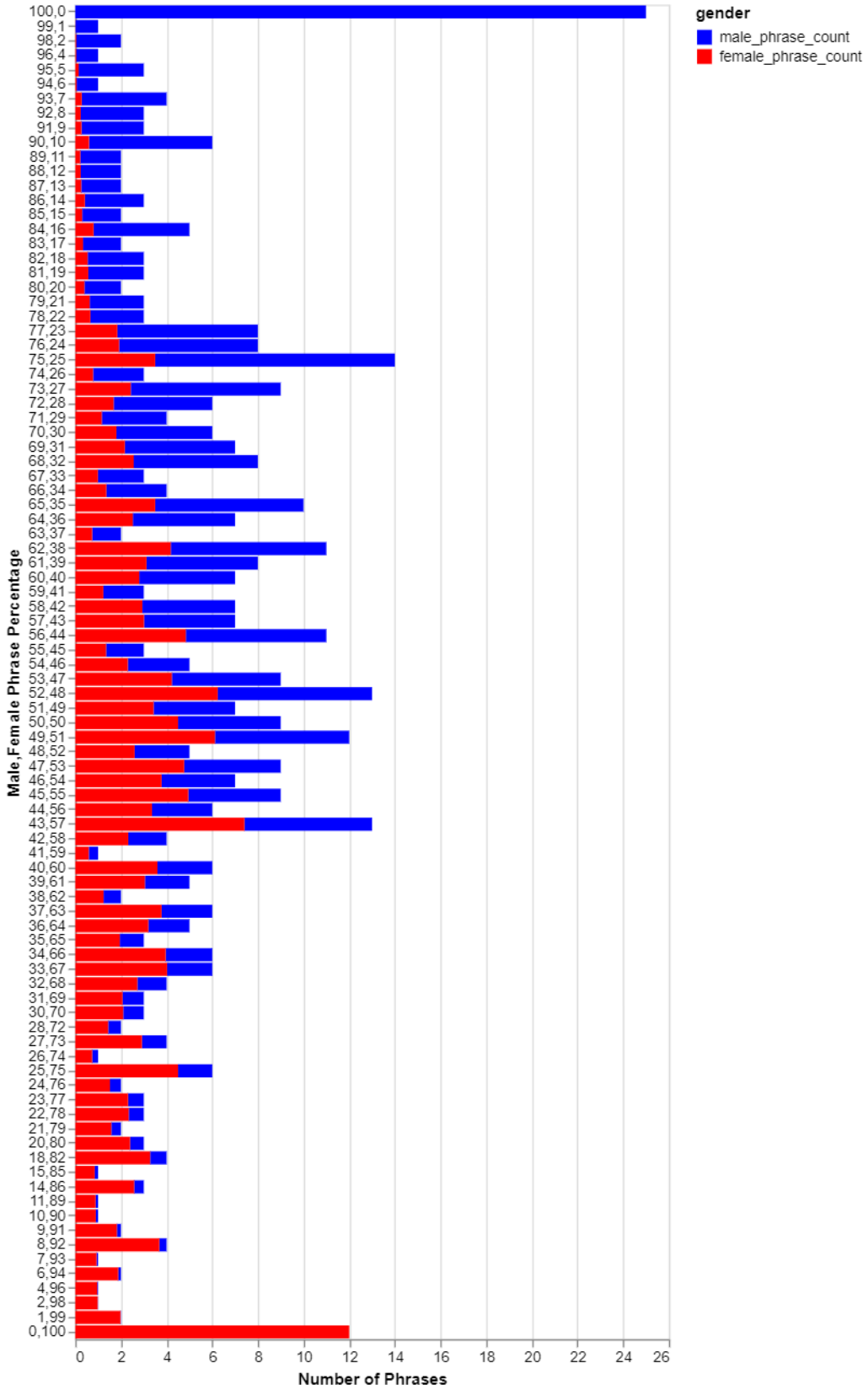

Рисунок 7. – Співвідношення чоловічих та жіночих реплік за п’єсами

На діаграмі представлено співвідношення у формі стовпців. Кожен стовпець налічує певну кількість п’єс з однаковим відсотковим співвідношенням чоловічих та жіночих реплік. Вісь ординат відповідає за це співвідношення, де перше число – це відсоток чоловічих реплік, а друге, через кому, відсоток жіночих. Вісь абсцис вказує на кількість п’єс у конкретному відсотковому співвідношенні (чим вищий стовпець, тим більше п’єс з цим співвідношенням).

Кількість п’єс з переважним відсотком чоловічих реплік – 271 (59,3% від усіх п’єс), кількість п’єс з майже однаковим співвідношенням (округленим до 50%) – 9 (1,97% від усіх п’єс), кількість п’єс з переважною більшістю жіночих реплік – 177 (38,7% від усіх п’єс).

Друга діаграма має аналогічну реалізацію, але демонструє співвідношення між чоловічими та жіночими словами, замість реплік.

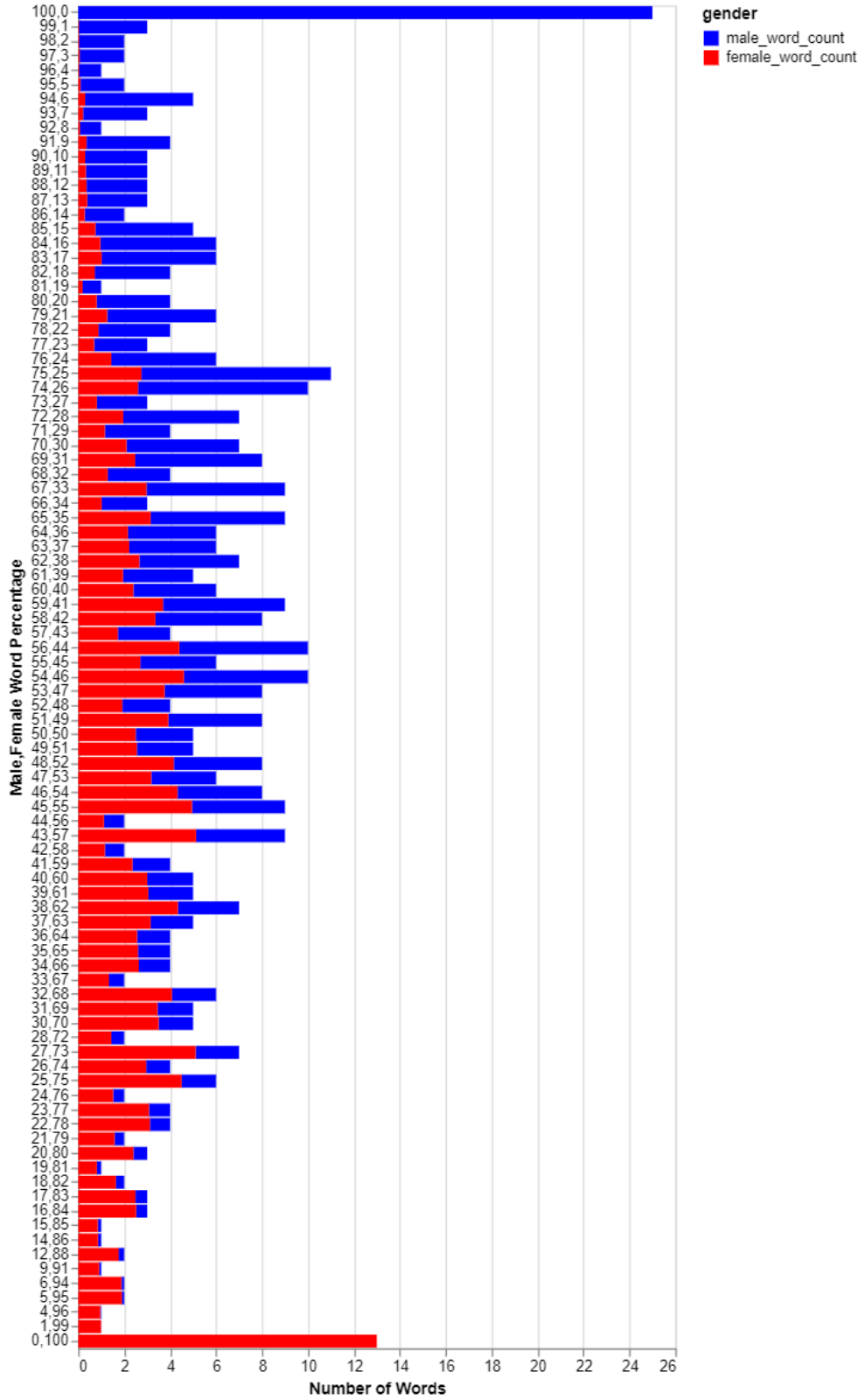

Рисунок 8. – Співвідношення чоловічих та жіночих слів за п’єсою

Кількість п’єс з переважним відсотком чоловічих слів – 279 (61% від усіх п’єс), кількість п’єс з майже однаковим співвідношенням (округленим до 50%) – 5 (1% від усіх п’єс), кількість п’єс з переважною більшістю жіночих реплік – 173 (37,9% від усіх п’єс).

Також проведено аналіз за творами видатних авторів, до яких належать: Іван Франко, Іван Котляревський, Леся Українка та інші. Сумарна кількість цих п’єс – 65/457.

Кількість п’єс, написаних чоловіками – 45 (69,23%), проти 20, написаних жінками (30,77%).

Результати проведеного частотного аналізу за класичними п’єсами:

- сумарна кількість чоловічих персонажів – 620 (65%)
- сумарна кількість жіночих персонажів – 331 (34,8%)
- сумарна кількість чоловічих реплік – 20995 (60,3%)
- сумарна кількість жіночих реплік – 13828 (39,7%)
- сумарна кількість чоловічих слів – 422087 (60,6%)
- сумарна кількість жіночих слів – 274163 (39,4%)
- середня кількість реплік на чоловічого персонажа – 33,86
- середня кількість реплік на жіночого персонажа – 41,78
- середня кількість слів на чоловічого персонажа – 680,79
- середня кількість слів на жіночого персонажа – 828,29

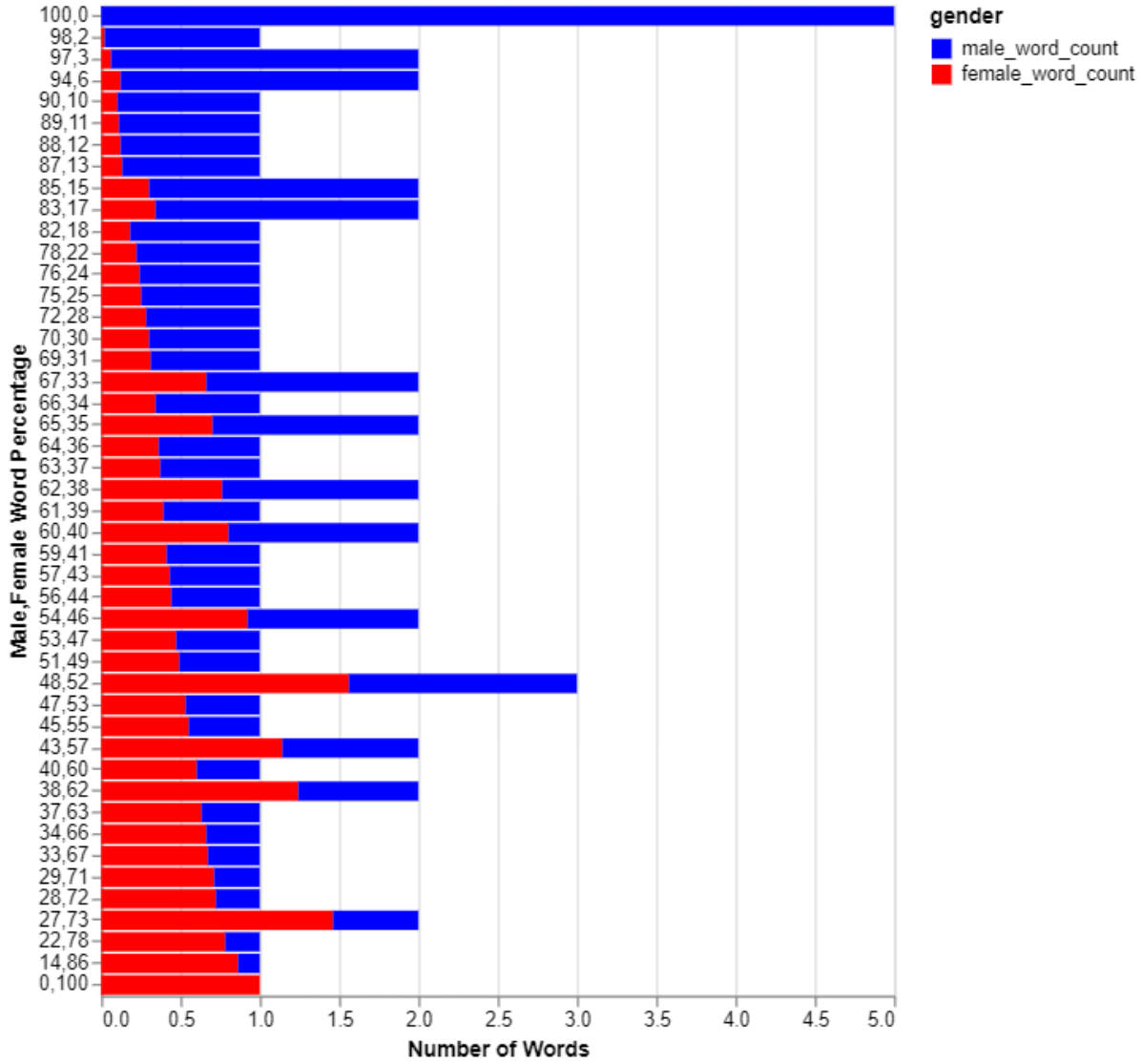

Рисунок 9. – Співвідношення чоловічих та жіночих слів за п’єсою (класичні п’єси)

Кількість п’єс з переважним відсотком чоловічих слів – 44 (67,7% від усіх п’єс), кількість п’єс з переважною більшістю жіночих реплік – 21 (32,4% від усіх п’єс).

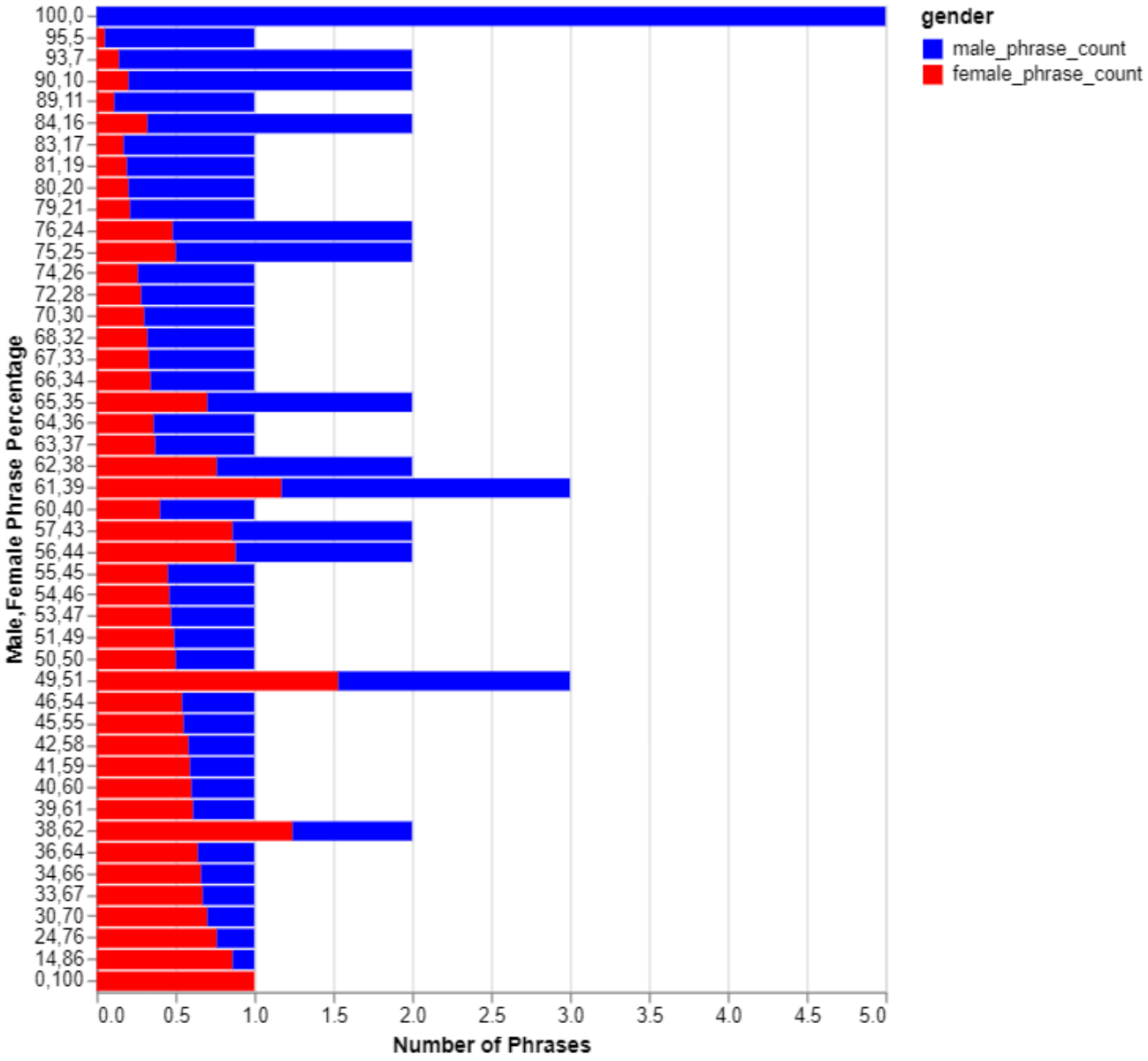

Рисунок 10. – Співвідношення чоловічих та жіночих реплік за п’єсою (класичні п’єси)

Кількість п’єс з переважним відсотком чоловічих реплік – 46 (70,7% від усіх п’єс), кількість п’єс з переважною більшістю жіночих реплік – 19 (29,2% від усіх п’єс). До цього переліку входять такі відомі п’єси:

- «Сто тисяч» (автор – Іван Карпенко-Карий; кількість чоловічих персонажів – 8; кількість жіночих персонажів – 2; співвідношення реплік – 90% чоловічі, 10% жіночі; співвідношення слів – 94% чоловічі, 6% – жіночі)

- «Мартин Боруля» (автор – Іван Карпенко-Карий; кількість чоловічих персонажів – 11; кількість жіночих персонажів – 2; співвідношення реплік – 82% чоловічі, 18% жіночі; співвідношення слів – 78% чоловічі, 22% – жіночі)
- «Наталка Полтавка» (автор – Іван Котляревський; кількість чоловічих персонажів – 4; кількість жіночих персонажів – 6; співвідношення реплік – 70% чоловічі, 30% жіночі; співвідношення слів – 63% чоловічі, 37% – жіночі)
- «Лісова пісня» (автор – Леся Українка; кількість чоловічих персонажів – 10; кількість жіночих персонажів – 7; співвідношення реплік – 46% чоловічі, 54% жіночі; співвідношення слів – 38% чоловічі, 62% – жіночі)
- «Мина Мазайло» (автор – Микола Куліш; кількість чоловічих персонажів – 3; кількість жіночих персонажів – 7; співвідношення реплік – 38% чоловічі, 62% жіночі; співвідношення слів – 29% чоловічі, 71% – жіночі)

Можна зробити висновок, що існує тенденція на переважну більшість чоловічих персонажів і відповідних їхніх показників. Слід зазначити, що ця різниця в класичних п’єсах обумовлена значно більшою кількістю п’єс, написаних чоловіками. Однак, якщо виокремити п’єси з сучасними авторами від загальної кількості, отримано невелику перевагу за кількістю п’єс, авторами яких є жінки, а саме – 185 жіночих п’єс проти 171 чоловічої. Віднявши показники всіх персонажів, реплік та слів для класичних п’єс, отримано результат, що тренд на домінування чоловічих персонажів за кількістю зберігається у решті корпусу.

5.2 Аналіз жіночих та чоловічих персонажів за дієсловами

Вилучено дієслова за статтю, їхні леми та відповідні ідентифікатори статі. Сумарна кількість таких дієслів у корпусі для 550 п’єс становить приблизно 127 тисяч.

Реалізовано частотний аналіз за всіма словами та отримано вибірку найпопулярніших слів.

Найпопулярнішими дієсловами для чоловіків є: «сказав» (використано 1877 разів); «хотів» (використано 1428 разів); «міг» (використано 993 рази). Найпопулярнішими дієсловами для жінок є: «сказала» (використано 1134 рази); «хотіла» (використано 892 рази); «бачила» (використано 747 разів).

З цього аналізу видно, що найпопулярніші слова мають перетин між чоловічими та жіночими персонажами, однак він не дає розуміння про стереотипну діяльність відповідних статей.

Метод оновлено шляхом прибирання дієслів, котрі мають однакове значення, але одночасно належать і чоловічим, і жіночим персонажам. Для цього використано базові форми слів (леми), завдяки яким однакові дієслова з різними ідентифікаторами статі видалено. Це дозволило зменшити сумарну кількість дієслів до 11 тисяч.

Після проведення повторного частотного аналізу з унікальними для обох статей дієсловами отримано оновлений результат. Слова з чоловічим ідентифікатором: «розлютився», «стерпів», «відрікся», «вистрелив», а з жіночим: «змарніла», «надумала», «пристала», «відшила», «виколисала» та «покохала».

У цьому аналізі чітко видно як вживані дієслова репрезентують відповідну стереотипну поведінку чоловічих та жіночих персонажів.

5.3 Аналіз п’єс за наявністю нецензурної лексики.

Поміж 550 п’єс, нецензурна лексика наявна в 378 файлах (68% від усієї кількості). Сумарна кількість нецензурних слів по цих текстах – 4730.

Проведено аналіз за статтю авторів, у творах яких наявна нецензурна лексика:

- кількість п’єс, написаних чоловіками – 199 (52%)
- кількість п’єс, написаних жінками – 148 (38,9%)
- кількість п’єс з невідомим автором – 33 (8,7%)

Вибірка з 10 п’єс із найбільшими показниками за нецензурною лексикою демонструє, що 5 з них написані чоловіками, 3 – жінками і 2 п’єси мають невідомого автора. Авторство твору з найбільшою кількістю ненормативної лексики (159 слів на п’єсу) належить жінці.

Проведено аналіз за наявністю нецензурної лексики у класичних творах. Виявлено, що попри багате різноманіття образливого лексикону в сучасних п’єсах, класичні твори демонструють більш стримані вислови та їхню меншу варіативність. Визначено, що переважна кількість авторів класичних п’єс часто вживають образливе слово по відношенню до євреїв. Це обумовлено культурними особливостями того часу.

Висновки

У роботі розглянуто сучасні підходи обробки природної мови для уможливлення автоматичного аналізу п’єс. Попри те, що використання інструментів аналізу, як-от UDPipe, не є безпомилковим, видно, що точність аналізу висока й технічні можливості цього інструменту оновлюють для моделі української мови. Це свідчить про те, що галузь має попит серед розробників та науковців в українській мережі Інтернет.

Продемонстровано методологію використання засобів обробки для вилучення текстової інформації пов’язаної з чоловічими та жіночими персонажами.

Позначено проблематику з відсутністю навчених моделей для встановлення статі за українськими іменами та запропоновано її вирішення.

Проаналізовано корпус сучасних та класичних п’єс на наявність стереотипів, пов’язаних зі статтю героїв. Аналіз охоплює порівняння співвідношення авторів п’єс, чоловічих та жіночих персонажів, їхніх реплік та слів і середніх значень цих показників. Попри те, що авторство текстів майже однаково розподілено між чоловіками та жінками, кількісна перевага надається чоловічим образам. Після порівняння середніх показників відповідних персонажів, отримано результат, який вказує на більшу кількість реплік для жінок і на те, що в середньому їхні репліки довші.

Проведено порівняння співвідношень для сучасних і класичних п’єс. Виявлено, що раніше було більше авторів чоловіків, що може бути одним із факторів значущої різниці в кількості чоловічих та жіночих персонажів. Однак, у сучасних п’єсах, показник авторства за статтю майже рівний, але тенденція зберігається.

Здійснено порівняння вживання дієслів, що можуть бути ідентифікованими за статтю. Продемонстровано наявність стереотипної поведінки в текстах на основі вживаних дієслів для чоловічих та жіночих

персонажів. Першочергово зроблено частотний аналіз усіх дієслів, де з’ясовано, що найпопулярніші дієслова мають перетин між статями. Однак, для глибшого розуміння стереотипної поведінки, метод удосконалено шляхом вилучення дієслів, які мають однакове значення, але використовуються, як чоловічими, так і жіночими персонажами. Застосування цього методу дозволило виділити дієслова, які є унікальними для кожної статі та чітко демонструють стереотипну поведінку персонажів у текстах.

Аналіз нецензурної лексики вказав на великий показник поширення серед українських п’єс. Виявлено, що незалежно від статі автора, вживання ненормативних слів поміж п’єс досить поширене. Також визначено, що наявність варіативної грубої лексики залежить від сучасності твору. Це свідчить про те, що сьогодні українська культура більше толерує використання табуйованих висловів, ніж у минулому.

Варто зазначити, що запропоноване в цій роботі використання методології для вирішення задач пов’язаних з обробкою живої мови, не обмежується лише українськими п’єсами, і за наявності достатнього обсягу корпусу з текстами українською мовою, аналіз може бути перероблено на інші тематики.

Список літератури

1. Словник NLP [Електронний ресурс] // medium.com. – 2021. – Режим доступу до ресурсу: https://medium.com/stinopys/%D1%81%D0%BB%D0%BE%D0%B2%D0%BD%D0%B8%D0%BA-nlp-b0fab1027551.

2. Film Dialogue from 2,000 screenplays, Broken Down by Gender and Age [Електронний ресурс] // The Pudding. – 2016. – Режим доступу до ресурсу: https://pudding.cool/2017/03/film-dialogue/.

3. She Giggles, He Gallops Analyzing gender tropes in film with screen direction from 2,000 scripts. [Електронний ресурс] // The Pudding. – 2017. – Режим доступу до ресурсу: https://pudding.cool/2017/08/screen-direction/.

4. What is natural language processing (NLP)? [Електронний ресурс] // IBM – Режим доступу до ресурсу: https://www.ibm.com/topics/natural-language-processing.

5. NLP: Tokenization, Stemming, Lemmatization and Part of Speech Tagging [Електронний ресурс] // medium.com. – 2021. – Режим доступу до ресурсу: https://medium.com/mlearning-ai/nlp-tokenization-stemming-lemmatization-and-part-of-speech-tagging-9088ac068768.

6. Stemming and lemmatization [Електронний ресурс] // stanford.edu – Режим доступу до ресурсу: https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html.

7. 8 NLP Techniques to Extract Information [Електронний ресурс] // analyticssteps. – 2022. – Режим доступу до ресурсу: https://www.analyticssteps.com/blogs/nlp-techniques-extract-information.

8. Part Of Speech Tagging for Beginners [Електронний ресурс] // towardsdatascience.com. – 2020. – Режим доступу до ресурсу: https://towardsdatascience.com/part-of-speech-tagging-for-beginners-3a0754b2ebba.

9. The Challenge of Building Corpus for NLP Libraries [Електронний ресурс] // defined.ai. – 2020. – Режим доступу до ресурсу: https://www.defined.ai/blog/the-challenge-of-building-corpus-for-nlp-libraries/#:~:text=What%20is%20a%20Corpus%20in,of%20the%20language%20or%20dialect.

10. Категорія: Українські п'єси [Електронний ресурс] // wikipedia. – 2022. – Режим доступу до ресурсу: https://uk.wikipedia.org/wiki/%D0%9A%D0%B0%D1%82%D0%B5%D0%B3%D0%BE%D1%80%D1%96%D1%8F:%D0%A3%D0%BA%D1%80%D0%B0%D1%97%D0%BD%D1%81%D1%8C%D0%BA%D1%96_%D0%BF'%D1%94%D1%81%D0%B8.

11. Бібліотека української драматургії [Електронний ресурс] – Режим доступу до ресурсу: https://ukrdramalib.com.ua/.

12. Портал сучасної української драматургії [Електронний ресурс] – Режим доступу до ресурсу: https://ukrdramahub.org.ua/search-plays.

13. Ukrainian drama translations [Електронний ресурс] – Режим доступу до ресурсу: https://ukrdrama.ui.org.ua/ua/poshuk-pyes.

14. Бібліотека української літератури [Електронний ресурс] – Режим доступу до ресурсу: https://www.ukrlib.com.ua/books/.

15. UDPipe Natural Language Processing – Text Annotation [Електронний ресурс]. – 2023. – Режим доступу до ресурсу: https://cran.r-project.org/web/packages/udpipe/vignettes/udpipe-annotation.html.

16. золотий морфосинтаксовий стандарт [Електронний ресурс] // mova.institute – Режим доступу до ресурсу: https://mova.institute/%D0%B7%D0%BE%D0%BB%D0%BE%D1%82%D0%B8%D0%B9_%D1%81%D1%82%D0%B0%D0%BD%D0%B4%D0%B0%D1%80%D1%82.

17. CoNLL-U Format [Електронний ресурс] – Режим доступу до ресурсу: https://urduhack.akkefa.com/en/stable/reference/conll.html#:~:text=CONLL%2DU%20is%20a%20standard,separated%20by%20single%20tab%20characters.

18. regex [Електронний ресурс] – Режим доступу до ресурсу: https://regex101.com/.

19. obscene_dictionary.txt [Електронний ресурс] // github.com. – 2019. – Режим доступу до ресурсу: https://github.com/saganoren/obscene-ukr/blob/master/obscene_dictionary.txt.