ШИФР «ТЕКСТОВІ КОРПУСИ»

РОЗРОБКА РОЗМІЧЕНОГО КОРПУСУ КРИМІНАЛЬНО-ЗНАЧУЩИХ ТЕКСТІВ

ЗМІСТ

ПЕРЕЛІК ПОЗНАЧЕНЬ ТА СКОРОЧЕНЬ

СППР - система підтримки прийняття рішень

КТ – комп’ютерна томографія

ПММ – приховані марківські моделі

LSA – Latent Semantic Analysis

ВСТУП

Слова, що описують кримінальну діяльність, мають свої характерні ознаки та можуть бути доволі специфічними, тому вони і є індикативною признакою, за якою здійснюється відбір документів, призначених для подальшої аналітичної обробки. Наступні словосполучення досить зрозумілі: підписка про невиїзд, відбувати покарання, вогнепальна зброя, речовий доказ і тому подіне. Але іноді бувають менш поширені та знайомі сполучення слів, які у свою чергу є більш ефективними у роботі з пошуку кримінально значущої інформації, наприклад «прес-хата». Кожне із наведених словосполучень має своє значення та свої асоціації у колах працівників, що пов’язані з правоохоронною діяльністю, саме тому їх наявність у тексті потребує детального дослідження самого тексту.

До цього часу ще не створено жодного корпусу кримінально значущих текстів, тому його розробка є актуальним завданням на сьогодні, що у подальшому спростить розуміння іншомовних текстів, спеціалізованих на кримінальній діяльності.

Метою роботи є створення розміченого корпусу кримінально-значущих текстів та подальше його опрацювання у вигляді створення словнику кримінально-окрашених слів на основі даного корпусу. Для досягнення даної мети будуть виділені методи створення корпусів, розроблено системи семантичної розмітки кримінально значущих текстів, розроблено програмне забезпечення автоматичної морфологічної розмітки корпусу та програмний додаток автоматизації семантичної розмітки корпусу, досліджені існуючі типи корпусів та види їх анотації.

Теоретичною основою для даної роботи стали методи і концепції в працях Джорджа Ципфа «Вибрані дослідження принципу відносної частоти в мові» та Карен Спарк Джонс «Аналогія між машинним перекладом та пошуком в бібліотеці».

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ

1.1 Аналіз існуючих підходів та методів корпусної лінгвістики

Корпусна лінгвістика є одним із найбільш перспективних напрямків сучасного теоретичного і прикладного мовознавства. Ця відносно нова лінгвістична галузь розпочала своє активне становлення у 60-х роках XX століття у зв'язку із інтенсивним розвитком комп'ютерних технологій. Сам термін "корпусна лінгвістика" міцно увійшов до наукового вжитку лише в останні десятиліття XX століття з публікацією у 1983 році збірника наукових праць «Corpus Linguistics: Recent Developments in the use of Computer Corpora in English Language Research». За матеріалами конференції ICAME «Conference On The Use of Computer Corpora in English Language Research». Звичайно, застосування комп'ютерів та спеціального програмного забезпечення суттєво змінило спосіб дослідження мови та значно полегшало роботу по збору лінгвістичних даних. Без зусиль, лише за декілька секунд, стало можливим здійснювати пошук у багатомільйонних текстових масивах (лінгвістичних корпусах), будувати конкорданс для будь-якого слова, одержувати дані про частоту словоформ, лексем, граматичних категорій, синтаксичних конструкцій, відстежувати зміни у частоті і контексті мовної одиниці у різні хронологічні періоди і таке інше.

Сьогодні дані корпусів масштабно використовуються в лексикографії, стилістиці, судовій лінгвістиці, лінгвістичній варіантології, перекладознавстві, соціолінгвістиці, методиці навчання і вивчення іноземної мови та в багатьох інших лінгвістичних дослідженнях.

Корпусна лінгвістика як галузь прикладного мовознавства займається визначенням загальних принципів побудови, обробки та експлуатації даних лінгвістичних корпусів(корпусів текстів) із використанням сучасних комп'ютерних технологій, розробленням методики збору реальних мовних явищ - писемних та усних текстів, а також способів їх збереження та аналізу. Під корпусом текстів розуміється значення за обсягом, представлений в

електронному вигляді, уніфікований, структурований, розмічений, філологічно компетентний масив мовних даних, створений для вирішення конкретних лінгвістичних завдань.

Головне завдання корпусної лінгвістики вбачається у повному й системному відображенні змістовного спілкування мовою. Важливою особливістю цього інформаційно-семіотичного напряму лінгвістичних досліджень є підхід до розгляду прикладних проблем лінгвістики конкретно в комунікативних процесах. При цьому в центрі уваги виявляється не мова як система, і не проблема її формалізації, а процес змістовного спілкування мовою, і по можливості точний її опис, який може бути використаний для розв'язання науково-технічних завдань інформатики. [2]

Застосування комп'ютерів миттєво обробити величезний обсяг мовного матеріалу і відібрати всі можливі у конкретному корпусі приклади вживання необхідних для аналізу одиниць. У розпорядження лінгвіста надаються об'єктивні кількісні дані, забезпечуючи досягнення більш ґрунтовних та переконливих висновків. Незважаючи на значні досягнення та популярність, статус корпусної лінгвістики у сучасній мовознавчій парадигмі ще не є однозначно вивченим дослідженням мовного матеріалу, на противагу таким традиційним галузям лінгвістики, як фонетика, синтаксис, семантика чи граматика. Більш того, корпусні методи можуть використовуватися для вивчення мовних одиниць будь-якого мовного рівня. Наприклад, синтаксичні явища можливо дослідити як із застосуванням корпусних методик, так і без них, так само існують корпусні і некорпусні семантичні дослідження.[1]

1.2 Огляд існуючих типів та видів корпусів

Сьогоднішня корпусна лінгвістика – це гетерогенна область дослідження мови, всередині якої виокремлюються окремі піднапрями, що різняться підходами до конструкції, експлуатації корпусів та аналізу корпусних даних. В основі виділення цих під напрямів знаходяться такі параметри :

–формат представлення текстів у корпусі (mode of communication). Корпуси можуть містити тексти, представлені в усній чи писемній формі. В

залежності від форми представлення текстів виділяють корпуси усного мовлення, корпуси писемного мовлення та корпуси змішаного типу. Нові типи корпусів, такі як мультимедійні корпуси та корпуси кінетичного мовлення, реєструють також і паралінгвістичні засоби, які супроводжують спілкування або є безпосереднім способом спілкування;

–корпуснобазовані (corpus-based) vs. корпуснокеровані (corpus-driven) дослідження. У корпуснобазованих дослідженнях дані корпусу використовуються для доведення, спростування чи уточнення визнаної на певному етапі розвитку наукової думки теорії чи гіпотези. Цей тип корпусних розвідок трактує корпусну лінгвістику як метод дослідження мови. Натомість корпуснокерована лінгвістика відмовляється від визнання корпусної лінгвістики як методу і стверджує, що корпус сам по собі є єдиним джерелом гіпотез про мову та втілює свою теорію мови;

–використання анотованих(annotated)/ неанотованих (unannotated) корпусів. Головною відмінністю сучасного корпусу є наявність анотації, тобто спеціальних міток, що приписуються словам у текстах корпусу та позначають різноманітні лінгвістичні категорії, наприклад, граматичні, синтаксичні і т. інш. Анотація може бути внесена безпосередньо до корпусу, а може супроводжувати корпус окремим документом;

–повне врахування (total accountability) vs відбір даних (data selection). Принцип повного врахування в корпусному дослідженні полягає у неприпустимості вмотивованого відбору даних із корпусу з метою уникнення фальсифікації відомостей для підтвердження досліджуваної гіпотези/теорії. Згідно принципу відбору пошук у корпусі здійснюється з метою підбору специфічного прикладу чи низки ретельно відібраних прикладів для спростування висунутої гіпотези;

–багатомовні (multilingual) vs одномовні (monolingual) корпуси. Іншим критерієм, що розрізняє типи корпусів є кількість мов, представлених у корпусі. Більшість корпусів є одномовними, в тому сенсі, що вони репрезентують

лінгвістичну варіативність певної однієї мови. Натомість багатомовні корпуси – це корпуси, що побудовані на матеріалі двох або більше мов;

–корпуси першого покоління. Ідея створення корпусу (уже в сучасному його розумінні) зародилася в 60-х роках XX століття під значним впливом здійснених масштабних емпіричних досліджень, про які ми вже зазначали. До кінця 1960-х існувало декілька невеликих корпусів, укладених на різних принципах. Першим комп'ютерним корпусом є одномільйонний Браунівський корпус (the Brown Corpus), укладений у Браунівському університеті (США) лінгвістами Нельсоном Френсісом та Генрі Кучероюю. Створення корпусу мало на меті дослідження лінгвістичних особливостей американського варіанту англійської мови. Корпуси другого покоління – це продукти Інтернету і характеризуються значним обсягом. Так, у кінці 80-х років у Великобританії був створений перший мега-корпус, що задав новий стандарт для представницьких корпусів – Британський національний корпус. Для корпусу використовувалася детальна класифікація документів за декількома параметрами: вид мовлення (писемне, усне), для писемного за тематикою, типом видання (книги, періодика і т.п.), параметром утворення очікуваної аудиторії (високий, середній чи довільний) та складністю мови (складний, середній, простий). За заданим Британським національним корпусом стандартом були укладені представницькі корпуси багатьох європейських мов. [9]

1.3 Морфологічні, семантичні та синтаксичні анотації корпусів

Анотованість — це та характеристика, яка відрізняє КТ від електронного збору текстів. Оскільки в Україні корпусна лінгвістика — порівняно новий напрям, її термінологічний апарат перебуває на стадії становлення. Тому паралельно послуговуються такими термінами, як розмітка, анотування, маркування, індексування на позначення процесу та результату присвоєння текстовим одиницям маркерів, в яких закодовано певну лінгвістичну інформацію.

Семантична розмітка текстів Корпусу української мови – це завдання нового наукового проекту, над яким працює сьогодні колектив лабораторії комп'ютерної лінгвістики Київського національного університету імені Тараса Шевченка. Семантичне розмічування текстів – четвертий етап представлення інформації про одиниці тексту у Корпусі української мови. Три попередні етапи стосуються граматичного розмічування з метою автоматичного визначення граматичних параметрів тексту. Перший етап, базовий для всіх наступних, – морфологічне розмічування, у межах якого кожній словоформі приписується морфологічний код частини мови і категорійних ознак, що дозволяє здійснювати пошук контекстів не тільки за заданим словом (хоча така опція також працює), а й контекстів до всіх слів за заданими морфологічними ознаками. Другий етап – синтаксичне розмічування, мета якого змоделювати синтаксичну структуру вхідного речення на рівні словосполучень і приписати інформацію про типи синтаксичних зв'язків, а також побудувати дерево залежностей речення. Третій етап – сегментування словоформ на морфи. Усі етапи анотації тексту є формалізованими і дають інформацію про кількісні характеристики лінгвістичних одиниць – абсолютну частоту (можливі статистичні підрахунки відносної частоти, дисперсії, коефіцієнта варіації, коефіцієнта стабільності).

Для морфологічного розмічування важливим є зняття граматичної й лексико-граматичної омонімії, яке здійснюється на 94 % автоматично, що забезпечує достовірність роботи всіх етапів автоматичного опрацювання текстів корпусу. Семантичне розмічування відрізняється від граматичного і ставить за мету надати можливість користувачеві одержувати списки слів за заздалегідь укладеними семантичними параметрами, наприклад, таксономічними, а також досліджувати лексику за різними аспектами: наповненість таксонів, мовна поведінка в контексті, зсуви у значеннях як прояв системних відношень у лексиці тощо. У майбутньому планується також перевірка сполучуваності слів за семантичними ознаками у словосполученнях тексту та формування на цій основі списку синтаксичних відношень, а також

перевірка можливості автоматичного визначення переносних значень. Характерним є й те, що значення слова обов'язково позначає лише дистинктивні риси об'єктів, тому в тлумачному словнику стільки різних значень, скільки слів, – поняття ж відображають глибші, більш істотні семантичні властивості слів. [10]

1.4 Кримінальна інформація та інформаційні технології

Розслідування злочинів є динамічною системою, основна функція якої полягає в ефективному протистоянні злочинній діяльності. Його можна розглядати як вид пізнавальної діяльності, що має специфічні риси. Кримінально-процесуальним законодавством України визначаються форми, засоби, строки діяльності, яку здійснюють органи досудового слідства і дізнання з розслідування злочинів. Зміст цієї діяльності складають процеси виявлення, фіксації, вилучення, зберігання та використання інформації, що має відношення до розслідуваної події, і встановлення істини у справі. Вказані процеси мають назву інформаційних і утворюють у пізнавальній діяльності гносеологічний ланцюг: факт – відображення – інформація – знання.

Об'єктивний інформаційно-пізнавальний характер процесу розслідування злочинів, універсальність законів пізнання, а також системна природа будь-якого знання складають передумови дослідження пізнавальної діяльності слідчого як інформаційної системи.

Відомо, що в рамках інформаційного підходу виділяються та досліджуються інформаційні аспекти об'єктів, явищ, процесів. Указаний підхід у криміналістиці знаходиться в зоні постійної уваги науковців. Найбільш продуктивним, на нашу думку, являється підхід М. С. Полевого, який розглядає діяльність з розкриття і розслідування злочинів як криміналістичну інформаційну систему, найважливішими компонентами якої є людина і його діяльність, пов'язана з розкриттям, розслідуванням або попередженням злочинів; криміналістична інформація, що є об'єктом такої діяльності; засоби і методи, які використовуються в цілях перетворення криміналістичної інформації у форми, необхідні для прийняття певного рішення. Метою вказаної

інформаційної системи, як зазначає М.С. Полевой, є отримання різносторонньої і максимально значущої інформації, яка у наступному у сукупності буде необхідною і достатньою для формування системи судових доказів, що забезпечують встановлення і доказування істини у справі, тобто для пізнання злочину.[4]

Під сукупністю засобів і методів збирання, обробки і передавання даних для отримання інформації нової якості про стан об'єкта, процесу або явища розуміють інформаційні технології. Метою інформаційної технології є вироблення інформації для її аналізу людиною і прийняття на його основі рішення. З інформаційної точки зору процеси розкриття та розслідування можна розглядати як процес послідовного прийняття рішень. Таким чином, основу пізнання події злочину слідчим складають криміналістична інформація та інформаційні технології, орієнтовані на підтримку прийняття рішень.

У контексті криміналістичної діяльності прийняття рішень можна розуміти як вид розумової (інтелектуальної) діяльності і вольовий акт слідчого з вибору обґрунтованих (або оптимальних) варіантів вирішення проблемних слідчих ситуацій – тактичних, процесуальних, плануючих, організаційних тощо з метою формування системи доказів, що дозволяють встановити особу винного і розкрити механізм злочинного діяння.

Варто відзначити, що вимога обґрунтованості є однією з основних вимог до якості будь-якого рішення. З урахуванням того, що проблемні слідчі ситуації утворюються сукупністю умов, характерних для так званої слабо структурованої проблеми (інформаційна невизначеність, пов'язана з браком інформації про подію, що розслідується, або відсутністю надійних джерел її отримання тощо). За таких умов у них неможливо з абсолютною впевненістю передбачити наслідки вибору (прийняття рішень в умовах ризику).

Звідси, метою статті є аналіз складових діяльності слідчого з розкриття і розслідування злочинів як інформаційної системи і визначення шляхів удосконалення процесу прийняття рішень слідчим. Її новизна полягає в наданих практичних рекомендаціях з формалізації криміналістичної інформації,

яка складає фактичну основу криміналістичної характеристики злочинів, що дозволяє її подальшу обробку із застосуванням технологій інтелектуального аналізу даних для висунення й оцінювання типових версій про особистість злочинця (на прикладі фактичних даних про умисні вбивства, вчинені з особливою жорстокістю).[5]

Виходячи з того, що одним із основних компонентів криміналістичної інформаційної системи є криміналістична інформація, для забезпечення вимог обґрунтованості рішень, що приймаються слідчим, нам уявляється корисним розмежувати два її рівні. Як слушно зазначає з цього приводу Б.І. Сазонов, для підготовки і прийняття рішень застосовується інформація пізнавального і керуючого характеру. Перша містить склад проблемної ситуації і з'являється в процесі інформаційно-аналітичної роботи, спеціальних досліджень, контролю, прогнозування. Використання пізнавальної інформації дозволяє сформулювати цілі рішення і визначити шляхи їх досягнення. Основу будь-якого рішення складають інформація пізнавального характеру, що визначає фактичну інформаційну базу рішення, і керуюча, що містить правила її обробки, і регламентує алгоритм, спосіб дій особи, що приймає рішення. Для слідчого в якості пізнавальної може виступати інформація про конкретну ситуацію розслідування, фактичні дані, що слідчий отримує в ході розслідування (доказова та орієнтуюча інформація) тощо. Керуюча інформація міститься у нормах кримінального, кримінально-процесуального законодавства, інших нормативних документах, розпорядженнях керівництва, матеріалах з узагальнення слідчої та судової практики, методичних рекомендаціях з розслідування тощо.

Так, проведене дослідження матеріалів слідчої та судової практики по кримінальних справах про вбивства, вчинені з особливою жорстокістю, свідчить про те, що особливу цінність у розслідуванні даної категорії справ набуває керуюча інформація кримінально-правового характеру та відомості про джерела доказів. Основні результати дослідження викладено нами у роботах.

Головними завданнями цих систем є формування версій, визначення напрямків розслідування, надання рекомендацій про подальші дії . Але й до цих пір залишається актуальною проблема дослідження можливостей обробки інформації, що складає фактичну базу криміналістичної характеристики злочинів, із застосуванням відповідних сучасних інформаційних технологій для вирішення завдань прийняття рішень слідчим, зокрема, про висунення типових версій.[4]

На наш погляд, однією з причин низького рівня реальної інформатизації процесу прийняття рішень слідчим є недостатній рівень формалізації криміналістичних знань. Перспективи формалізації криміналістичних знань, як зазначає В.І. Шаров, полягають у максимальному виявленні області застосування формалізованих методів, поданні у формалізованому вигляді постановки завдань і визначенні методологічного напрямку їх вирішення.

Криміналістична інформація (керуюча та пізнавальна) та сучасні інформаційні технології є складовими прийняття обґрунтованих рішень слідчим. Одним із перспективних шляхів удосконалення діяльності з розкриття та розслідування злочинів є її інтелектуалізація, яка полягає у формалізації подання пізнавальної криміналістичної інформації та її обробці із застосуванням сучасних інформаційних технологій, орієнтованих на підтримку прийняття рішень і дозволяють легко інтерпретувати отримані результати. Це дозволить слідчому не тільки приймати обґрунтовані рішення, але й розширювати пізнавальні можливості.[5]

2 СТВОРЕННЯ РОЗМІЧЕНОГО КОРПУСУ КРИМІНАЛЬНО-ЗНАЧУЩИХ ТЕКСТІВ

2.1 Структура та контент корпусу кримінально-окрашених текстів

Існуючі текстові корпусу мають різні структури. Найбільш прості не мають ніякої структури і являють собою просту колекцію текстів. Часто, тексти згруповані за категоріями, які можуть відповідати жанрами, джерелами, авторам, мов і т.д. Іноді, ці категорії перетинаються, особливо в разі тематичних категорій, коли текст може бути релевантним більш ніж однієї категорії. Іноді, текстові колекції мають тимчасову структуру, найбільш загальним прикладом якої є колекції новин.

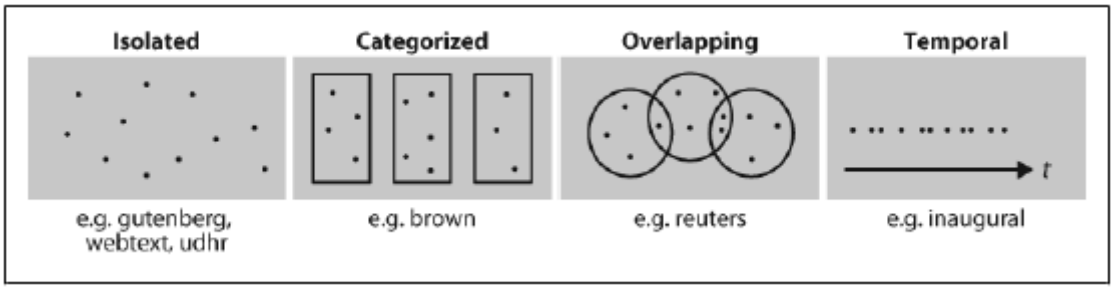

Рисунок 2.1 - Загальні структури для текстових корпусів

Найбільш простий тип корпусу - це колекція ізольованих текстів без будь-якої практичної організації; деякі корпуси структуровані за категоріями, наприклад, за жанрами (Brown Corpus); деякі категорії перетинаються, наприклад, тематичні категорії (Reuters Corpus); інші корпуси представляють мову в залежності від часу . Пакет NLTK підтримує ефективний доступ до різних корпусах і може бути використаний для роботи з новими корпусами.[3]

Найбільш популярним на даний момент є Британський національний корпус . Він був створювався протягом 1991-1994 рр. дослідниками з Оксфордського Університету і Університету Ланкастер. Його обсяг становить 100 млн. Слововживань і він значно більше, ніж його попередники. Тексти належать до кінця XX століття і являють різні жанри. У ньому можна зустріти газетні статті, науково-популярну літературу, приклади ділового листування, тексти на релігійну тематику, транскрибоване записи неофіційної мови, радіо-

шоу, урядової мови та ін. Варто відзначити, що саме Британський корпус отримав статус «національний» першим.[2]

Структура Британського Національного корпусу:

1) Основний корпус (прозові письмові тексти XVIII - початку XXI століття).

2) Синтаксичний корпус (в якому для кожної пропозиції побудована повна морфологічна і синтаксична структура).

3) Газетний корпус.

4) Паралельні корпуси.

5) Корпус діалектних текстів.

6) Корпус поетичних текстів.

7) Навчальний корпус англійської мови.

8) Корпус усного мовлення.

9) Мультимедійний корпус .

10) Корпус історії англійського наголосу.[1]

Як можна бачити з структури, Національний корпус охоплює майже всі грані мови. Для створення корпусу кримінально-значущих текстів буде відібрано інформацію з детективів англійською мовою, адже вони мають лексику, що буде корисною при створенні даного корпусу. В детективах ведеться детальний опис підготовки до злочинів, його здійснення, процесу розслідування і тому інше. Автори намагаються найточніше відтворити лексикон кримінального світу, щоб створити реальне відображення ситуації. Корпус буде представляти зібрання творів Агати Крісті, а саме наступних детективів: «Вбивство у Східному експресі», «Десять негренят», «Зло під сонцем», «Смерть настає в кінці», «Вбивства за алфавітом», «Безкінечна ніч», «Побачення зі смертю», «Таємниця блакитного потягу», «Тіло у бібліотеці» . Ці твори будуть поміщені в окремі файли з назвами, що відповідають назвам творів. В свою чергу файли будуть зберігатися в єдиній папці, що відповідає імені автора. Всі тексти, що входять в основний корпус, проходять процедуру морфологічної, синтаксичної та семантичної розмітки. Морфологічна та

синтаксична розмітка здійснюється автоматично, за допомогою спеціальних програм автоматичного морфологічного аналізу, а для втілення семантичної розмітки буде розроблений власний алгоритм мічення.

2.2 Автоматичний POS-tagging розробленого корпусу

POS tagging (part-of-speech tagging, розмітка частин мови) - етап автоматичної обробки тексту, завданням якого є визначення частини мови і граматичних характеристик слів в тексті (корпусі) з приписуванням їм відповідних тегів. POS tagging є одним з перших етапів комп'ютерного аналізу тексту.

У корпусній лінгвістиці, POS-tagging, що також називається граматичне мічення, це процес розмітки слова в текстовому (корпусі) як відповідні певну частину мови, заснований як на його визначенні, так і на контексті, т. е. на його зв'язку з суміжними і пов'язаними словами у фразі, реченні або абзаці, Спрощена форма цього зазвичай викладається дітям шкільного віку при ідентифікації слів як іменників, дієслів, прикметників, говірок і т.д.

Маркування частин мови є більш складним процесом, ніж просто складання списку слів і приписання їх відповідності до певних частин мови, тому що деякі слова можуть належати до більше однієї частини мови в різний час, а також тому, що деякі частини мови є складними або немовними. Це не рідкість - в природних мовах (на відміну від багатьох штучних мов) великий відсоток словоформ є неоднозначним.[6]

При тегуванні частини мови на комп'ютері зазвичай виділяється від 50 до 150 окремих частин мови для англійської мови. Наприклад, за корпусом Брауна, тег NN використовується для загальних іменників однини, NNS для загальних іменників множин, NP для іменників власних назв . Робота над стохастичними методами позначки Коіна Гріка (DeRose 1990) використовувала більше 1000 частин мови і виявила, що приблизно стільки ж слів були неоднозначними, скільки їх існує взагалі в англійській мові.

Більш просунуті (вищого порядку) моделі вивчають ймовірності не тільки пар, але і потрійних або навіть великих послідовностей. Так, наприклад, якщо

ви тільки що побачили іменник, за яким слід дієслово, наступний елемент може бути прислівником, артиклем або іменником, але набагато рідше - іншим дієсловом.

Коли кілька неоднозначних слів зустрічаються разом, можливості множаться. Однак легко перерахувати кожну комбінацію і призначити відносну ймовірність кожної з них, множачи разом ймовірності кожного вибору по черзі. Потім вибирається комбінація з найбільшою ймовірністю. Європейська група розробила CLAWS, програму мічення, яка зробила саме це і досягла точності в діапазоні 93-95%. [7]

Для виконання морфологічної розмітки тексту існують готові бібліотеки та програми-морфоаналізатори. Наприклад, Stanford Log-linear Part-Of-Speech Tagger - частина програмного забезпечення, яка зчитує текст на деякій мові і привласнює частини мови кожному слову (і іншому токену), таким як іменник, дієслово, прикметник і т.д., хоча, як правило, обчислювально додатки використовують більш дрібні POS-теги, такі як «іменник-множина». Це програмне забезпечення являє собою Java-реалізацію лог-лінійних тегерів частин мови.

Іншим прикладом морфоаналізатору може бути Stanford Parser, який представляє собою програму, яка розробляє граматичну структуру речень, наприклад, які групи слів об'єднуються (як «фрази») і які слова є предметом або об'єктом дієслова. Даний морфоаналізатор використовує знання мови, отриманих з розібраних вручну речень, для створення найбільш ймовірного аналізу речень.

2.3 Система семантичної розмітки кримінально-значущих текстів

У створеному корпусі кримінально-значущих текстів є морфологічна розмітка, однак семантичне анотування для майбутніх досліджень структури і семантики мови, зокрема питань сполучуваності лексичних одиниць, важливий крок у розвитку корпусу як інформаційно-довідкового ресурсу.

Семантична розмітка дуже важлива для вирішення завдань дослідження лексики, зокрема проблем сполучуваності слова, його синтактики. Створення

різних, призначених для користувача запитів, з урахуванням семантики дозволить на великому масиві прикладів уточнити, виявити правила відбирання необхідних кримінально-окрашених слів.

Одним із методів виконання семантичної розмітки корпусу кримінально-значущих текстів є підхід «мішок слів». Його суть полягає в тому, що для нас не важливий порядок слів у документі, в яких морфологічних формах вони представлені, а важливо тільки кількість входжень конкретних слів. Припустимо, що кожну тему можна охарактеризувати певним набором слів і частотою їх появи. Якщо в тексті конкретний набір слів вживається з певними частотами, то текст належить до певної теми.[3]

Грунтуючись тільки на цій інформації, будується таблиця «слово-документ». Де рядки відповідають словам (а точніше, їх леми), а стовпці - документам. У кожному осередку зберігається 1, якщо слово є в документі, і 0 - якщо ні. Хоча такий варіант і найпростіший, але не найкращий. Замість 0 і 1 можна використовувати, наприклад, частоту слова в документі або tf-idf слова. Такий спосіб представлення текстів у вигляді таблиці (або матриці) називається векторною моделлю тексту. Тепер, для того щоб порівняти два документи, потрібно визначити міру схожості двох стовпців таблиці. Для LSA частота входження в конкретному документі часто є якраз у вигляді індексу tf * idf, що розшифровується як «term frequency * inverse document frequency».

$$\mathrm{Tf} - \mathrm{idf}(\mathrm{t}, \mathrm{d}, \mathrm{D}) = \mathrm{tf}(\mathrm{t}, \mathrm{d}) \times \mathrm{idf}(\mathrm{t}, \mathrm{D}), \quad (2.1)$$

де Tf - частота терміна - розраховується як кількість входжень конкретного терміна в конкретний документ, поділене на загальну кількість слів у цьому документі:

idf -Document frequency – обернена частота документа - це кількість документів, в яких цей термін зустрічається, поділене на загальну кількість документів;

t – власне термін;

d – документ, в якому зустрічається даний термін;

D - загальна кількість документів;

Існує два основні підходи до семантичного пошуку, та й взагалі до порівняння документів за змістом. Перший підхід заснований на ручному наділення об'єктів деякими атрибутами і обробці саме атрибутів, а не самих об'єктів. Сюди можна віднести тегування, ручну каталогізацію, онтології і, звичайно ж, концепцію Web 3.0.

Другий підхід, про який і піде мова, заснований на протилежній ідеї: замість складних логічних правил використовується проста математична модель, замість тисяч годин ручної роботи - статистичний аналіз вже існуючих текстів. Початок цей підхід бере в роботах над методом LSA (Latent Semantic Analysis, неявний семантичний аналіз). Пізніше метод зазнав безліч модифікацій і отримав досить широку популярність. Досить сказати, що сьогодні Google і ряд інших великих пошукових систем використовують один з параметрів даного методу (а саме так званий індекс tf * idf) при ранжируванні результатів.[14]

Є припущення, що слова, що зустрічаються в одних і тих же контекстах (документах), а також контексти (документи), що містять одні й ті ж слова, є близькими за змістом. Інформацію про спільну зустрічальності можна організувати у вигляді матриці або, що те ж саме, у вигляді набору векторів в деякому багатовимірному просторі. Наприклад, якщо взяти матрицю, яку ми будували для LSA, то її рядки будуть являти собою T векторів в D-вимірному просторі, де T - кількість термінів, а D, відповідно, - кількість документів. Це і становить суть моделей векторного простору

Модель векторного простору має наступні обмеження:

–довгі документи погано представлені, тому що вони мають погані значення подібності (маленький скалярний продукт і велика розмірність);

–ключові слова для пошуку повинні точно відповідати умовам документа; подстроки слова можуть привести до «помилковому позитивному відповідності»;

–семантична чутливість; документи зі схожим контекстом, але з іншим терміном словник не будуть зв'язані, що призведе до «помилкового негативного відповідності»;

–порядок появи термінів в документі втрачається в поданні векторного простору;

Однак багато хто з цих труднощів можуть бути подолані шляхом інтеграції різних інструментів, включаючи математичні методи, такі як розкладання по сингулярним значенням і лексичні бази даних, такі як WordNet.[5]

Інший метод базується, спираючись на частотні словники. Частотний словник (або частотний список) - набір слів даної мови (або підмови) разом з інформацією про частоту їх зустрічальності. Частотні словники забезпечують можливість порівняти два корпуси, щоб визначити слова, найбільш характерні для кожного з них. Словник може бути відсортований по частоті, по алфавіту (тоді для кожного слова буде вказана його частота), по групах слів (наприклад, перша тисяча найбільш частотних слів, за нею друга і т. д.), по типовості (слова, частотні для більшості текстів), і т. д. У нашому випадку словник буде відсортовано по типовості, тобто буде створена вибірка зі слів, що найбільш характерні для даних текстів, та по частоті, тобто уже в порівнянні з іншим корпусом. Наприклад, якщо слово «shotgun» - «мисливська зброя» зустрічається в корпусі кримінальних текстів 10 разів, а у корпусі загальновживаної лексики лише один раз, то це слово буде пропонуватися користувачу як те, що має кримінальну окрашеність.

У зв'язку з тим, що розміри корпусів можуть бути різні, більш надійна оцінка частоти слів грунтується на приведення їх до ЧМС (частота на мільйон словоформ, англ. Ipm, instances per million words.

2.4 Алгоритм роботи створеного корпусу кримінально-значущих текстів

1 Формування корпусу кримінально-значущих текстів;

2 впровадження морфологічно-синтаксичної за допомогою існуючих бібліотек;

3 впровадження семантичної розмітки на основі існуючих методів та підходів;

4 порівняння спеціалізованих та неспеціалізованих корпусів на основі частотного словника ;

5 ручне коректування семантичної розмітки корпусу;

6 відбір слів за показником tf*idf;

7 створення словника кримінально-окрашених слів.

3 ОПИС ПРОГРАМНОЇ РЕАЛІЗАЦІЇ РОЗРОБЛЕНОГО ДОДАТКУ

3. 1 Опис методів розробки

У ході створення корпусу кримінально-значущих текстів було відібрано тексти кримінального характеру, а саме детективи XX ст. Для реалізації анотації корпусу було розроблено програмне забезпечення з можливостями морфо-синтаксичного та семантичного аналізу. Нижче наведено основний перелік методів, використаних для реалізації анотації текстів:

–annotate () – запускає процес аналізу тексту;

–listFiles() – повертає список файлів в папці (в нашому випадку папка - корпус, а файли - текстові документи);

–readLine() – зчитує те, що ввів користувач;

–startsWith() – перевіряє, чи починається рядок вказаною послідовністю символів;

–equals() – використовується для порівняння двох рядків;

–endsWith() – перевіряє, чи закінчується рядок вказаною послідовністю символів;

–replace() – дозволяє виконати заміну символу (або підрядка) в рядку;

–contains() – дозволяє перевірити, чи містить рядок інший підрядок або будь-який інший елемент;

–put() – слугує для внесення елементу до колекції;

– get() – слугує для отримання доступу до елементу колекції через ключ;

– remove() – слугує для видалення елементу із масиву;

– readFile() – дозволяє зчитувати необхідний файл;

– append() – дозволяє оновлювати значення об’єкту;

– println() – слугує для виведення необхідної інформації на консоль;

–setValue() – дозволяє встановити бажане значення об’єкту;

–containsKey() - перевіряє на наявність ключа, що передається, у списку.

3.2 Інструкція користувача

При запуску програми користувачу виводиться графічний інтерфейс.(рис.3.1)

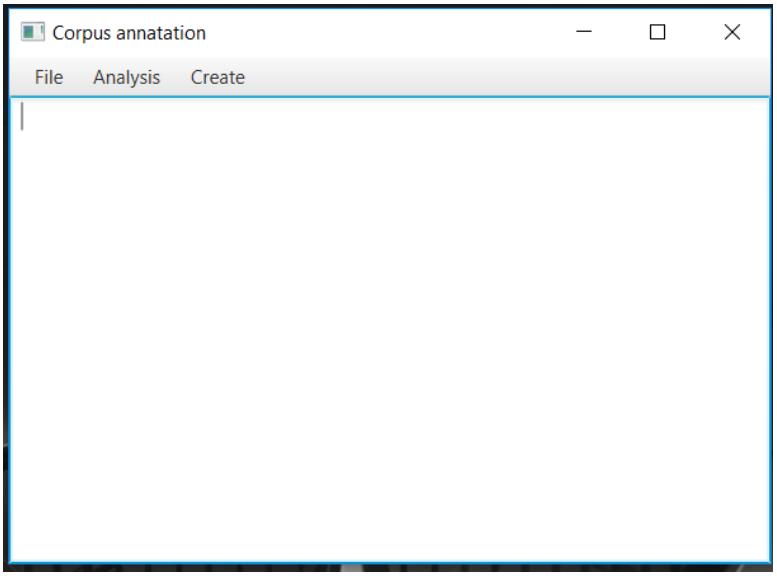

Рисунок 3.1 – Графічний інтерфейс користувача

При натисненні кнопки File можна вибрати корпус, що має бути розмічений у подальшому.(рис.3.2)

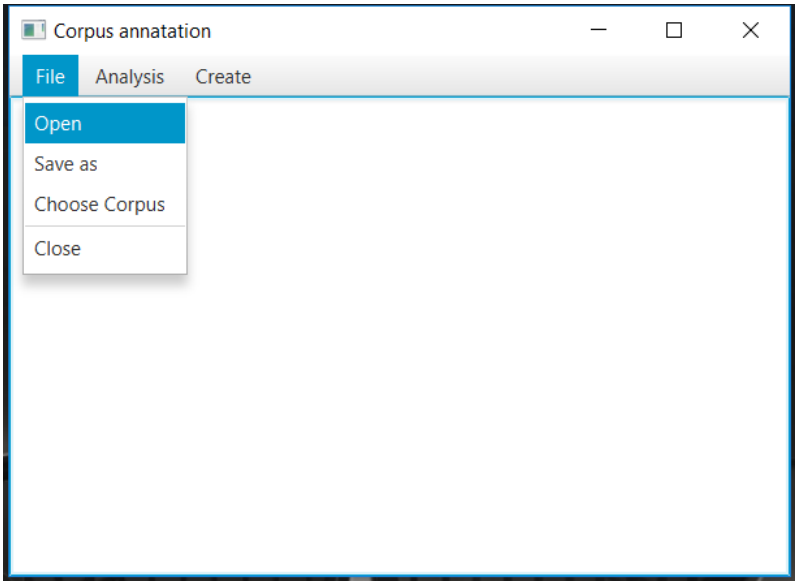

Рисунок 3.2 – Результат натискання кнопки File

При натисненні кнопки Analysis користувач може обрати тип анотування, що має бути застосований для даного корпусу.(рис.3.3)

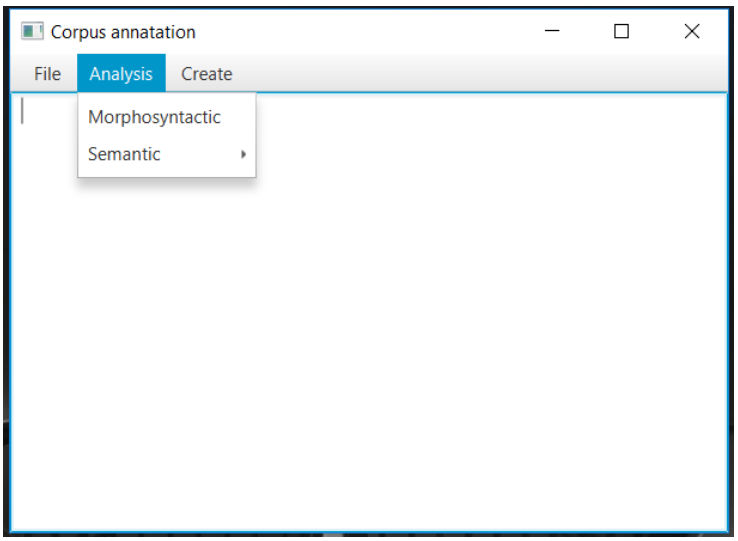

Рисунок 3.3 – Результат натискання кнопки Analysis

При виборі аналізу під назвою Morphosyntactic результатом буде побудова зв'язків членів речення, а також їх належність до відповідних частин мови.(рис.3.4)

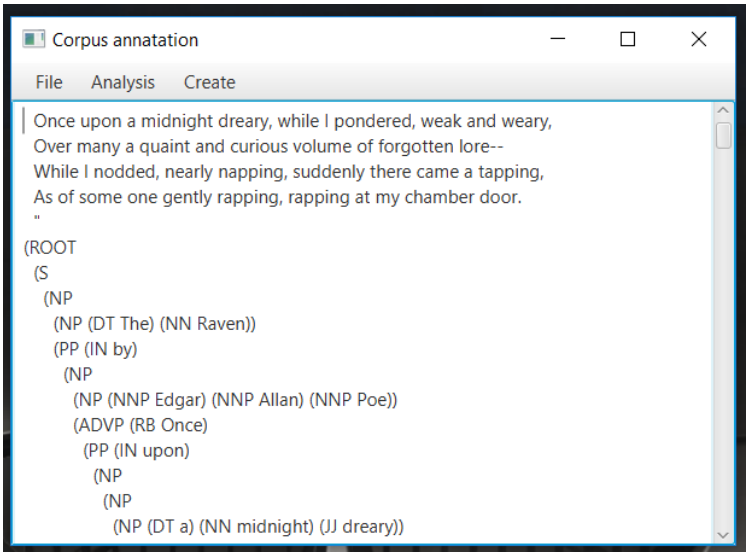

Рисунок 3.4 – Результат аналізу Morphosyntactic

При виборі аналізу під назвою Semantic користувач має обрати : вивести результати попередньо опрацьованого корпусу або обробити корпус заново. (рис.3.5)

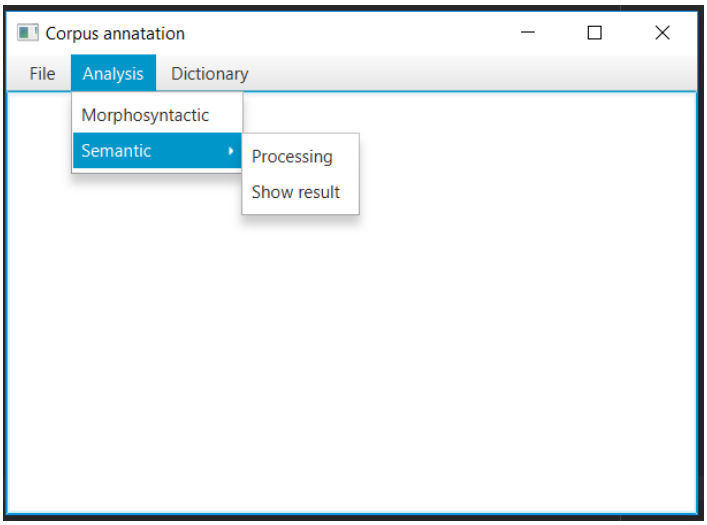

Рисунок 3.5 – Результат натискання кнопки Semantic

При натисканні кнопки Show result результатом буде відсортований список іменників з ваговими коефіцієнтами. (рис.3.6)

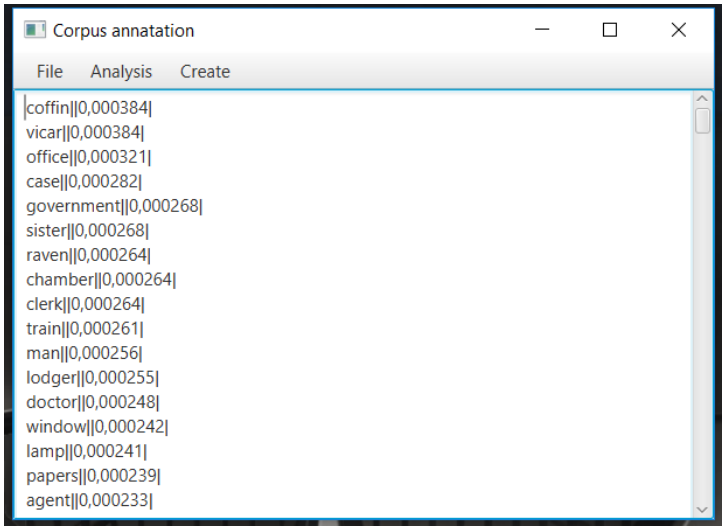

Рисунок 3.6 – Результат натискання кнопки Show result

При натисканні кнопки Dictionary користувач може створити словник кримінально-окрашених слів з попередньо відібраних лексем, або вже відобразити раніше створений словник.(рис. 3.7 – 3.8)

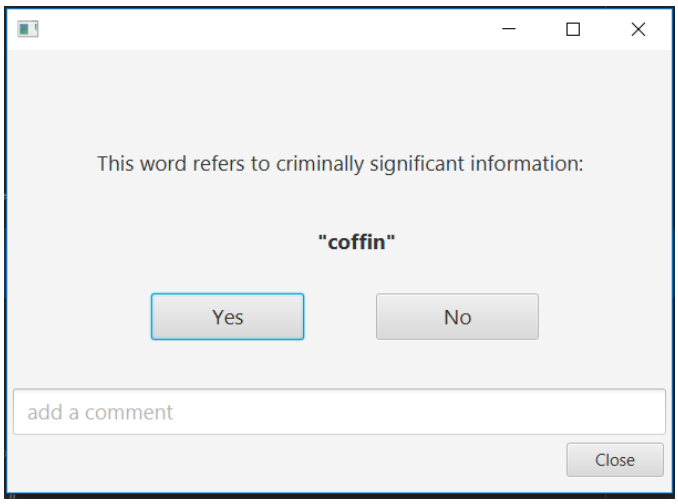

Рисунок 3.7 – Результат натискання кнопки Add

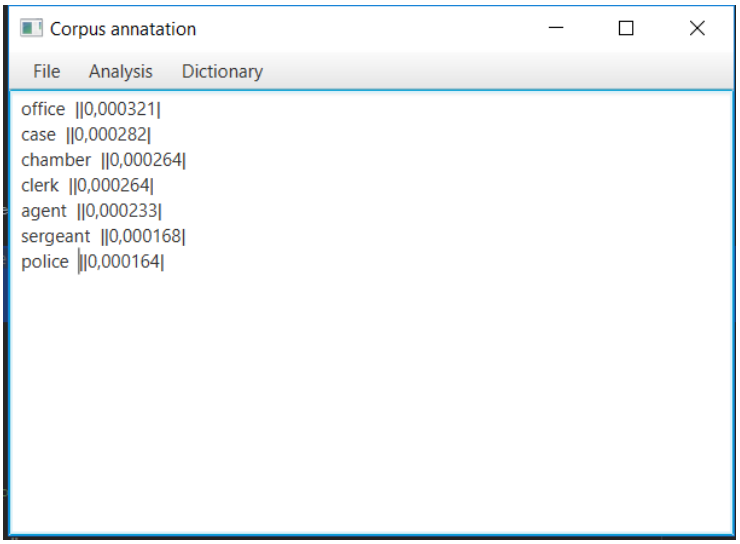

Рисунок 3.8 - Результат натискання кнопки Show

ВИСНОВКИ

Результати проведеного дослідження дають підставу зробити наступні висновки. Першим завданням нашого дослідження було опрацювати різні типи корпусів, виявити між ними схожість і різницю. На основі аналізу корпусів біло виявлено, що до цього часу ще не було створено жодного корпусу кримінально-значущих текстів. Також було проаналізовано різні типи розмітки: морфологічна, синтаксична та семантична. Було відібрано кримінально-значущу інформацію.

Наступним кроком було описано структуру та контент корпусу, визначено деякі особливості у структуризації корпусу. В ході роботи було розроблено алгоритм роботи розміченого корпусу кримінально-значущих текстів на основі існуючих підходів до реалізації даної задачі.

Дана тема потребує ще глибшого дослідження для створення більш точної роботи системи розмічення.

СПИСОК ДЖЕРЕЛ

1 Гальперін І.Р. «Текст как объект лингвистического исследования». – М.: Видавництво «Едиториал» УРСС, 2005.

3 Доповіді наукової конференції «Корпусная лингвистика илингвистические базы данных» / Під ред. А.С. Герда. СПб., 2002.

4 Зубов А.В. «Информационные технологии в лингвистике»: Навч. посібник. – М.: Видавничий центр «Академия», 2004.

5 Інформаційні системи (документальний пошук): Навч. посібник. – СПб.: Видавничий центр СПбГУ, 2002.

6 Інформаційні технології: Основи роботи з програмними засобами і інформаційними ресурсами Інтернет: Практикум / Сост. Захаров В.П., Арбатская О.А.; Под ред. В.П. Захарова; Санкт-Петербург. Держ. ун-т культури та мистецтв. – СПб.: Видавництво СПбГУКИ, 2003.

7 Коваль С.А. «Лингвистические проблемы компьютерной морфологии.» – СПб.: Видавництво С.-Петерб. ун-та, 2005.

8 Балл Г. Психологія праці та професійної підготовки особистості: навч. посібник / Академія педагогічних наук України; Інститут педагогіки і психології професійної освіти / Г. Балл, П.С. Перепелиця, В.В. Рибалко. – Хмельницький : Універ, 2001. – 330с.Калініна Т. Фізіологія і психологія праці: конспект лекцій / Т. Калініна – Х. : ХНЕУ, 2005. – 268с.

9 Крушельницька Я. Фізіологія і психологія праці: підручник / Я.Крушельницька – К. : КНЕУ, 2003. – 367с.

10 Рибалка В. Психологія праці особистості: навч.-метод. посібник / В. Рибалка – К. : КМПУ ім. Б.Д.Грінченка, 2006. – 159с.

11 Тимош І. Основи фізіології та психології праці: навч. посібник для студ. екон. спец. вузів / І. Тимош – Т. : Економічна думка, 1999. – 167с.

12 Траверсе Т. Психологія праці: навч.-метод. посіб. / Т. Траверсе – Інститут післядипломної освіти Київського національного ун-ту ім. Тараса Шевченка. – К., 2004. – 116с.

13 Баклицький І. Психологія праці: підручник / І.Баклицький – К. : Знання, 2008. – 655с.

14 Альфред, В. Ахо Компиляторы. Принципы, технологии и инструментарий / Альфред В. Ахо и др. - М.: Вильямс, 2015. - 689 с.

15 Савитч, Уолтер Язык Java. Курс программирования / Уолтер Савитч. - М.: Вильямс, 2015. - 928 с.

16 Шилдт, Герберт Java 8. Руководство для начинающих / Герберт Шилдт. - М.: Вильямс, 2015. - 720 с.

17 Эккель, Брюс Философия Java / Брюс Эккель. - М.: Питер, 2016. - 809 с.

18 Давыдов, Станислав IntelliJ IDEA. Профессиональное программирование на Java / Станислав Давыдов , Алексей Ефимов. - М.: БХВ-Петербург, 2005. - 800 с.

19 Нотон Java. Справочное руководство. Все, что необходимо для программирования на Java / Нотон, Патрик. - М.: Бином, 1996. - 448 с.

20 Баранов, А.Г. Моделирование применения корпусных методов для локальных лингвистических исследований // Материалы международной конференции «Диалог-2010». Режим доступа: http://www.dialog21.ru/dialog2010/materials/pdf/Baranov.pdf (29.09.2010).

21 Baroni, Marco, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston, and Marco Mazzoleni. 2004. Introducing the La Reppublica Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian. In Proceedings of LREC 2004. Lisbon: ELDA. Pp. 1771–1774.

22 Бабина, О.И. Автоматический морфологический анализ флективных языков / О.И. Бабина, Н.Ю. Дюмин // Наука ЮУрГУ: материалы 62-й научной конференции. Секции естественно-научных и гуманитарных наук. – Челябинск: Издательский центр ЮУрГУ, 2010. – Т. 2. (в печати)

23 Лінгвістична анотація корпусів [Електронний ресурс]: - https://www.academia.edu/Syntactic_annotation_for_Selkup_Evenki_and_Ker_corpora_

24 Корпусна лінгвістика: поняття та підходи [Електронний ресурс]: - https://myfilology.ru/177/korpusy-i-korpusnaya-lingvistika-osnovnye-ponyatiya/

25 Плунгян В.А. «Почему современная лингвистика должна быть лингвистикой корпусов? (Публичная лекция, прочитанная 01.10.2009)». [Электронний ресурс]. – Режим доступу: http://www.polit.ru/article/2009/10/23/corpus/

26 І.С. Місуно, Д.А. Рачковський « Пошук текстової інформації за допомогою векторних уявлень» [Електронний ресурс]. – Режим доступу: http://www.cl.uni-heidelberg.de/~sokolov/pubs/misuno05searching.pdf

27 Д.А. Поспєлов «Латентно-семантичний аналіз» [Електронний ресурс]. – Режим доступу: https://habr.com/post/326380/

28 Коваль С.А. «Роль корпуса в создании реалистичных моделей словоизменительной морфологи» [Электронний ресурс]. – Режим доступу: http://skowal.narod.ru/research/corpora2006/Koval_Corpora.2006.htm.