Міністерство освіти і науки України
Національний педагогічний університет імені М. П. Драгоманова

На правах рукопису

ВОЛОШИНОВСЬКА ІРИНА АНАТОЛІЇВНА

УДК 81'322.2:519.765:38:811.11

СТИЛЬОВА, ТЕМАТИЧНА Й АВТОРСЬКА АТРИБУЦІЯ НАУКОВИХ І ХУДОЖНІХ ТЕКСТІВ
(на матеріалі англійської, німецької та української мов)

10.02.15 – загальне мовознавство

Дисертація на здобуття наукового ступеня кандидата філологічних наук

Науковий керівник:
Толчеєва Тетяна Станіславівна,
доктор філологічних наук, доцент

Київ – 2013

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

КЛД – дивергенція Кульбака-Лайблера (Kullback-Leibler Divergence)

n-грам – послідовність n елементів

PC – головна компонента (Principal Component)

PCA – аналіз головних компонент (Principal Component Analysis)

ВСТУП

Дисертаційне дослідження присвячене вивченню стильової, тематичної і авторської атрибуції англо-, німецько- та україномовних текстів. У роботі розроблено нову методику здійснення стильової атрибуції текстів на основі аналізу частоти вживання слів в англійській, німецькій та українській мовах; установлено розбіжності між рангово-частотними закономірностями вживання слів у науковому та художньому текстах, що дає змогу диференціювати їх функціонально-стильовий різновид; доведено, що послідовність частоти вживання одного слова / двох слів ідентифікує тексти різних функціональних стилів за їх тематикою, тимчасом послідовність частоти вживання трьох і чотирьох слів свідчить про імовірну приналежність тексту певному авторові.

Сучасне теоретичне і прикладне мовознавство демонструє тенденцію до міждисциплінарної дескрипції тих об’єктів наукового спостереження, які становлять інтерес не лише для представників гуманітарного знання, а й перебувають у фокусі уваги дослідників точних наук, зокрема математики (А. Рогов, А. Романов, Т. Суровцова, О. Шевелев, S. Argamon, J. Binongo, M. Koppel), фізики (Ю. Головач, А. Ровенчак, I. Popescu, J. Rudman) тощо. З-поміж таких об’єктів аналізу варто назвати *атрибуцію текстів* (П. Вашак, Г. Мартиненко) – інтегрований філолого-математико-статистичний феномен групування текстів за ознаками стилю, часу, тематики, жанру, автора, статі, мови, літературної школи, ідейної течії.

Традиційно атрибуцію текстів здійснюють за допомогою *статистичних методів*: хі-квадрат (указує на статистичну однорідність текстів щодо певного мовного явища), критерій Стьюдента (показує на істотні/неістотні розбіжності середньої частоти появи певних одиниць мови у двох довільних зіставлюваних текстах) (В. Левицький, М. Марусенко, В. Перебийніс, Р. Піотровський, Ю. Тулдава, Г. Хетсо); *математичних методів*, які враховують багатовимірність простору спостережуваних

об’єктів (Д. Хмельов, D. Hoover, P. Juola, E. Stamatatos), зокрема методу аналізу головних компонент, що одночасно дає змогу проводити моніторинг розташування текстів та слів відповідно до їх подібності за тематикою або автором (H. Baayen, J. Binongo, J. Burrows, D. Holmes), та власне *лінгвістичних*: структурного з його методиками аналізу – трансформаційною (Н. Хомський, Л. Теньєр) і дистрибутивною (Л. Блумфільд, З. Харріс).

Особливий інтерес у завданнях пошуку інформації становлять наукові тексти з огляду їх важливості для ідентифікації наукової школи, приналежності до наукового напряму (J. Swales). Натомість і донині у функціональній стилістиці не вирішеною залишається проблема визначення автора наукової статті, особливо написаної у співавторстві (І. Колегаєва, Т. Радзієвська, Н. Разінкіна). Для текстів художніх творів вирізняються індивідуальні стильові й авторські ознаки, проте чітко ідентифікувати автора художнього твору можна лише шляхом застосування інтегрованого підходу із залученням методів математики, статистики і лінгвістики. З огляду на це постає необхідність стандартизації методів атрибуції текстів (M. Jockers, J. Rudman) та надання математичним параметрам аналізу тексту лінгвістичного змісту (I. Popescu). Така постановка проблеми актуалізує вивчення феномена атрибуції текстів з мовознавчих позицій.

Актуальність дисертаційного дослідження зумовлена його спрямуванням на пошуки тих процедур вивчення мовних явищ, які на тлі сучасних різноманітних комплексних методів і прийомів аналізу здатні забезпечити максимальну об’єктивність здобутих результатів. Комплексне поєднання формалізованих методів точних наук із класичними і новітніми лінгвістичними методиками є необхідним передусім для обчислення й обробки якісних характеристик і показників мовного матеріалу, з-поміж якого тексти різних функціональних стилів і різних мов найбільше потребують вдосконалення наявних процедур їх опису, особливо в зіставно-типологічному аспекті.

Зв'язок роботи з науковими програмами, планами, темами. Дисертацію виконано відповідно до тематичного плану науково-дослідних робіт Національного університету "Львівська політехніка" в межах держбюджетної теми "Пріоритети сучасної прикладної лінгвістики" (державна реєстрація № 0107U006226), а також Національного педагогічного університету імені М. П. Драгоманова за напрямом "Дослідження проблем гуманітарних наук". Дисертаційна робота є складовою наукової теми кафедри загального мовознавства та германістики Інституту іноземної філології Національного педагогічного університету імені М. П. Драгоманова "Зіставно-типологічне вивчення мов у синхронії і діахронії" (тему дисертації затверджено на засіданні Вченої ради Інституту комп'ютерних наук та інформаційних технологій Національного університету "Львівська політехніка", протокол № 6-2005/06 від 15 лютого 2006 року; перезатверджено на засіданні Вченої ради Національного педагогічного університету імені М. П. Драгоманова, протокол № 3 від 23 жовтня 2012 року).

Метою дисертації є виявлення закономірностей і відмінностей у здійсненні стильової, тематичної та авторської атрибуції англо-, німецько- та україномовних наукових і художніх текстів.

Поставлена мета передбачає вирішення таких **завдань**:

– визначити теоретичні засади вивчення атрибуції текстів у сучасному мовознавстві;

– розробити методику аналізу стильової, тематичної та авторської атрибуції англо-, німецько- та україномовних наукових і художніх текстів;

– виявити критерії розмежування англо-, німецько- та україномовних текстів наукового і художнього стилів на основі методу рангово-частотного розподілу слів;

– здійснити тематичну атрибуцію англомовних наукових текстів із залученням методу одночасного моніторингу групування текстів і відповідних їм слів та частоти вживання одного й більше слів;

– охарактеризувати процедуру виконання авторської атрибуції англомовних наукових текстів шляхом поєднання методу одночасного моніторингу групування текстів і відповідних їм слів із параметром послідовності вживання чотирьох слів у цих текстах;

– установити оптимальний розмір послідовності вживання одного та більше слів для авторської атрибуції англо-, німецько- та україномовних художніх текстів.

Об'єкт дослідження становлять англо-, німецько-, та україномовні наукові та художні тексти.

Предметом аналізу є стильова, тематична та авторська атрибуція англо-, німецько-, та україномовних наукових і художніх текстів, здійснена шляхом застосування методу частотного розподілу слів та методу одночасного моніторингу групування текстів і відповідних їм слів із залученням параметра послідовності вживання одного та більше слів.

Фактичним матеріалом дисертації є: а) наукові англомовні праці ("Crystal Design: Structure and Function" by Gautam R. Desiraju, "Lecture notes in Statistics: Bayesian spectrum analysis and parameter estimation" by Bretthorst, "Mathematical models for speech technology" by Stephen E. Levinson, "PLS Toolbox 3.5 for use with MATLAB" by Barry M. Wise), дисертаційні праці (Ch. Bostedt, Y. Kuzminykh, L. Pieterson, D. Talapin, M. True, R. Wegh), журнали (Physical Review B), а також вибірка наукових статей чотирьох авторів: проф. д-р. Pieter Dorenbos та проф. д-р. Andries Meijerink (голландська фізична школа), д-р. Gregory Stryganyuk (українська фізична школа) та проф. д-р. Georg Zimmerer (німецька фізична школа); німецькомовні праці (Wolfgang W. Osterhage Studium Generale Physik. Ein Rundflug von der klassischen bis zur modernen Physik, Michael Komma Moderne Physik mit Maple: von Newton zu Feynman, Rainer Scharf Ausgezeichnete Physik); дисертаційні праці (C. Granzow, A. Guesmann, T. Latz, C. Rotsch), україномовні праці (Електрика і магнетизм Т. Г. Січкар, А. В. Касперський, Конспект лекцій з фізики, Оптика М. О. Романюк), журнали ("Український

фізичний журнал", "Вісник ЛНУ, серія Фізична", "Фізика конденсованих високомолекулярних систем"), дисертаційні праці (В. Вістовський, А. Пушак, П. Савчин, Г. Стриганюк); б) *художні* тексти XIX-XXI століть: англомовні (M. Albom, J. Austen, Ch. Bronte, L. Carroll, A. Conan Doyle, Ch. Dickens, H. Fielding, J. Harries, S. King, J. Rowling, L. Tolstoy); німецькомовні (T. Fontane, A. Friedrich, K. Gier, T. Mann, J. Rudiger, W. Raabe, F. Shätzing, T. Storm, P. Süskind); україномовні (Ю. Андрухович, В. Винниченко, Л. Дереш, О. Забужко, О. Кобилянська, Л. Костенко, Б. Лепкий, П. Мирний, І. Нечуй-Левицький, Ю. Покальчук, І. Франко, В. Шкляр).

Методи дослідження. *Метод частотного розподілу слів* (закон Ципфа) використано для опису розподілу слів у тексті та розмежування наукового і художнього стилю; *метод одночасного моніторингу групування текстів і відповідних їм слів* (аналіз головних компонент) апробовано для тематичної та авторської атрибуції текстів; поєднано *метод ентропії* з параметром послідовності сполучуваності одного та більше слів для авторської атрибуції наукових текстів. Елементи *зіставно-типологічного методу* використано для зіставлення наукового і художнього функціональних стилів англійської, німецької та української мов та здійснення стильової атрибуції наукових і художніх текстів трьох мов; за допомогою *описового методу* узагальнено та систематизовано основні лінгвістичні параметри, придатні для проведення атрибуції текстів.

Наукова новизна визначається тим, що у роботі *вперше*: 1) *розроблено* комплексну методику здійснення стильової, тематичної та авторської атрибуції англо-, німецько- та україномовних наукових і художніх текстів; 2) *виявлено* відмінність рангово-частотних розподілів слів для наукових і художніх текстів (у науковому тексті стрімкість спаду ймовірності появи слова є меншою, ніж у художньому), *доведено*, що ця відмінність є статистично значимою (межі визначених інтерквантильних інтервалів для наукових і художніх текстів не перетинаються) та *запропоновано* її використання для стильової атрибуції текстів; *визначено* найчастіше вживані

слова у наукових і художніх текстах досліджуваних мов і проаналізовано загальні тенденції їх вживання (серед найчастіше вживаних слів у наукових текстах є загальнонаукові терміни, службові частини мови; у художніх текстах – слова на позначення частин тіла та періодів дня, займенники, службові частини мови); 3) *оптимізовано* процедуру проведення тематичної атрибуції текстів за допомогою методу одночасного моніторингу групування текстів і відповідних їм слів, а також показано її ефективність у разі одночасного аналізу текстів статей і відповідних їм тез доповідей, заголовків та анотацій. *Набула подальшого розвитку* методологія авторської атрибуції наукових текстів в аспекті поєднання таких методів і методик: методу ентропії разом із методом одночасного моніторингу групування текстів і відповідних їм слів із залученням параметра послідовності вживання чотирьох слів у текстах одного автора. *Укладено* словник найчастіше вживаних послідовностей чотирьох слів (науковий текст) та трьох слів (художній текст), а також установлено закономірності послідовності найчастіше вживаних слів у наукових і художніх текстах.

Практичне значення одержаних результатів полягає в можливості їхнього застосування у викладанні навчальних дисциплін: “Загальне мовознавство” (розділ “Методи дослідження мови”), “Прикладна лінгвістика” (розділи “Методи прикладної лінгвістики”, “Прикладні аспекти квантитативної лінгвістики”), “Стилістика” (розділи “Практична стилістика англійської мови”, “Стилістика німецької мови”, “Стилістика української мови”, “Функціональні стилі”, “Жанри наукового стилю”), “Теорія та практика перекладу” (розділ “Переклад науково-технічних текстів”), “Лінгвістичний аналіз художнього тексту” (розділ “Образ автора – категорія комплексного дослідження мови художнього тексту”). Положення та результати роботи, розроблене програмне забезпечення можуть бути використані для укладання тематичних, термінологічних та частотних словників, словників мови окремих авторів.

Апробація результатів дослідження. Основні положення дисертації висвітлено у доповідях на *дев'яти* міжнародних наукових конференціях: "Комп'ютерні науки та інформаційні технології" (Львів, 2008), "Граматичні читання" (Донецьк, 2009, 2011), "Горизонти прикладної лінгвістики і лінгвістичних технологій" (Київ, 2009), "Іноземна філологія у XXI столітті" (Запоріжжя, 2010), "Інтелектуальні системи прийняття рішень і проблеми обчислювального інтелекту" (Крим, 2010), "Актуальні проблеми сучасної філології" (Київ, 2012), "Научная дискуссия: вопросы филологии, искуствоведения и культурологии" (Москва 2013), "Шевченківська весна: 2013" (Київ 2013); на *одній* всеукраїнській науковій конференції "Пріоритети сучасного германського та романського мовознавства" (Луцьк, 2008). Дисертаційна робота обговорювалася на засіданнях кафедри прикладної лінгвістики Інституту комп'ютерних наук та інформаційних технологій Національного університету "Львівська політехніка" і кафедри загального мовознавства та германістики Інституту іноземної філології Національного педагогічного університету імені М. П. Драгоманова.

Публікації. Проблематику, теоретичні і практичні результати дисертаційного дослідження викладено в *одинадцятьох* публікаціях: у шести статтях, опублікованих у фахових наукових виданнях України, в *одній* статті – у міжнародному журналі та тезах доповідей *чотирьох* наукових конференцій.

Обсяг і структура роботи. Дисертація складається з переліку умовних скорочень, вступу, п'ятьох розділів, висновків, списку використаної літератури (262 найменування, із яких 134 іноземними мовами), списку довідникових джерел (3 найменування), додатків (8). Повний обсяг дисертації – 244 сторінки, основний зміст викладено на 182 сторінках, у тому числі 15 рисунків та 31 таблиця, додатки займають 34 сторінки.

У **першому** розділі **"Теоретичні засади вивчення атрибуції текстів різних функціональних стилів"** охарактеризовано особливості стильової, тематичної та авторської атрибуції текстів як складових сучасного загального мовознавства. Проаналізовано співвідношення термінів "стилеметрія",

“атрибуція”, “авторизація”, “класифікація”, “кластеризація”. Розглянуто особливості атрибуції наукових текстів. Розроблено класифікацію основних лінгвістичних параметрів для атрибуції текстів, обгрунтовано доцільність використання параметра послідовності вживання одного та більше слів.

У **другому** розділі **“Методологічна база аналізу атрибуції текстів різних функціональних стилів”** здійснено аналіз сучасних методів та підходів до атрибуції текстів, серед яких виділено: частотний розподіл слів у текстах (закон Ципфа), метод одночасного моніторингу групування текстів і відповідних їм слів (аналіз головних компонент) та метод ентропії (дивергенція Кульбака-Лайблера). Запропоновано комплексну методику аналізу стильової, тематичної та авторської атрибуції наукових і художніх текстів. Описано особливості реалізації програм, розроблених для розв’язання поставлених завдань у цій дисертаційній праці.

У **третьому** розділі **“Стильова атрибуція наукових і художніх англо-, німецько- та україномовних текстів”** проведено стильову атрибуцію наукових і художніх текстів за допомогою модифікованого закону розподілу частоти вживання слів у текстах. Розбіжності між параметрами атрибуції проаналізовано з позицій їх статистичної значимості в межах 2σ довірчого інтервалу.

У **четвертому** розділі **“Тематична та авторська атрибуція англомовних наукових текстів”** показано функціональні можливості методу одночасного моніторингу групування текстів і відповідних їм слів, та методу ентропії у поєднанні з параметром послідовності вживання слів для тематичної та авторської атрибуції англомовних наукових текстів.

У **п’ятому** розділі **“Авторська атрибуція англо-, німецько- та україномовних художніх текстів”** із залученням методу одночасного моніторингу групування текстів і відповідних їм слів англо-, німецько- та україномовні художні тексти згруповано відповідно до їх авторів. Укладено і проаналізовано словник послідовностей трьох слів для англо-, німецько- та україномовних авторів художніх текстів.

РОЗДІЛ 1
ТЕОРЕТИЧНІ ЗАСАДИ ВИВЧЕННЯ АТРИБУЦІЇ ТЕКСТІВ РІЗНИХ ФУНКЦІОНАЛЬНИХ СТИЛІВ

1.1 Атрибуція та суміжні терміни в сучасному мовознавстві

У сучасному мовознавстві сформувалися два підходи до визначення термінів “атрибуція” та “авторизація”: в одному – ці терміни не розрізняються, в іншому, – навпаки. Так, терміни “атрибуція” й “авторизація” в “Енциклопедії української мови” визначаються як синоніми [91, с. 36]. Атрибуція – встановлення авторства тексту на основі композиції, способів текстотворення, почерку, мови змісту і позатекстових відомостей про його походження та історію. Атрибуція здійснюється типологічним зіставленням неавторизованого твору з авторизованим або спеціально дібраним неавторизованим, щоб на підставі подібності (відмінності) довести прийняту, передбачувану чи дискутовану гіпотезу про авторство. Остання обставина часто є причиною того, що атрибуцію називають ще й авторизацією [там само, с. 36]. У “Словнику чужомовних слів” [264, с. 8] та “Словнику іншомовних слів” [265, с. 19], на відміну від “Великого тлумачного словника української мови” [263, с. 8] та “Енциклопедії української мови” [91, с. 36], зміст терміна “авторизація” передбачає надання автором уповноважень, згоди в будь-якій справі; ухвалу справи від особи, що цю справу сама виконала. У великому тлумачному словнику української мови: авторизація – підтвердження авторства, авторського права [263, с.8]; атрибуція – визначення достовірності, автентичності художнього твору, його автора, місця й часу створення [263, с. 45; 265, с. 93]. Згідно зі “Словником чужомовних слів” “атрибуція” – приписування анонімного художнього твору якомусь певному авторові [264, с. 54].

Г. Я. Мартиненко розглядає атрибуцію в ширшому розумінні. Поряд із авторською виділяє ще й “не авторську” атрибуцію, мета якої – віднесення

даного тексту до певної мови, стилю, періоду часу, літературної школи, літературного напрямку, ідейної течі, суспільного класу і т.д. Г. Я. Мартиненко не бачить принципової різниці між авторською і “не авторською” атрибуцією, адже в обидвох випадках завдання полягає у віднесенні тексту до наперед зафіксованої безлічі на основі подібності лінгвостилістичних характеристик [74, с. 6]. Крім того, він розглядає атрибуцію в контексті загальної теорії класифікації [74, с. 160]. Такої ж думки дотримується і П. Вашак щодо визначення терміна “атрибуція”, розрізнюючи “авторську” та “не авторську” атрибуцію [цит. за пр.: 47, с. 30].

С. Н. Бук викристалізовує такі різновиди “атрибуції”: авторська (з’ясування авторства тексту), стильова (з’ясування стилю тексту), тематична (з’ясування тематики тексту), часова (з’ясування часу написання тексту) тощо [13, с. 53]. Саме такий широкий підхід до визначення терміна “атрибуція” характерний для авторів сучасних дисертаційних праць, присвячених проблемам класифікації, кластеризації текстів [108; 121], а також для авторів програмних продуктів, що здійснюють атрибуцію текстів [101]. У дисертаційній роботі використовується термін “атрибуція”, який зводиться не тільки до визначення автора тексту, але і до встановлення стильової та тематичної приналежності тексту.

П. Берков виділяє три основні види критеріїв атрибуції: 1) біографічні, 2) ідеологічні та 3) стилістичні, які задають мінімальний набір ознак [8, с. 183]. Ці критерії повинні розглядатися в комплексі. Так, використання лише біографічних критеріїв атрибуції може призвести до переоцінки різних хронологічних періодів творчості авторів (місця їх перебування; сфери діяльності тощо). В. В. Виноградов запропонував поділити принципи атрибуції на дві групи, в яких ознаки класифікуються на суб’єктивні (суб’єктивно-комерційні, суб’єктивно-кон’юнктурні, суб’єктивно-естетичні, суб’єктивно-психологічні, суб’єктивно-ідеологічні) та об’єктивні (документальні, рукописні або ж археологічні, історичні, історико-

ідеологічні, історико-стилістичні, художньо-стилістичні, лінгвостатистичні) [17, с. 3–5].

У підходах до авторської та тематичної атрибуції текстів наявні розбіжності. Найуживаніші службові слова (артиклі, прийменники, займенники) вважаються одними з найкращих критеріїв визначення автора [133], у той час як для тематичної атрибуції ці слова можуть бути неефективними. Так, для авторської атрибуції анонімних повідомлень, написаних англійською мовою, дослідники працювали з наборами у 150 службових слів [129, p. 68], 303 службових слова [132, p. 476], 480 службових слів [193].

Поняття атрибуції текстів тісно пов’язане з категоріями класифікації та кластеризації текстів. Класифікація є найбільш розповсюдженою операцією інтелектуального аналізу даних. За її допомогою виявляють ознаки документів, що характеризують групу, до якої належить той чи інший об'єкт [1, с. 12–74]. Класифікацію здійснюють за наперед заданим алгоритмом – правилом віднесення тексту до того чи іншого класу. Алгоритм апробують за набором текстів, які завчасно належать заданому класу. Для перевірки ефективності класифікації використовують тексти, які не бралися для апробації алгоритму [207, p. 575–607; 235, p. 32]. Одним із завдань класифікації є визначення автора текстів. У цьому випадку набір текстів одного автора складають один клас і т.д. Процедура класифікації за автором відома також як *авторська атрибуція*.

Атрибуцію можна здійснювати не лише у процесі класифікації об'єктів, що передбачає їх розподіл (текстів, слів тощо) за класами (спільною темою, автором, періодом часу) відповідно до наперед визначених ознак, але й без заздалегідь заданих критеріїв [180, p. 119–163; 203, p. 537]. У цьому розумінні атрибуцію можна прирівняти до кластерного аналізу, де об'єкти вибірки (тексти, слова тощо) ділять на підмножини (тексти, об’єднані спільною темою та автором; слова, що входять до спільної синонімічної групи тощо), що називають кластерами. Кожен виділений кластер

складається з подібних об'єктів, а об'єкти різних кластерів суттєво відрізняються [229, p. 23–40].

Кластеризація відрізняється від класифікації тим, що попереднє навчання системи не проводять, і ознаки груп заздалегідь не задають [1, с. 75–133]. Її виконують без попередньої класифікації об'єктів і формулювання певного набору правил. У процесі кластеризації за допомогою статистичних методів моделювання, засобів інтелектуальних обчислень апріорно виділяють різні кластери даних, а потім шукають ознаки спільності у кластері [207, p. 495–528; 254, p. 69–75]. Отже, класифікація та кластеризація – це ті процедури, за допомогою яких можна здійснювати атрибуцію текстів.

Розглянуті терміни (атрибуція, класифікація, кластеризація) тісно пов’язані з терміном “стилеметрія”. Стилеметрія – прикладна філологічна дисципліна, яка визначає стильові характеристики з метою систематизації та упорядкування текстів та їх частин за типологією, атрибуцією, хронологією, діагностикою, реконструкцією тощо [73, с. 119; 75, с. 420]. Стилеметрія оперує кількісними параметрами певного стилю мови чи мовлення певних авторів. У стилеметрії текст є вихідним об’єктом дослідження, де увага зосереджена перш за все на кількісній організації тексту [33, с. 451; 74, с. 17]. Низка основних напрямів досліджень у комп’ютерній стилеметрії, виділених О. Н. Грінбаумом [33, с. 452], узгоджується із завданнями, які розв’язуються у процесах атрибуції та кластеризації текстів:

1) первинна обробка лінгвістичних даних: обчислення статистичних параметрів, статистичне оцінювання, побудова моделей;

2) обробка багатовимірних даних з використанням стандартних алгоритмічних процедур: факторного, дискримінантного, кластерного та інших методів;

3) створення частотних і алфавітно-частотних словників, словників письменників, асоціативних словників, словників ключових слів, словників-мінімумів;

4) автоматичний пошук текстів (авторський, жанровий, історико-хронологічний, бібліографічний).

1.2 Лінгвістичні параметри атрибуції наукових текстів

Кожна авторська праця характеризує самого автора або ж авторський колектив, даючи змогу відтворити найбільш вагомі фактори (часові, історичні, політичні, соціальні тощо), які впливали на процес написання роботи. Встановлення авторства тексту – це не лише філологічна проблема, але й міждисциплінарна, пов'язана з використанням потенціалу прикладних історико-філологічних дисциплін, застосуванням методів природничо-технічних наук – статистики, теорії ймовірностей, теорії комунікації, теорії штучного інтелекту тощо [74, с. 16–105]. У питаннях атрибуції постановка завдання, вибір параметрів аналізу тексту має мовознавчий аспект, подальша ж процедура опрацювання текстів вимагає застосування математичних методів дослідження.

Авторизація – це, з одного боку, традиційне філологічне завдання ідентифікації автора художнього твору, а з іншого, – це може бути і перевірка автентичності сумнівного зізнання, встановлення автора анонімного листа, аналіз електронного листування та спілкування в Інтернет-форумах, визначення автора програмного забезпечення [130, p. 7:10; 166, p. 81; 195, p. 246; 248, p. 58; 260, p. 383]. Атрибуція анонімного тексту або сумнівного авторства потребує зіставлення досліджуваного тексту з текстом-зразком, авторство якого є відомим [6, с. 41–51; 7; 194, p. 1261–1263].

За кількістю авторів, яких необхідно ідентифікувати, авторизація може бути поділена на 5 категорій: 1) бінарна авторська атрибуція, що розглядає лише дві кандидатури авторства; 2) багатокласова авторська атрибуція, де кандидатів більше, ніж два; 3) перевірка достовірності авторства (верифікація), за умови розгляду лише одного потенційного автора; 4) встановлення авторства анонімного тексту без знання потенційних авторів;

5) ідентифікація співавторства (вивчає документи, написані у співавторстві кількома авторами) [231, p. 151; 232, p. 613]. У випадку визначення автора статті виникають труднощі, пов’язані зі співпрацею авторів у процесі написання статті, наприклад: автор пише речення, абзац, розділ; один з авторів редагує статтю або кожен з авторів редагує свою частину тексту [226, p. 128]. Залежно від того, чи досліджують розбіжності між текстами різних авторів, чи різницю між текстами одного автора, виділяють інтер-авторську та інтра-авторську атрибуцію [218, p. 33–40].

Проблема авторської атрибуції наукових текстів є набагато складнішою, ніж художніх творів, адже арсенал мовних засобів наукових статей у багатьох випадках прямує до уніфікації. Мова і стиль же художнього твору тісно пов’язані з особистістю автора, певним особистісним поглядом, який пронизує твір і об’єднує його в єдине ціле. Для наукових статей властивим є зменшення емоційно-експресивних елементів [50, с. 173–179]. Це є наслідком властивості наукового мислення – пізнання світу через його логічне осмислення, шляхом перетворення фактів пізнання в логічні категорії, поняття, позбавлені експресивного забарвлення та емоційної оцінки [18, с. 3–17]. Характеристикою наукового стилю є академічний виклад з підкресленим інформативним напрямом, адресований спеціалістам. Авторська індивідуальність наукового стилю певною мірою виявляє себе в різних жанрах наукового стилю. До основних жанрів наукового стилю можна віднести: монографію, статтю, дисертацію, наукову доповідь, анотацію, реферат, тези, підручники, лекції, методичні розробки, навчальні програми [81; 96; 127]. Порівняння жанрів наукового стилю дозволило зробити висновки про можливість атрибуції робіт за жанрами в межах одного стилю, оскільки між самими жанрами існує розбіжність за змістом, композицією, граматичними, синтаксичними особливостями [115]. Завдання ускладнюється, якщо наукова стаття написана колективом авторів. У цьому випадку необхідно перевірити, чи сучасні мовознавчі підходи та математичні методи здатні виявити внесок співавторів у написання статей. Однак

простішим завданням атрибуції для праць із декількома співавторами є їх сортування за науковою тематикою або приналежністю до певного наукового колективу (наукової лабораторії), встановлення наукових зв'язків між дослідниками [212, p. 29].

Серед жанрів сучасного наукового стилю провідне місце займає наукова стаття. Значна частина наукових статей публікується в англомовних журналах з метою як найшвидшого поширення наукової інформації серед наукової спільноти. Журнальна наукова стаття, опублікована в престижному науковому журналі із високим індексом цитування, може мати більшу вагу, ніж монографія [242]. Наукова стаття була і залишається цікавим об'єктом мовознавчого дослідження як провідний жанр наукового дискурсу [53, 62; 93; 94; 95; 127; 161; 217; 240; 242].

Виділяють п'ять головних субжанрів наукової статті: 1) стаття, що звітує про конкретне дослідження (research report); 2) стаття суто теоретичного характеру; 3) оглядова стаття; 4) полемічна стаття; 5) коротка стаття-повідомлення [128, с. 51]. Такий поділ на субжанри не вичерпує внутрішньо жанрового розмаїття англомовної наукової статті.

Авторський стиль у науковому тексті аналізувала І. М. Колегаєва, яка відзначала підпорядкування індивідуально-стильового компонента функціонально-стильовому, жанровому, видавничому та іншим "колективним нормам" [62, с. 37]. А. П. Коваль підкреслює, що науковий текст – це інформація, що зараховується до так званого "колективного" (а не індивідуального) стилю [61, с. 3]. Свідоме невиконання колективних норм є проявом масштабності та неординарності особистості вченого [62, с. 37]. З цієї ж позиції О. М. Гніздечко розглядає можливість проведення авторизації наукового дискурсу з огляду на порушення або дотримання наукової риторики, що властиво авторитарній або неавторитарній мовній особистості [30].

Діяльність автора наукового тексту зводиться до: 1) передачі наукового повідомлення; 2) коментування повідомлення. Прояв авторського стилю

можна очікувати у коментарі до повідомлення стосовно значимості власного повідомлення [62, с. 48]. І. М. Колегаєва відзначає, що індивідуальний авторський стиль у науковому тексті може проявитись у структуруванні тексту через використання таких маркерів, як тематичні заголовки, цифрова і буквенна індексація з подальшим дробленням, використання цифри з крапкою, букви з дужкою і т.д. [там само, с. 65]. Т. В. Радзієвська вважає, що науковим статтям також притаманна індивідуальність, оскільки наукова стаття вимагає чіткості, логічності, однозначності, що залежить від мовної та фахової компетенції автора [93].

Н. М. Разінкіна [94, с. 32–33] та А. П. Коваль [60; 61, с. 27] вказують, що:

1) у науковому стилі всякий прояв елементів індивідуального є припустимим, але це не органічна якість стилю;

2) елементи індивідуального "вкраплюються" в текст наукового викладу, залишаючись при цьому чужостильовими (у художньому стилі вони визначають його).

З огляду на вище сказане атрибуція наукових статей за автором є складною процедурою. Атрибуція стає складнішою у випадку наукових жанрів, які мають великий ступінь стандартизації та уніфікації, таких, як анотації, каталоги, інструкції, проспекти, патенти. Поряд з цим, вивчення мови наукових праць розкриває нові сторони, пов'язані з проблемою становлення і розвитку тієї чи іншої області дослідження [94, с. 27].

О. С. Герд [28, с. 68–90] виділяє лінгвістичні компоненти структури наукового тексту, які можна використовувати для атрибуції текстів:

– макроструктура та загальна архітектоніка (композиція, типи й розташування розділів, параграфів, схем, рисунків);

– лексика (загальна і загальнонаукова) та вузька термінологія;

– морфологічні та словотворчі засоби;

– синтаксис;

– семантика, що пронизує всі частини тексту.

Індивідуальний авторський стиль вивчається як система, що виникає з реального багатоманіття різних ознак, стильових елементів, які відносяться як до змісту, так і до стилю аналізованого тексту [71]. Обов'язковою передумовою таких досліджень є виявлення специфічних ознак-маркерів стилю автора, характерних винятково для нього, які повинні усувати можливість збігу з особливостями іншого автора [218, p. 25]. У зв'язку з цим уводиться спеціальний термін "linguistic fingerprint" (лінгвістичний відбиток пальця), який би мав відображати індивідуальні лінгвістичні ознаки автора [205, p. 353]. Це передбачає створення стандартного переліку ознак стилю, своєрідну стильову анкету, заповнення якої дозволило б надійно і без громіздкого аналізу розв'язувати питання атрибуції. Проте М. П. Штокмар зауважує, що окремі елементи стилю, які можуть бути використані для визначення авторства конкретного документа, недоступні стандартизації й повинні щоразу виділятись у результаті серйозного і різностороннього аналізу стильової системи в аспекті її специфічності [122, с. 100–145]. У сучасних оглядах наголошують на відсутність усталеного набору критеріїв, який можна було б запропонувати для стандартизації процесів атрибуції текстів за автором [181, p. 215–223; 230, p. 352].

Таким чином, неможливо заздалегідь сказати, які ознаки виявляться чітко індивідуальними в творчості певного автора, особливо, коли це стосується атрибуції наукових текстів.

1.3 Лінгвістичні параметри атрибуції художніх текстів

Параметрами оцінки тематики або авторського стилю тексту можуть виступати одиниці будь-якого рівня мови – фонемного, морфемного, лексичного, синтаксичного, – а також структурні властивості тексту. Кількість вибраних параметрів для аналізу текстів може бути досить великою. Незважаючи на те, що для вирішення питань атрибуції текстів перед дослідниками ставиться завдання зменшення кількості параметрів, є й

тенденція до їх збільшення. Й. Рудман виділяє близько 1000 параметрів, які можна використовувати для розв'язання завдань стилеметрії [230, p. 351–359].

Нільс Е. Енквіст виділяє такі особливості контексту, які можуть бути основою для проведення стилістичного аналізу тексту:

1) текстовий контекст:

– лінгвістична основа (фонетичний контекст, фонемний контекст, морфемний контекст, синтаксичний контекст, лексичний контекст, особливості орфографії та пунктуації);

– композиційна основа (особливості вступу, закінчення, поділу на абзаци, типографічні особливості, взаємозв'язок даного тексту з іншими частинами тексту тощо);

2) екстратекстовий контекст:

– епоха та час написання тексту, приналежність до літературного жанру, взаємозв'язок між автором та читачем з огляду на їх соціальне положення, вік, стать, освіту, життєвий досвід тощо [45, с. 261]. Перераховані пункти можна розглядати як стилістичні характеристики, як потенціальні маркери стилю.

Слід звернути увагу, що вибрані параметри атрибуції тексту можуть зазнавати змін. На цю обставину звернув увагу Б. Я. Слепак. Він відзначає варіабельність параметрів трьох порядків:

1) варіабельність першого порядку – наявність в окремих авторів статистичного набору синтаксичних характеристик, які найбільшою мірою відрізняють одного автора від іншого;

2) варіабельність другого порядку – наявність внутрішньо авторських відмінностей, одне і те ж явище в частотному співвідношенні може по-різному використовуватись автором упродовж його творчості;

3) варіабельність третього порядку – наявність внутрішньо текстових відмінностей (встановлюються при вивченні функціонування різних синтаксичних явищ в окремих художніх текстах), які проявляються залежно

від частотного використання синтаксичних одиниць у сюжетно-композиційній побудові тексту [104, с. 103].

Багато дослідників зазначає, що для проведення тематичної та авторської атрибуції тексту успішною є робота з синтаксичними параметрами тексту [48; 51; 63; 102]. Такими параметрами можуть бути довжина речення, тип речення (розповідний, окличний, спонукальний, запитальний), структура речення (прості – складні, сурядні – підрядні), порядок слів у реченні (підмет + присудок, означення + означуване слово). Важливими для розпізнавання є також словосполучення (n-грами слів) як тип синтаксичного зв'язку. І. П. Севбо використовує аналіз синтаксичних конструкцій і будує дерева залежностей для визначення автора тексту [102]. Л. П. Іванова-Маркова аналізує вживання простих та складних речень в українських і російських поетичних творах однакової тематики, вказуючи на те, що ці твори характеризуються подібним вживанням складних речень, але відрізняються за вживанням різних типів односкладних речень [54, с. 250]. Ю. Бойко проводить авторську атрибуцію на основі аналізу залежності частоти підрядних речень та глибини складнопідрядного речення у творах англійських й американських письменників першої половини XX ст. [10, с. 293]. П. Вашак досліджував довжину речення у словах та довжину слова у графемах для атрибуції 17 робіт Я. Неруди, які були впорядковані відповідно до хронології його творчості. Для цього він аналізував довжину речення не просто від крапки до крапки, а згрупував речення на такі, що вводять непряму мову, мову автора, мову героїв, вставні речення тощо [15, с. 315].

Характеризуючи тексти на лексичному рівні, дослідники аналізують довжину слів [144, p. 13; 186, p. 495] та багатство словника [167, p. 105]. Виявлено, що довжина слова є показником стилю та жанру автора [164, p. 55]. Проте деякі дослідники не погоджуються, що довжина слова та речення може бути хорошим параметром для проведення авторської та тематичної атрибуції текстів [241, p. 211–212]. Показано, що не лише слова, що мають високу частоту вживання, але й слова з малою частотою вживання hapax

dislegomena (слова, що вжиті двічі у тексті), hapax legomena (слова, що вжиті один раз у тексті) придатні для атрибуції текстів [237, p. 45].

Н. П. Дарчук виявила, що найкращими параметрами авторського стилю у художніх творах можуть бути: 1) лексико-семантичні групи, які створюються на основі аналізу загальновживаної лексики (так, найкраще розрізнює авторський стиль лексико-семантична група говоріння, гірше – лексико-семантична група розміру); 2) лексико-граматичні категорії (найбільші процентні розбіжності виявлено для дієслів, іменників, прислівників); 3) високочастотна лексика (вплив тематики, авторського стилю, композиційної структури повідомлення); 4) низькочастотна лексика (становить лексичне багатство словника) [35].

Дж. Керрол для аналізу стилю художньої прози виділив 29 суб'єктивних та 39 об'єктивних характеристик. До об'єктивних характеристик він відніс кількість: абзаців, складів, речень, простих речень, інфінітивів, герундіїв, неозначених артиклів, перехідних/неперехідних дієслів, дієслів латинського походження тощо [68, с. 183–197]. Аналіз використання середньої довжини слова та розподілу довжини слова, середньої довжини фрази та її розподілу, ймовірнісних характеристик графем у різних функціональних стилях української мови показав, що ці параметри можуть бути статистичними параметрами стилів (белетристичний, політичний, науково-технічний) і відрізняти один функціональний стиль від іншого [11, с. 86].

Г. Хетсо запропонував методику атрибуції літературних творів на основі наступних семи параметрів: середня довжина слова в буквах, що обчислюється на основі вибірок розміром 500 слів; загальний розподіл довжини слова; середня довжина речення в словах, порахована на основі вибірок розміром в 30 речень; загальний розподіл довжини речення; лексичний спектр тексту на рівні словника; лексичний спектр тексту на рівні тексту; індекс різноманіття лексики [116, с. 48–61; 42]. За допомогою такої методики Г. Хетсо провів дослідження ряду анонімних статей, опублікованих у журналах "Время" та "Эпоха", з метою підтвердження їх авторської

приналежності Ф. М. Достоєвському. Він провів також аналіз творів М. Шолохова, зокрема, роману “Тихий Дон”, підтвердивши авторство М. Шолохова [116, с. 48–61].

1.4 Лінгвістичні параметри атрибуції текстів інших стилів

Відповідно до головних функцій мови можна виділити побутовий стиль (функція спілкування); діловий, офіційно-документальний та науковий (функція повідомлення); публіцистичний і художньо-белетристичний (функція впливу) [19, с. 6]. За сферами функціонування виділяють публіцистичний, науковий, офіційно-діловий, розмовний та художньо-літературний стилі [5].

Одним із параметрів для визначення автора тексту уважають розподіл частин мови (особливо дієслів) у тексті [162, p. 166]. Дослідження текстів белетристичного, наукового і соціально-політичного стилів української мови за частотою в них різних частин мови показало, що різним стилям властива різна частота частин мови. Для науково-технічного стилю, наприклад, характерною є висока частота іменників та прикметників, а частота дієслів у ньому значно менша [80, с. 153–158].

Для різних стилів характерним є вживання певних груп суфіксів, а от співвідношення між частотою *k*-суфіксальних і односуфіксальних словоформ не може виступати параметром розмежування стилів [80, с. 144–153]. Пропонується також для атрибуції текстів аналізувати односкладові слова та слова, що розпочинаються з голосної букви [156, p. 45; 170, p. 114]. Поряд із цим Д. Холмс вказує, що кількість складів у слові може виступати не тільки параметром визначення автора тексту, але і бути критерієм розрізнення мов [168, p. 89].

Результативними у вирішенні питань авторської атрибуції в межах одного стилю вважаються дослідження службових слів або ж знаків пунктуації [135, p. 29–37]. Дослідження частоти розділових знаків у науково-

технічній літературі, прозі, драматургії, поезії та суспільно-політичній літературі показують, що розділові знаки можуть бути певним критерієм відмежування одного функціонального стилю від іншого, а також – автора в межах одного жанру або стилю [106, с. 165–177]. Всебічне опрацювання статистичних параметрів стилів для всіх рівнів мовної структури української літературної мови проведено у колективній монографії "Статистичні параметри стилів", за редакцією В. І. Перебийніс. Зокрема, показано, що такі розділові знаки, як знак питання, оклику, три крапки є параметрами відмежування науково-технічного стилю від інших, адже в цьому стилі ймовірність появи цих розділових знаків наближається до нуля [там само, с. 165–172]. Аналіз вживання розділових знаків у межах одного жанру різними авторами також може бути параметром авторської атрибуції тексту [там само, с. 173–177]. В. І. Критська крім загальноприйнятих знаків пунктуації досліджувала також пробіл, абзацний відступ, параграф, дефіс на текстах з кібернетики, а саме на рефератах, анотаціях, резюме авторів. Вона дослідила, що знак питання не є характерним для реферативного наукового тексту [66].

Дослідження орфографічних та граматичних помилок у нередагованому тексті можна також використовувати як параметри для проведення авторської атрибуції текстів [140, p. 197–199; 193, p. 69–72; 248, p. 55–64]. Так, щоб провести авторську атрибуцію електронних листів аналізують помилки щодо повторення літер, заміни літер, інверсії літер, пропущення літер, злиття слів [193, p. 69].

До структурних параметрів тексту відносять: а) структуру оформлення тексту, як, наприклад, заголовок, розподіл інформації за розділами, абзацами, цитування, посилання, таблиці, графіки, ілюстрації; б) форматування тексту: шрифт, колір, виділення [248, p. 60]. Вказується на результативність вибору таких структурних параметрів, як розмір параграфів та пробіли між ними, використання підписів, скорочень, символів, підрахунок абзаців, що розпочинаються з великої та малої букви тощо для проведення авторської

атрибуції електронних листів [150; 248, p. 55–54]. Зокрема, було виділено 55 таких параметрів для авторської атрибуції електронних листів [150].

З метою авторської атрибуції 84 юридичних документів (грамот) XIV ст., М. М. Пещак провела аналіз композиційної структури тексту. Дослідниця показала, що наявність або відсутність таких восьми підрозділів оформлення тексту як нестандартизований текст, формула адресанта, формула дати і місця написання документа, формула назви свідків та підтвердження основного змісту грамоти, формула початку, формула закляття та формула імені писаря може виступати одним з параметрів для проведення авторської атрибуції текстів [88, с. 218; 89].

Здебільшого аналіз текстів відбувається на основі декілька видів параметрів або кількісного співвідношення між ними. Прикладом такого комплексного використання параметрів для атрибуції текстів є використання 56 основних і 56 похідних параметрів (довжина слів, речень, співвідношення між певними групами слів, частинами мови тощо) для класифікації 24 псевдонімних статей В. В. Маяковського [76, с. 34–35, 71–74]. Ефективність цього підходу була підтверджена і на опрацюванні текстів І. О. Буніна, О. І. Купріна [там само].

Найбільш оптимальними параметрами для проведення авторської атрибуції текстів є [163, p. 251–270]: а) одночасне врахування частоти слова і знаків пунктуації (ефективність авторської атрибуції для чотирьох авторів – 89%); б) послідовність із 2 букв (88%), послідовність із 3 букв (88%), послідовність із 4 букв (85%). Гіршу ефективність атрибуції демонструє використання послідовності вживання з: 5 букв (79%), 6 букв (72%), 7 букв (64%), 2 слів (54%). Такі параметри, як середня довжина слова (46%), речення (45%) виявились неефективними для авторської атрибуції. Аналізуючи ефективність використання різних текстових параметрів для авторської атрибуції, можна зробити висновок, що параметр послідовності вживання n-грам букв та слів може бути найбільш успішним для авторської атрибуції наукових текстів. Саме на цей параметр і звернуто увагу у цій

дисертаційній роботі. Вибір оптимальних параметрів для атрибуції тексту є важливим з огляду подальшого їх опрацювання сучасними математичними методами.

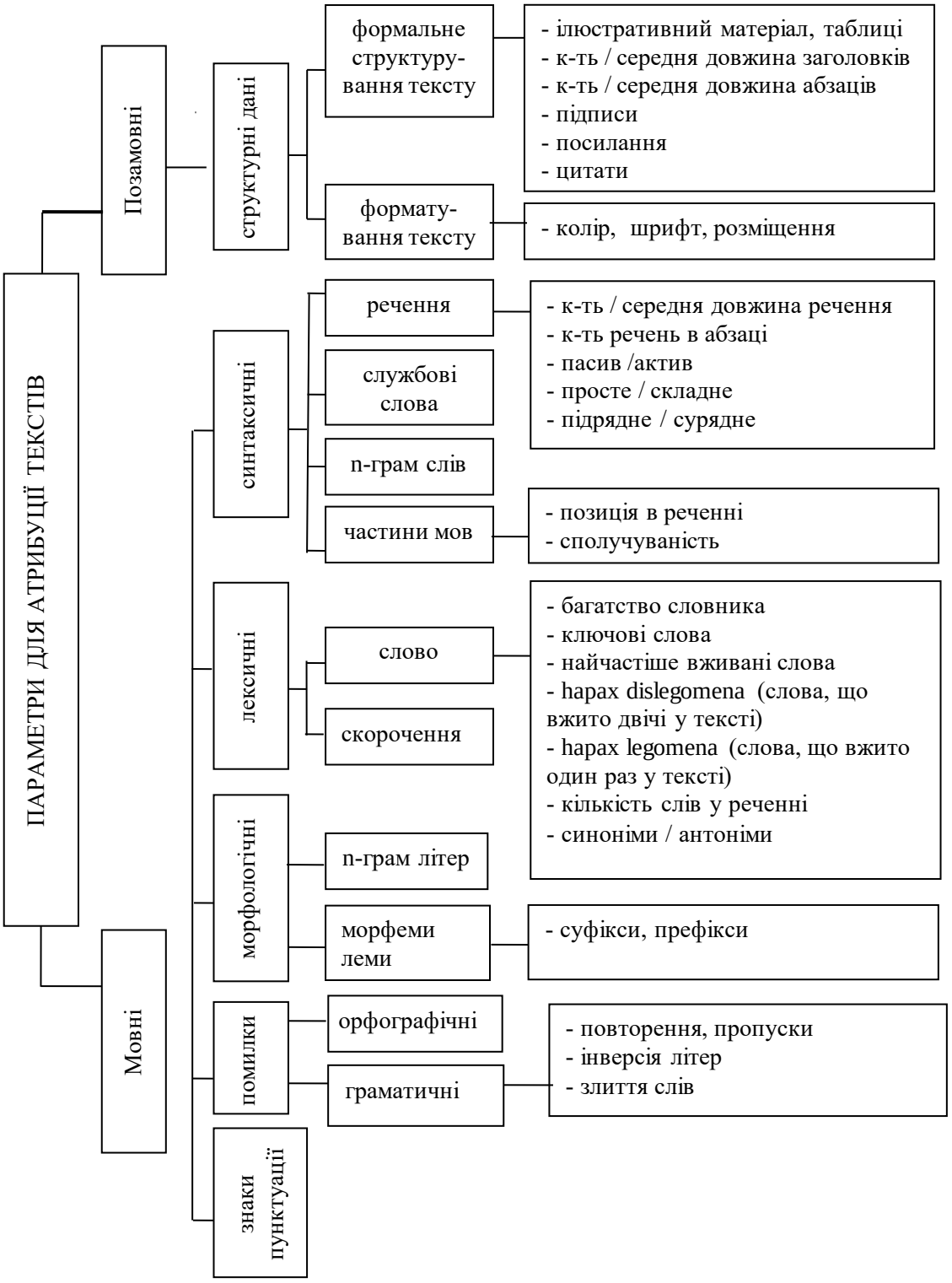
На основі узагальнення точок зору, приведених вище, зроблена спроба класифікації основних параметрів тексту (табл. 1.1), які можна використовувати для проведення атрибуції текстів. У цій таблиці параметри згруповані за лексичними, синтаксичними, морфологічними, структурними групами. Деякі автори виділяють ще контентно-специфічну групу слів, до якої належать слова, написані іноземною мовою, скорочення та акроніми, ключові тематичні слова і т. д. [101]. Однак ця група параметрів може бути складовою лексичної групи параметрів, що відносяться до словника.

1.5 Застосування лінгвістичного параметра послідовності вживання літер або слів для атрибуції текстів

Використання параметра послідовності вживання літер або слів до атрибуції текстів у поєднанні з різними математичними методами є одним із напрямів підвищення ефективності атрибуції, на чому наголошується в цій дисертації. У більшості випадків у текстах аналізуються окремі слова. Частотні параметри цих слів є основою для подальшого опрацювання математичними методами. Однак у деяких випадках більшу інформативність має частота послідовності повторюваності літер або слів, відомої як параметри послідовності вживання літер або слів. Цю послідовність також називають n-грам, де n – кількість послідовних літер або слів [151; 188; 190]. Слова, що стоять у тексті поряд, поєднані між собою синтаксично чи семантично. Одним із базових і найпростіших статистичних методів є аналіз статистики послідовності вживання літер або слів, який є простим у формулюванні та легким у застосуванні [151, p. 161].

Таблиця 1.1

Параметри для проведення атрибуції текстів



Параметром послідовності вживання літер або слів називають довільну послідовність з наперед заданою кількістю літер або слів. Кількість літер або слів, які необхідно аналізувати, дослідник вибирає та задає самостійно. Наприклад, послідовність з n слів (значення n може змінюватись від одного, двох, трьох і більше слів) означає, що потрібно аналізувати частоту вживання одного, двох, трьох або більше підряд вживаних слів у тексті [34, с. 69].

Ймовірність появи одного слова (одного-грама) є незалежною подією і має ймовірність:

$$P(w_i), \qquad (1.1)$$

де P – ймовірність появи події (літери, буквосполучення, слова, словосполучення, речення);

w_i – слово (літера, буквосполучення, слово, словосполучення, речення).

У випадку n-грам (n = 2, 3, ... n) ймовірність появи послідовності n з n кількістю слів/літер буде певним чином пов’язана з попередньою подією (літерою, буквосполученням, словом, словосполученням):

послідовність з двох слів: $P(w_i \mid w_{i-1})$, (1.2)

послідовність з n слів: $P(W) = P(w_1, w_2, .., w_n) = \prod_{i=1}^{n} P(w_i \mid w_1, w_2, ..., w_{i-1})$, (1.3)

У випадку послідовності з одного слова (літери) припускається, що кожне слово (літера) з’являється незалежно, без урахування контексту. Для послідовності з двох слів ймовірність нового слова залежить від ймовірності попереднього слова. Для послідовності з трьох слів ймовірність нового слова залежить від ймовірності комбінації попередніх двох слів. Проте, насправді, ймовірність появи певного слова залежить не лише від комбінації попередніх кількох слів. Легко побудувати безглузде речення або з неправильною граматикою, але таке, що складається з цілком прийнятних три-грам. Розглядаються також n-грами вищих порядків. Біграм також відомий як Марківська модель першого порядку. Вона базується на припущенні, що слово або літера залежать від попереднього слова або літери. Біграми можна

розширити до триграм, чотири-грам і т. д., називаючи їх ще Марківськими моделями другого, третього порядків тощо.

Проілюструємо сегментацію тексту на послідовність із одного, двох, трьох, чотирьох та п'яти слів (n=1, 2, 3, 4, 5) на прикладі речення: *"Many other examples together with a full discussion of the relation between depression value and the crystal structure will be presented in forthcoming papers"*. (Dorenbos P. The 5d level positions of the trivalent lanthanides in inorganic compounds / P. Dorenbos // Journal of Luminescence. – 2000. – Vol 91. – P.155 – 176). Ця послідовність виглядатиме так:

– послідовність вживання одного слова (1-грам слів): *many, other, examples, together, with, a, full, discussion, of, the, relation, between, depression, value, and, the, crystal, structure, will, be, presented, in, forthcoming, papers;*

– послідовність вживання двох слів (2-грам слів): *many other, other examples, examples together, together with, with a, a full, full discussion, discussion of, of the, the relation, relation between, between depression, depression value, value and, and the, the crystal, crystal structure, structure will, will be, be presented, presented in, in forthcoming, forthcoming papers;*

– послідовність вживання трьох слів (3-грам слів): *many other examples, other examples together, examples together with, together with a, with a full, a full discussion, full discussion of, discussion of the, of the relation, the relation between, relation between depression, between depression value, depression value and, value and the, and the crystal, the crystal structure, crystal structure will, structure will be, will be presented, be presented in, presented in forthcoming, in forthcoming papers;*

– послідовність вживання чотирьох слів (4-грам слів): *many other examples together, other examples together with, examples together with a, together with a full, with a full discussion, a full discussion of, full discussion of the, discussion of the relation, of the relation between, the relation between depression, relation between depression value, between depression value and, depression value and the, value and the crystal, and the crystal structure, the*

crystal structure will, crystal structure will be, structure will be presented, will be presented in, be presented in forthcoming, presented in forthcoming papers;

– послідовність вживання п’яти слів (5-грам слів): *many other examples together with, other examples together with a, examples together with a full, together with a full discussion, with a full discussion of, a full discussion of the, full discussion of the relation, discussion of the relation between, of the relation between depression, the relation between depression value, relation between depression value and, between depression value and the, depression value and the crastal, value and the crystal structure, and the crystal structurewill, the crystal structure will be, crystal structure will be presented, structure will be presented in, will be presented in forthcoming, be presented in forthcoming papers.*

1.5.1. Апробація застосування лінгвістичного параметра послідовності вживання літер або слів для атрибуції текстів у різних мовах. Параметр послідовності вживання літер або слів можна застосовувати у лексикографії для укладання частотних словників. За допомогою цього параметра можна легко отримати статистику “графічних слів” [119, с. 122–149]. Так, А. Я. Шайкевич уперше у практиці статистичної лексикографії уклав словник бінарних словосполучень (послідовність вживання двох слів) Ф. Достоєвського, де подано усі словосполучення графічних слів, що зустрічались більше одного разу [там само]. Він наголошує, що дослідження вживання бінарних словосполучень у тексті розкриває перед дослідником багатий матеріал для дослідження комбінаторики слів у тексті. Розглянутий метод аналізу тексту розширює сучасні підходи до укладання частотних словників нових типів, які б містили нову інформацію про мовні об’єкти [55, с. 16–33; 56, с. 19–22].

Параметр послідовності вживання літер/слів можна використовувати як параметри для класифікації та кластеризації текстів. Запропоновано використання характерних n-грам для тематичного сортування текстової інформації [138, p. 678; 204, p. 543; 208, p. 1447]. Такий підхід до тематичної

атрибуції вимагає формування набору характерної послідовності вживання одної і більше літер або слів для кожної з тематик.

Аналізуючи твори 45 письменників, Д. В. Хмельов запропонував для визначення автора тексту метод аналізу послідовності літер у тексті (ланцюгів Маркова) [117, с. 115]. Д. В. Хмельов показав, що навіть мала послідовність вживання літер (послідовність із 2 літер) може забезпечити авторську атрибуцію із ефективністю 73,3% [190, p. 305]. Крім послідовності літер як характеристики авторського стилю методом Марківських ланцюгів можна аналізувати і послідовність граматичних класів слів у текстах, написаних російською мовою [67, с. 98, 104].

При дослідженні ефективності використання послідовності з певної кількості літер для авторської атрибуції у текстах китайської, грецької та англійської мов було встановлено, що у грецькій мові оптимальною є послідовність із 3 літер, яка забезпечує ефективність атрибуції – 90% [222, p. 270]. Для англійської мови необхідно працювати з послідовністю із 6 літер, щоб досягти 98% ефективності. Для китайської мови, як і для грецької, найкращий результат атрибуції текстів досягнуто при послідовності з 3 літер із ефективністю 96%. Однак Дж. Грів, аналізуючи залежність ефективності авторської атрибуції художніх текстів, виявив дещо інші тенденції, ніж Ф. Пенг [163, p. 257]. Заявлена ефективність близько 98% простежувалась тільки у випадку атрибуції двох авторів. У процесі збільшення кількості авторів цей показник ефективності значно зменшувався і складав 60% для атрибуції 40 авторів. В. Кешел отримав дещо гіршу ефективність (67%), ідентифікуючи тексти англійською мовою 8 авторів для послідовності з 2 літер [188, p. 258].

Використовуючи метод послідовності вживання літер, проводили авторську атрибуцію в основному для художніх текстів. Для текстів, подібних за обсягом, кількістю авторів, мовою, методом 5-грам літер досягнуто максимальну ефективність 79% [163, p. 264]. Вважається, що авторська атрибуція проведена успішно, якщо її ефективність становить

близько 75% [там само]. В. Кеселж, Р. Пенг та Р. Клемент отримали задовільні результати щодо авторської атрибуції текстів для послідовності з 4, 5 та 6 літер [152; 188; 222]. Однак, Дж. Грів отримав оптимальні результати з ефективністю 88% для n=2 [163, p. 264]. Причому, у всіх згаданих публікаціях авторська атрибуція в межах n=8, 9, 10 була неефективною. Відмінність ефективності атрибуції залежно від розмірів параметру послідовності з двох та більше літер може бути пояснено так. В. Кеселж, Р. Пенг та Р. Клемент опрацьовували тексти авторів різної тематики, і в цьому випадку тематика тексту могла частково перешкоджати авторській атрибуції. Тоді як вибірка текстів Дж. Гріва, а це здебільшого газетні статті подібної тематики, могла відрізнятись насамперед індивідуально-авторським стилем.

Сушко С. О дослідив частоти повторюваності букв і біграм української мови на основі випадково вибраних україномовних текстів різних функціональних стилів. Для української мови він виявив такі характерні послідовності вживання із двох літер як: *на; но; ст; ов; ко; ро; ни; ер; ан; ом; пр*. Найчастотнішими літерами української мови є: *О, А, Н*. А загалом мало вживана літера *Ф* в україномовних текстах є досить частою в технічних текстах, бо використовується в таких словах, як функція, диференціал, дифузія, коефіцієнт і т. п.а [109, с. 100].

Використання параметра послідовності вживання літер або слів разом із кластерним аналізом дозволяє покращити результати атрибуції [175; 176; 234]. Було показано, що найкраще атрибуція “працює”, коли задано перших 200 найчастіше вживаних 2-грам. Із текстів восьми авторів успішно було ідентифіковано тексти семи авторів з ефективністю – 87,5%. Збільшення розміру n-грам слів веде до погіршення ефективності атрибуції. У згаданих вище роботах пропонується взагалі не розглядати послідовність сполучуваності слів, починаючи з 3 слів. Це пояснюють тим, що збільшення розміру послідовності сполучуваності слів веде до зменшення їх частоти у тексті (наприклад, частота появи *the* становить 52000, *of the* – 5930, *out of the*

– 362, *out of the room* – 76), зменшуючи у такий спосіб ефективність кластерного аналізу. Здебільшого, максимальне значення числа послідовності сполучуваності слів або літер в n-грамах досягає 4-5. Для японської та китайської мови створена прграма підрахунку n-грам, де послідовність сполучуваності літер досягає 255 [215, p. 611]. Значення n може містити як лексичну, так і синтаксичну інформацію. Для англійської мови, наприклад, при n=2, 3 на лексичному рівні виявляються “the”, “to”, а на синтаксичному – “-ing”, “-ed” тощо. Найбільш вдалою для англомовних текстів вважається послідовність вживання з 3 слів [199, p. 74].

З огляду на неузгоджені висновки різних авторів щодо ефективності атрибуції залежно від розміру послідовності сполучуваності літер або слів, у цій дисертаційній роботі звернуто увагу на ефективність атрибуції текстів залежно від розмірів послідовності вживання слів для *n* у діапазоні 1, ...10.

Висновки до розділу 1

У дисертаційній праці проаналізовано співвідношення термінів стилеметрія, атрибуція, авторизація, кластеризація, класифікація. Стилеметрія – прикладна філологічна дисципліна, яка вимірює стильові характеристики текстів та їх фрагментів з метою систематизації та упорядкування (типології, атрибуції, датування, діагностики, реконструкції і т. ін.). Атрибуція – це визначення достовірності, автентичності художнього твору, його автора, місця й часу створення. Поряд із цим у сучасній науковій літературі термін “атрибуція” стосується процесів кластеризації або класифікації текстів за стилем, тематикою, автором, часом, мовою, літературним напрямом. Останній підхід є найбільш коректним і прийнятним для розв’язання завдань дисертації, адже класифікація і кластеризація стосуються групування об’єктів, причому перша процедура (класифікація текстів, слів) відбувається за попередньо заданими ознаками, тоді як

кластеризація об’єктів – за апріорними ознаками, які виявляють об’єктивно за допомогою математичних методів.

На відміну від авторської атрибуції текстів художніх творів, особливості авторської атрибуції наукових текстів неможливо визначити за тими методиками, які розроблені на сьогодні у загальному теоретичному і прикладному мовознавстві. Це зумовлено тим, що авторська атрибуція наукових текстів має свою специфіку. Наукова стаття не завжди є одноосібною, її може писати колектив авторів, і арсенал мовних засобів, на відміну від художнього твору, прямує до уніфікації. За таких умов відсутня чітка межа між тематичними та авторськими ознаками для атрибуції наукових праць.

У роботі розроблено класифікацію основних лінгвістичних параметрів тексту, що придатні для здійснення авторської атрибуції тексту. Ці параметри класифіковано за двома типами: мовні (синтаксичні, лексичні, морфологічні, знаки пунктуації, помилки) та позамовні (структурні дані).

Вдосконалення наявних методик атрибуції текстів є можливим у випадку обчислення послідовності вживання літер і слів як лінгвістичних параметрів тексту. Так, наукові дослідження показують ефективність застосування параметра послідовності вживання літер до авторської атрибуції художніх текстів, у яких математичними методами у випадку двох авторів досягається ефективність розмежування 94%, а для 40 авторів – близько 60%. Аналіз ефективності застосування параметра послідовності вживання літер та слів для наукових текстів загалом науковцями ще не проводився. Очікується, що використання послідовності вживання з двох та більше слів дозволить врахувати зв’язки між словами, а відтак вдосконалити методику авторської атрибуції наукових текстів.

Основні положення цього розділу висвітлено у працях автора [250; 251].

РОЗДІЛ 2
МЕТОДОЛОГІЧНА БАЗА АНАЛІЗУ АТРИБУЦІЇ ТЕКСТІВ РІЗНИХ ФУНКЦІОНАЛЬНИХ СТИЛІВ

2.1 Сучасні методи атрибуції текстів різних функціональних стилів

Застосування статистичних методів стало невід'ємною частиною машинного автоматичного перекладу, інформаційного пошуку, атрибуції текстів і т. д. [207]. Зародження статистичного аналізу тексту датується кінцем XIX – початком XX століття. Ще у 1847 р. російський математик В. Я. Буняковський висловив думку про можливість застосування статистичних методів для розв'язання теоретичних проблем та практичних завдань мовознавчої науки [цит за пр.: 107, с. 1317–1318]. У 1887 р. Т. Менденхол зробив перші спроби використання кількісних методів для аналізу текстів п'єс В. Шекспіра, здійснивши простий підрахунок довжини речень і слів [210, p. 373–377]. А. Марков досліджував порядок літер у творах російської літератури [72, с. 239], Дж. Ципф – статистичні закономірності рангово-частотного розподілу слів у текстах [261].

Для виконання мовознавчих завдань застосовуються різні розділи математики: алгебра, теорія множин, математична логіка, теорія інформації, теорія ймовірностей та математична статистика. Основи принципів використання статистичних методів для дослідження мови наведено у багатьох працях 2; 13; 29; 32; 36; 69; 70; 86; 92]. Г. В. Ермоленко дає перелік понад 1000 літературних джерел з області статистичних досліджень мови [46, 25-146].]. Застосування статистичних методів в українському мовознавстві розглянуто в роботах В. В. Левицького, М. П. Муравицької та В. І. Перебийніс [69; 82; 87]. І. Краус наводить базові підходи квантитативної стилістики з виявленням загальних властивостей кількісних даних тексту та їх класифікації, вивчення питання квантитативних характеристик функціональних та індивідуальних стилів [65, с. 32]. Встановлення

англійського наукового стилю із урахуванням його статистичних характеристик XVII-XVIII ст. висвітлено в праці Т. С. Тетеріної [110]. На основі лінгвостатистичного аналізу піддано сумніву авторство Гомера і стверджується, що всі 24 пісні поеми “Іліади” за походженням – самостійні пісні, об’єднані в епос [58, с. 433].

Методи дослідження для проведення стильової, тематичної та авторської атрибуції текстів можна поділити на *контрольовані* (supervised analysis) та *неконтрольовані* (unsupervised analysis) [184; 185]. Контрольований метод вимагає наперед заданої інформації про досліджуваний текст. Систему, яка працює з цим методом, необхідно спочатку навчити, а потім вже проводити атрибуцію [165, p. 29]. За такою схемою працюють: лінійний дискримінантний аналіз [219, p. 460; 221, p. 179], машини опорних векторів [247, p. 138–170]. Неконтрольовані методи, натомість, не вимагають апріорних даних про аналізований текст, система самостійно навчається виконувати поставлене завдання без втручання експертів [165, p. 486]. До неконтрольованого методу дослідження відносять аналіз головних компонентів [12, с. 72–131], кластерний аналіз [174, p. 421–444].

Триває дискусія стосовно методів, які можна використовувати для ефективного виконання завдань атрибуції текстів [181, p. 215; 170, p. 111–117]. Методологія, що є успішною для розв’язання одного із завдань атрибуції, може виявитись непридатною для інших, тому кожен конкретний випадок вимагає свого індивідуального підходу й аналізу [221, p. 351; 230, p. 175]. Огляд літератури з питань стильової, тематичної й авторської атрибуції текстів дозволяє виділити такі основні математичні підходи до розв’язання цього класу завдань.

1. ***Рангово-частотний розподіл слів та закон Ципфа.*** Між частотою вживання слова в тексті f та рангом (порядковим номером) k цього слова – його місцем у впорядкованому за спаданням частоти списку всіх слів тексту – існує співвідношення відоме як закон Ципфа: $f = \frac{c}{k^s}$, де c – стала нормування,

s – показник ступеня, який певний час дорівнював одиниці і незалежав від автора, жанру, часу написання тексту (ґрунтовніше про це див.: розділ 3).

Спроби зіставлення текстів різних мов, використовуючи закон Ципфа, проводились неодноразово багатьма авторами. До прикладу, можна навести рангові розподіли слів для для англійської та китайської мов [201, p. 92], англійської та ірландської мов [202, p. 66]. Автори показали якісну різницю між частотними розподілами слів для англійської та ірландської мови, однак не змогли її кількісно описати в рамках закону Ципфа. Аналіз кривих рангового розподілу слів у текстах, написаних англійською, німецькою та угорською мовами, показав, що закон Ципфа не дозволяє повністю відтворити ймовірнісний розподіл слів для цих мов [216, p. 388–390]. Отже, виникає потреба пошуку апроксимаційних формул, які б дали змогу кількісно описати рангово-частотні розподіли слів для різних мов, стилів, авторів.

Перша спроба дослідити рангово-частотний розподіл слів в українській мові на основі закону Ципфа була зроблена у працях С. Бук [143, p. 161]. С. Бук та А. Ровенчак проаналізували ранговий розподіл слів залежно від жанру української мови. Однак відсутність чітко вираженого прямолінійного нахилу та використання однопараметричної моделі Ципфа не дозволило кількісно порівняти ці жанри. Вдалося тільки встановити, що для розмовного жанру s=1.09. Ю. Головач та В. Пальчиков апроксимували криву Ципфа для творів Ів. Франка “Лис Микита” та “Абу-Касимові капці” із параметром s=1.00 (рис. 2.1) [31, с. 23–24]. Як видно з рис. 2.1, неможливо повністю описати розподіл слів (на рисунку позначені кружечками) за допомогою закону Ципфа (суцільна чорна лінія). Неописаними залишаються слова, які мають високу частоту вживання у тексті, де суцільна апроксимаційна лінія показує значне відхилення від частотного розподілу слів у тексті.

Неможливість проведення кількісного порівняння рангово-частотних розподілів слів текстів, написаних різними мовами, в рамках однопараметричної моделі Ципфа зумовлює необхідність пошуку

багатопараметричних ципфоподібних функцій для опису рангово-частотного розподілу слів [252].

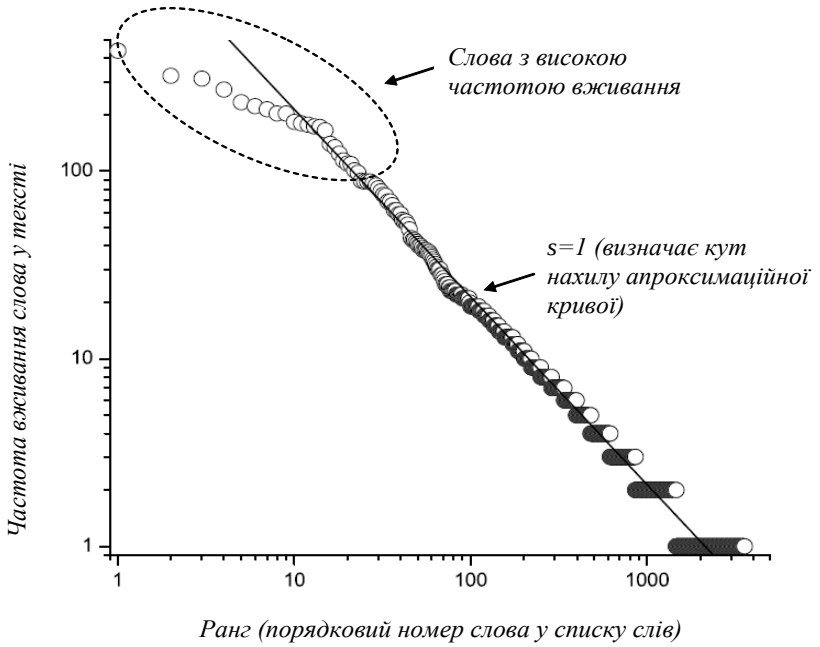


Рис. 2.1. Залежність частота-ранг для “Лиса Микити” І. Франка. Суцільна пряма – апроксимація ступеневою функцією із показником *s*=1.00.

2. ***Метод графів***. У цьому методі структуру досліджуваного об’єкта представляють як графічну побудову, де основні досліджувані мовні об’єкти наочно зображено у вигляді вузлів, вершин, точок, а взаємозв’язки між цими об’єктами – у вигляді відрізків, ребер, дуг. Таке графічне зображення полегшує аналіз взаємозв’язків і дозволяє проводити процедури класифікації та кластеризації. Про подібність текстів судять, виходячи із подібності самих графів, наявності спільних вершин, ребер, дуг. Теорія графів успішно використана для проведення атрибуції текстів, датованих з 1015 р. до середини XVIII століття [84, с. 15–18]. Дослідники створили дві системи кодів: 1) систему зі 150 кодів, яка враховувала всі граматичні класи слів, та 2) систему зі 135 кодів, яка не враховувала сполучники, прийменники, частки [там само]. Опрацювання частот зустрічі граматичних класів слів проводили за допомогою побудови графа сильного зв’язку. Вершини такого графа відповідають синтаксичному класу, а кожному сильному зв’язку

(синтаксичному зв’язку з високими частотами) приписується дуга графа. Зіставлення побудованих графів текстів, зокрема, за кількістю вузлових вершин, використано для атрибуції текстів.

3. ***Штучні нейронні мережі.*** Цей метод є складним у реалізації, чутливим до браку даних. Відзначимо тільки деякі роботи з використанням нейронних мереж, щоб продемонструвати, які параметри використовуються для аналізу. Б. Кджел [191, p. 120] використав 16 найбільш інформативних буквосполучень (*ed, er, ou, ve, to, of, he, et, na, me, ar, ur, it, de, se, ov*). Серед буквосполучень із 2-5 букв краща класифікація текстів досягається із двох букв. Дж. Гурн аналізував ефективність використання нейронних мереж для буквосполучення із двох та трьох букв [173, p. 319]. Ф. Твіді на вході нейронних мереж виміряв частоти одинадцяти функціональних слів англійської мови (*an, any, can, do, every, from, his, may, on, there, upon*), які на виході отримували вектор, відповідальний за авторський стиль [246, p. 2]. Метод нейронних мереж використано для дослідження приналежності творів “The Two Noble Kinsmen” В. Шекспіру чи Флетчеру, та “Едвард III” – В. Шекспіру чи К. Марлоу [209, p. 203; 211, p. 3].

Ефективність методу суттєво залежить від попереднього навчання [192, p. 299]. З метою полегшення процесу навчання системи поєднують штучні нейронні мережі з методом аналізу головних компонент для зменшення розмірності даних [244, p. 429].

4. ***Дерева рішень*** – це метод, де створюють ієрархічну структуру правил класифікації на зразок “якщо... тоді... ”, що має вигляд дерева. Для того щоб вирішити, до якого класу віднести певний об'єкт або ситуацію, відповідаємо на питання, що стоять у вузлах цього дерева, починаючи з його кореня. Питання можуть виглядати так: “Значення параметра А більше Х?” або “Значення змінної В належить підмножині ознак С?”. Якщо відповідь позитивна, переходимо до правого вузла наступного рівня, якщо негативна – до лівого вузла; потім знову відповідаємо на запитання, зв'язані з відповідним вузлом. Таким чином, можна досягти одного з кінцевих вузлів,

де є вказівка, до якого класу треба віднести розглянутий об'єкт. Цей метод цінний тим, що таке представлення правил наочне і його легко зрозуміти [142, p. 140; 165, p. 295–337].

Згідно з результатами О. Шевельова мінімальне значення обсягу тексту для задовільної класифікації за автором методом дерева рішень складає 30 000-40 000 символів, або 5000 – 6000 слів, або 400-600 речень. У цілому ж, результат класифікації методом дерев рішень, залежно від вибраних параметрів, лежить у межах 20 – 80 % [121, с. 11]. Класифікація за жанром методом дерев рішень розглянута на прикладі газетних статей різних жанрів: інформаційно-публіцистичного, офіційно-ділового і т.ін. [там само]. Особливість газетних статей – їх малий розмір. У зв'язку з цим мінімальний газетний текст складав 6000 слів. Дерева рішень дають низькі показники частот правильних класифікацій за жанром на всіх ознаках.

5. ***Дискримінантний аналіз*** (discriminant analysis), відомий ще як лінійний дискримінантний аналіз (linear discriminant analysis). В основі дискримінантного аналізу текстів – зменшення розмірності простору параметрів, які необхідні для опису системи. Перед застосуванням дискримінантного аналізу доцільно провести попередньо класифікацію текстів методом головних компонент. Це дозволяє швидко вилучити тексти, які не належать до досліджуваних класів.

Саме так зробили Р. Пенг і Н. Хенгартнер, класифікуючи тексти 9 авторів із сайту "Project Guttenberg" за 69 службовими словами. Перед застосуванням дискримінантного аналізу було використано метод аналізу головних компонент, щоб вилучити з аналізу тексти, які суттєво відрізняються між собою за стилем. Максимальна ефективність атрибуції досягала 92% [221, p. 178]. Дискримінантний аналіз було використано для дослідження зміни літературного стилю у часі на матеріалі тетралогії Я. Кемала. Для цього проводили аналіз найчастотніших голосних букв, слів, складів, частин мови, довжини речення в словах, довжини слова в тексті, довжини слів у словнику письменника [219, p. 461–462]. Метод лінійного

дискримінантного аналізу з урахуванням ентропії частотних характеристик слів використано для класифікації робіт студентів за тематикою та автором, диференціацією авторів за віком та освітою [136, p. 30].

6. ***Метод машини опорних векторів*** (support vector machines, SVM) орієнтований на аналіз невеликої кількості класів [145, p. 121; 247, p. 138–170]. Серед параметрів, що аналізують, – середня довжина слова і речення, число речень, число цифр, табуляцій, окремі букви та буквосполучення із 2 і 3 букв, граматичні класи слів, їх біграми. Цей метод у лінгвістиці застосовують для розпізнавання почерку [249], категоризації тексту, авторської атрибуції [100, с. 36; 153, p. 109; 160, p. 42]. С. Думейс [154, p. 148–149] відзначає покращення класифікації методом опорних векторів завдяки бінарному представленню параметрів.

7. ***Метод стискування даних***, за яким приналежність до класу визначається обсягом стиснутого тексту (архівом). Перевага методу в тому, що класифікація проводиться за самим текстом без аналізу додаткових параметрів тексту. Такий підхід не вимагає додаткового аналізу тексту чи його підготовки до класифікації у вигляді формування матриці векторів параметрів [118; 159, p. 200]. Кукушкіна та ін. [67, с. 96–109] аналізують ефективність архіваторів (zip, arj, ccompress, gyip, ha, ppm, rar та інші) та ентропійний метод Хмельова [117, с. 115–126] для класифікації 358 текстів за 82 авторами. За допомогою архіватора rar досягнуто ефективність 86,85%, а ентропійного методу – 84,14%.

8. ***Ентропійні методи.*** Термін ентропія в мовознавчу практику був уведений із теорії інформації та означає кількість бітів, необхідних для передачі інформації, закодованої у слові [233; 236]. Ентропія в лінгвістиці розглядається як міра рівномірності розподілу частоти слів у тексті. Вона використовується для порівняння частотних спектрів різних текстів [112, с. 141]. У лінгвістиці ентропію можна розглядати як міру невизначеності появи, наприклад, букви, слова, граматичної конструкції тощо. Так, імовірність появи комбінації двох букв є меншою, ніж імовірність появи в

тексті однієї букви. Вважається, що для першого випадку невизначеність є меншою. Ентропія є тим меншою, чим більшою є визначеність появи заданої комбінації символів, слів тощо. Кореляцію ентропії текстів можна розглядати як доказ подібності цих текстів.

Ентропія тексту *H(S),* де мовна одиниця x_j (літера, слово) з'являється в тексті S з імовірністю $p(x_j)$, визначається формулою:

$$H(S) = \sum_{j=1}^{n} p(x_j) \times \log_2 p(x_j). \quad (2.1)$$

Дивергенція ентропії обчислюється для оцінки подібності текстів S_i до наперед визначеного (опорного) тексту S_1. Опорним текстом називаємо текст, з ентропією якого порівнюється ентропія всіх інших текстів. Така міра подібності текстів відома як дивергенція Кульбака-Лайблера (Kullback-Leibler divergence, KLD) [196, p. 79–86; 197, p. 340–341]:

$$KLD(S_1; S_i) = \sum_{j=1}^{n} p_1(x_j) \times \log_2 \frac{p_1(x_j)}{p_i(x_j)}, \quad (2.2)$$

де $KLD\ (S_1;S_i)$ – дивергенція текстів Кульбака-Лайблера; n – кількість слів у тексті; $p_i(x_j)$ – ймовірність появи одиниці x_j (слова, фонеми, морфеми) в S_i тексті; $p_1(x_j)$ – ймовірність появи одиниці x_j в опорному S_1 тексті.

Відмінність методу ентропії від інших методів, де відбувається порівняння статистичних параметрів, може бути інтерпретована так: ентропія – це певна міра кількості інформації, яка передається за допомогою повідомлення (слова, речення, тексту тощо). У цьому розумінні ентропія є "важливішою" характеристикою, ніж просто частота вживання слова в тексті. З уведенням поняття ентропії з'являється можливість увести інформаційну міру використання слова.

Метод ентропії (дивергенція Кульбака-Лайблера) успішно використано для авторизації текстів [257, p. 92–105], для проведення авторської атрибуції, коли необхідно ідентифікувати роботи двох авторів [259, p. 59–68]. Для такої бінарної класифікації метод дивергенції демонструє високу точність

атрибуції робіт. Так, ефективність атрибуції (процент правильно віднесених документів) становить близько 89% для аналізу 25 текстів. Якщо аналізувати малу кількість документів, то метод ентропії є ефективнішим, ніж метод мереж та опорних векторів, і лише для великої кількості документів метод опорних векторів є придатнішим [258, p. 381–392]. Метод дивергенції Кульбака-Лайблера у поєднанні з параметром послідовності вживання літер використано для проведення авторської атрибуції текстів італійського політика, філософа, журналіста А. Грамши [137, p. 125211]. Аналіз 50 текстів А. Грамши та 50 текстів інших авторів показав, що для успішної атрибуції текстів за автором слід використовувати послідовність з 8 літер. Одна з переваг методу дивергенції Кульбака-Лайблера – відносна простота складання програми для обчислення дивергенції за формулою (2.2).

Подібний до цього методу є метод Хмельова, який також для групування текстів використовує порівняння ентропії [117, с. 115]:

$$L = \sum_{i,j}^{k} m_{1ij} \times \ln\left\{ \frac{m_{1ij}}{n_{1i}} \Big/ \frac{m_{2ij}}{n_{2i}} \right\}, \qquad (2.3)$$

де m_{1ij} – частота *i* словоформи в *j* тексті, що аналізується; n_{1i} – загальна кількість словоформ *i* елементу; m_{2ij} та n_{2i} – аналогічні числа матриці автора, з яким проводиться порівняння. Міру порівняння *L* називають мірою Хмельова. Значення *L* тим менше по модулю, чим менша різниця між текстами.

Для кожного автора підраховується частота появи досліджуваного параметра (літери, слова) і будується відповідна матриця-еталон (де кожному текстові у відповідність ставиться частота вживання у ньому кожного слова) на основі аналізу текстів автора. Така ж матриця будується для тексту автора, що розпізнається. Порівняння матриць здійснюється на основі порівняння ентропії текстів [121; 257].

Структура даного типу класифікації на основі застосування мір порівняння текстів така:

1) створюється матриця еталон для елементів тексту автора;

2) створюється така ж матриця для тексту, що класифікується;

3) проводиться порівняння матриць на основі математичних процедур.

Методи, що працюють за цією схемою, можуть відрізнятися вибором параметрів тексту та математичним методом, що використовується для порівняння матриць. Це можуть бути методи на основі обчислення ентропії текстів чи інші статистичні міри порівняння.

Розглянута формула (2.3) є подібною до формули Кульбака-Лайблера (2.2) для обчислення дивергенції ентропії текстів. Однак у формулі Кульбака-Лайблера (2.2) замість частоти m_{1ij} фігурує множник $p_1(x_j)$, який має зміст імовірності, що значно спрощує вимоги до розміру аналізованих текстів, вони не обов'язково повинні бути рівні за обсягом. Міри порівнянь "направлені" від матриці аналізованого тексту до матриці еталону. Можливо, і навпаки, – направленість на аналізований текст.

Для порівняння матриць використовують також підрахунок статистики хі-квадрат (χ^2) [64, с. 485]:

$$\chi^2 = n_1 n_2 \sum_{i,j}^{k} \frac{1}{m_{1ij} + m_{2ij}} \times \left(\frac{m_{1i}}{n_{1i}} - \frac{m_{2ij}}{n_{2i}} \right)^2 \qquad (2.4)$$

Значення хі-квадрат використовують як відстань між досліджуваними текстами, а не як міра статистичної однорідності текстів. Ефективність класифікації текстів із використанням різних мір порівняння та врахуванням особливостей текстів довів О. Шевельов [121].

Найбільша ефективність класифікації текстів при застосуванні ентропійних та хі-квадрат мір порівняння досягнуто у випадку використання пар букв. Використання такого параметра, як довжина речення для класифікації текстів дає втричі гірший результат правильних класифікацій текстів. Значимої різниці в ефективності класифікації, використовуючи ентропійні чи хі-квадрат методи, не виявлено. Різниця спостерігається тільки для середніх значень на користь ентропійних методів. О. Шевельов не аналізує залежність ефективності класифікації текстів залежно від розміру n-грам. У формулах Хмельова (2.3) та χ^2 (2.4) для порівняння текстів

використовують частоти появи параметрів. Ця обставина зумовлює необхідність порівнювати тексти однакових розмірів або вносити поправки у значення частот відповідно до обсягу текстів, використовуючи нормативні множники. Цього недоліку позбавлена формула Кульбака-Лайблера (2.2), оскільки в ній замість частот появи параметрів використовують відності частоти (ймовірності) появи параметрів.

9. ***Оцінка якості процедури класифікації текстів.*** У випадках, коли порівняння текстів виконують, аналізуючи частоти параметрів, актуальною є оцінка якості процедури, оскільки частота є статистичною оцінкою появи ознаки. Подібність числових значень частот параметрів ще не означає подібності ймовірностей їх появи. Тому порівняння частот параметрів повинно враховувати певні статистичні критерії. У цьому випадку для статистичної оцінки можна застосовувати критерій хі-квадрат, t-критерій Стьюдента [69, с. 115–158; 86, с. 72–82].

У випадку застосування цих критеріїв перевіряється виконання так званої нульової гіпотези, яка передбачає, що досліджувані тексти є статистично однорідними стосовно вибраного параметра. Якщо зіставляються стилі текстів, то це означає, що досліджувані тексти належать до одного стилю. Альтернативна гіпотеза – тексти відрізняються за стилем стосовно аналізованої ознаки. Коли використовують критерій хі-квадрат для встановлення статистичної однорідності текстів, то значення хі-квадрат обчислюють за відповідними формулами [69, с. 133; 86, с. 73]. Знайдене експериментальне значення χ^2 порівнюють із табличним значенням $\chi^2_{кр}$, що відповідає відповідному ступеню свободи та імовірності (рівня значимості). Якщо $\chi^2 < \chi^2_{кр}$, то приймається нульова гіпотеза про статистичну однорідність досліджуваних об’єктів стосовно виділеної ознаки. У протилежному випадку перевага віддається на користь альтернативи – досліджувані тексти статистично несумісні і, наприклад, не можуть належати одному автору.

У цій дисертаційній праці пропонуємо підходи, де порівняння текстів відбувається не на основі апріорно вибраного чи спеціально протестованого

за критерієм χ^2 набору лінгвістичних параметрів (наприклад, службових слів), а за набором усіх слів, що зустрічаються у тексті. Оскільки всі слова тексту одного автора не можуть задовольняти умові статистичної однорідності, то критерій χ^2 у випадку опрацювання всіх слів тексту не буде вказувати на статистичну однорідність двох текстів одного й того ж автора. У цьому випадку про статистичну успішність проведеного експерименту можна говорити, провівши певну кількість досліджень з визначення авторства статті, що дозволить виявити статистичні характеристики об'єктів – середнє арифметичне та стандартне відхилення, необхідні для знаходження довірчих інтервалів.

2.2 Метод моніторингу класифікації та кластеризації текстів (їх слів, словосполучень) для здійснення атрибуції

Збільшення обсягів загальнодоступної інформації в умовах стрімкого розвитку інформаційних технологій вимагає розробки процедур пошуку та сортування безлічі текстових документів відповідно до їх тематичної, авторської або ж хронологічної спорідненості. Такі завдання не можуть бути розв'язані без сучасних математичних підходів до обробки статистичних даних, отриманих у процесі опрацювання текстів.

Одним із таких методів, який на даному етапі активно вивчається до застосування задач класифікації та кластеризації документів, є ***метод аналізу головних компонент*** (одночасного моніторингу групування текстів та відповідних їм слів, словосполучень) [123]. В англомовній літературі цей метод відомий як "principal component analysis" (PCA) [1, с. 134–165; 123, с. 3–15; 182]. Вперше опис цього методу був представлений К. Пірсоном [220, p. 559–572] та Х. Хотелінгом [178, p. 417–441]. Проте, ще за 12 років до опублікування статті К. Пірсона, Дж. Сильвестр у своїй статті описав математичний апарат методики аналізу головних компонент [243].

Діапазон застосувань методу аналізу головних компонент широкий: хімія, генетика, біологія, демографія, екологія, економіка, психологія, агрономія, мовознавство тощо. Метод одночасного моніторингу групування текстів та відповідних їм слів, словосполучень з успіхом застосовують зарубіжні вчені для вирішення проблем мовознавства 136; 139; 146; 172]. Піонером у застосуванні цього методу в літературознавстві та мовознавстві був Й. Бероуз [146, p. 61–70], який провів авторську атрибуцію текстів, застосувавши метод одночасного моніторингу групування текстів та відповідних їм слів, словосполучень до аналізу службових слів. У вітчизняній мовознавчій літературі даний метод згадується у монографії М. А. Марусенко [76, с. 106] та використовується С. Н. Андрєєвим і Ю. А. Тулдавою для аналізу ознак афіксальних дієслів англійської мови [4, с. 5–10]. Широкого застосування цей метод отримав у працях, де досліджують особливості кластеризації, тематичної та авторської атрибуції текстів [136; 139; 147; 149; 184].

Метод використовується для аналізу внутрішніх закономірностей у великих масивах даних і дає можливість виявляти імпліцитні зв’язки між характеристиками текстового масиву [179; 182]. Суть методу полягає у зменшенні кількості змінних (слів, словосполучень), які є характерними для тексту [182, p. 1–8]. Загалом метод одночасного моніторингу групування текстів та відповідних їм слів, словосполучень базується на ідеї, що велика кількість взаємно корелюючих параметрів (“зовнішніх”, “видимих”), насправді, виражаються через невелику кількість “внутрішніх”, “невидимих” параметрів [49, с. 149]. Велика кількість ознак є лише індикатором певних властивостей явища, які безпосередньо не вимірюються. Так, рівень життя людей може визначатись такими параметрами, які можуть безпосередньо вимірюватись – величина заробітної плати (x_1), середній вік життя (x_2), тривалість відпустки (x_3) і т.д. У просторі даних {**X**} вони виокремлюється у групу ознак, які корелюють між собою через опис певного явища (рівень життя людей). У моделі методу головних компонент змінні x_1, x_2, x_3

утворюють головну компоненту (PC-1). У просторі даних може бути декілька компонент (PC-1, PC-2, ..., PC-n), які описують явище з різних сторін. Отже, завдання методу полягає у виділенні з великої кількості досліджуваних змінних (слів, текстів) невеликої кількості факторів, які найбільшою мірою визначають всі інші змінні. У результаті застосування методу головних компонент (одночасного моніторингу групування текстів та відповідних їм слів, словосполучень) великий масив даних представляють у вигляді малого масиву, тим самим виділяють певну закономірність, що відповідає за сутність досліджуваного явища [123, с. 12–26].

При розгляді текстів різних авторів за допомогою методу одночасного моніторингу групування текстів та відповідних їм слів, словосполучень простежується зміна частоти вживання слів при переході від одного тексту до іншого. При цьому слова групуються за подібною закономірністю зміни частоти їх вживання в текстах. Так можна виявити набори характерних слів та словосполучень, зміна частоти вживання яких корелює при переході від одного тексту до іншого. Якщо група слів, що вживалася дуже часто в одному тексті, відсутня в іншому, то можна сказати, що характеристика, описана даною групою слів, не є притаманною другому текстові. Такою характеристикою може виступати тематика тексту або ж авторський стиль. Так, наприклад, один автор використовує групу слів та словосполучень, що не притаманні іншому автору. Така група слів та словосполучень не обов’язково повинна налічувати велику кількість елементів, які, в свою чергу, не повинні бути найчастіше вживаними. Саме такі групи слів та словосполучень формують основні характеристики аналізованого набору текстів, розділяючи їх за автором, тематикою тощо. Кількість таких характеристик та значущість кожної з них залежать від спорідненості аналізованих текстів. Для наочного відображення результатів аналізу методом головних компонент тексти представляються у координатній системі, вісям якої відповідають виявлені характеристики та групи слів, що їх описують. Розташування тексту в напрямку кожної вісі визначається мірою

прояву однієї з виявлених характеристик – частоти вживання в даному тексті слів відповідної групи. При математичному описі процедури аналізу головних компонент виявлені основні характеристики називаються змінними (головними компонентами). Таким чином, кількість змінних, якими описується текст (спершу – це є обсяг словника), зменшується методом аналізу головних компонент до кількості виявлених основних характеристик (головних компонент). Таких нових змінних (основних характеристик масиву текстів) може бути, як правило, 10-20. Головна компонента, з погляду мовознавства, – це характеристика масиву текстів за мірою вживання певної групи слів у кожному з текстів. Наявність вибірки слів для кожної компоненти дозволяє проводити її аналіз і робити висновок про відповідність головної компоненти певній ознаці, характеристиці тексту.

Вхідні дані для такого аналізу формуються у вигляді матриці **A**:

$$\mathbf{A} = \begin{vmatrix} a_{11} & a_{12} \ldots & a_{1j} & \ldots & a_{1n} \\ a_{21} & \ldots\ldots\ldots\ldots & & & a_{2n} \\ \cdot & & & & \\ a_{i1} & \ldots\ldots\ldots & a_{ij} & \ldots\ldots & a_{in} \\ \cdot & & & & \\ a_{m1} & \ldots\ldots\ldots\ldots & & & a_{mn} \end{vmatrix}$$

НАЗВА ТЕКСТУ	*СЛОВА*				
	а	*так*	*іти*	*сонце*	*при*
Т1	17	6	8	5	10
Т2	22	7	5	0	13
Т3	19	9	0	0	15
Т4	20	7	6	7	15
Т5	20	8	2	4	12

Елемент a_{ij} матриці **A** відображає абсолютну частоту *j*-го слова в *і*-му тексті (наприклад, частота слово "*іти*" у тексті "*Т4*"). Кожен рядок матриці – це вектор, який відображає частоту слів у певному тексті. А стовпчик матриці – це вектор, який відображає частоту слова в кожному тексті.

У методі одночасного моніторингу групування текстів та відповідних їм слів, словосполучеень зв'язок між первинними даними, що задаються матрицею **A**, та головними компонентами записується у наступному розкладі частотної матриці **A**:

$$\mathbf{A} = t_1 p_1^{\mathrm{T}} + t_2 p_2^{\mathrm{T}} + \ldots + t_k p_k^{\mathrm{T}} + \mathbf{E}, \qquad (2.5),$$

$\boldsymbol{p}$ **(**$\boldsymbol{p_1}$**, ... ,** $\boldsymbol{p_k}$**)** – вектор головних компонент; $\boldsymbol{k}$ – кількість головних компонент; $\boldsymbol{t}$ – проекція вектора даних на осі головних компонент, тобто задає проекції

текстів на напрямки виділених характеристик у просторі головних компонент; вектор **E** – залишок.

Вектори $\mathbf{p}_i$ знаходяться як власні вектори коваріаційної матриці, тобто для кожного $\mathbf{p}_i$

$$\mathrm{cov}(\mathbf{A})\mathbf{p}_i = \lambda_i \mathbf{p}_i \,, \tag{2.6}$$

де λ_i є власними значеннями, що відповідають власним векторам $\mathbf{p}_i$**.** Коваріаційна матриця формується на основі матриці даних **A** за правилом:

$$\mathrm{cov}(\mathbf{A}) = (\mathbf{A}^{\mathrm{T}}\mathbf{A})/(m\text{-}1) \,, \tag{2.7}$$

де значення m визначається кількістю зразків (текстів, слів і т.д.).

Для вхідної матриці даних **A** та пари векторів $\mathbf{t}_i$, $\mathbf{p}_i$ виконується умова:

$$\mathbf{A}\mathbf{p}_i = \mathbf{t}_i \,, \tag{2.8}$$

тобто вектори $\mathbf{t}_i$ є проекціями вхідних даних на напрямки $\mathbf{p}_i$**.**

Власні значення λ_i є прямопропорційними до дисперсії вхідних даних, що описуються відповідною парою векторів $\mathbf{t}_i$, $\mathbf{p}_i$. Послідовність λ_i є спадною і у більшості випадків вхідні дані описуються (без значної втрати інформації) кількома першими парами $\mathbf{t}_i$, $\mathbf{p}_i$, кількість яких є значно меншою від кількості вхідних змінних. Таким чином, модель головних компонент трансформує багатовимірний (m) ознаковий простір у k-вимірний простір головних компонент ($k<m$). Використання аналізу головних компонент дозволяє створити набори словоформ у межах однієї змінної (головної компоненти). У межах цієї головної компоненти групується до 20–30% всіх словоформ тексту, зв’язаних між собою спільними ознаками. Практично, одна така компонента зі своїм наповненням створює “новий лінгвістичний об’єкт”, що може дати якісно нову інформацію про мову методами комп’ютерної лінгвістики [57, с. 28–29]. Результати застосування методу одночасного моніторингу групування текстів та відповідних їм слів, словосполучень, зазвичай, представляють у графічній формі у просторі головних компонент. У побудованому просторі головних компонент координати тексту вказують на міру прояву відповідної характеристики у тексті, а координати словоформи відображають її вклад в опис відповідної характеристики.

Переваги методу одночасного моніторингу групування текстів та відповідних їм слів над іншими методами в тому, що: а) результати аналізу є автоінформативними, тобто не базуються на попередньо заданих критеріях щодо числа груп, результатах класифікації тощо; б) рангування параметрів відбувається з урахуванням усіх первинних параметрів та всіх зв'язків між параметрами; в) багатовимірний простір первинних ознак трансформується у простір меншої вимірності, що полегшує процедуру атрибуції текстів.

Метод одночасного моніторингу групування текстів та відповідних їм слів був використаний для вивчення питання приналежності авторства 15-ї книги "Королівська книга країни Оз" із циклу книг про чарівну країну Оз (Лаймен Френк Баум, чи його послідовниці Руті Пламлі Томпсон) [139, p. 9–17]. Проаналізувавши твори кожного автора на основі частотності 50 службових слів, обмеживши першою та другою головними компонентами, розділено твори Л.Ф. Баума та Р.П. Томпсон на два незалежні кластери (рис. 2.2) і показано, що 15 книга потрапляє в область робіт Р. Томпсон. Для Ф. Баума характерними виявились слова *which, that, so, to*, а для Р. Томпсон – *up, on, down, over* (рис. 2.3) [139, p. 13].

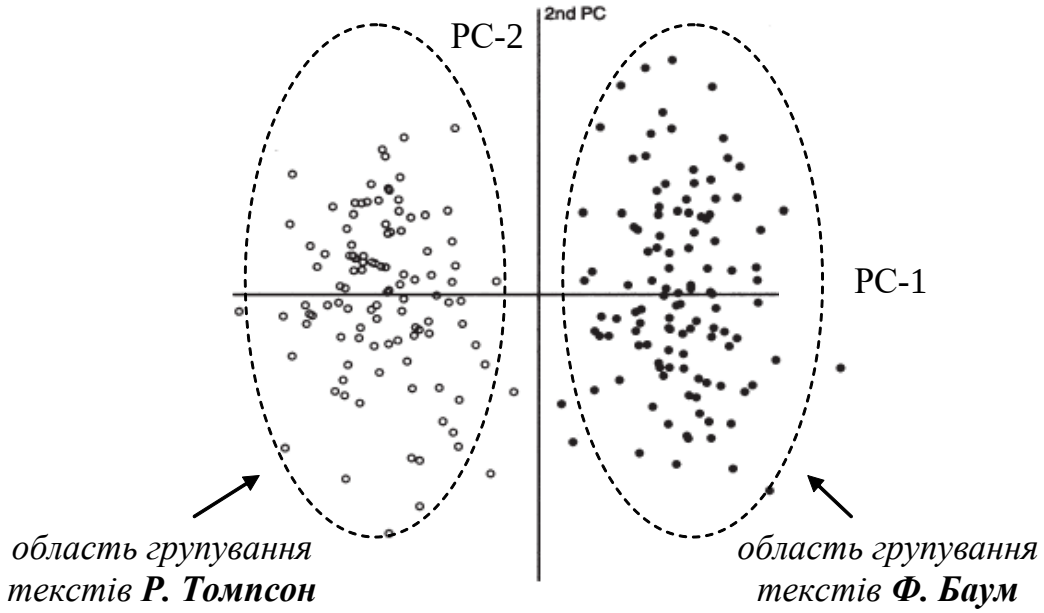


Рис. 2.2. Розподіл робіт Ф. Баума (замальований кружечок) та Р. Томпсон (незамальований кружечок) у просторі першої та другої головних компонент PC1 та PC2.

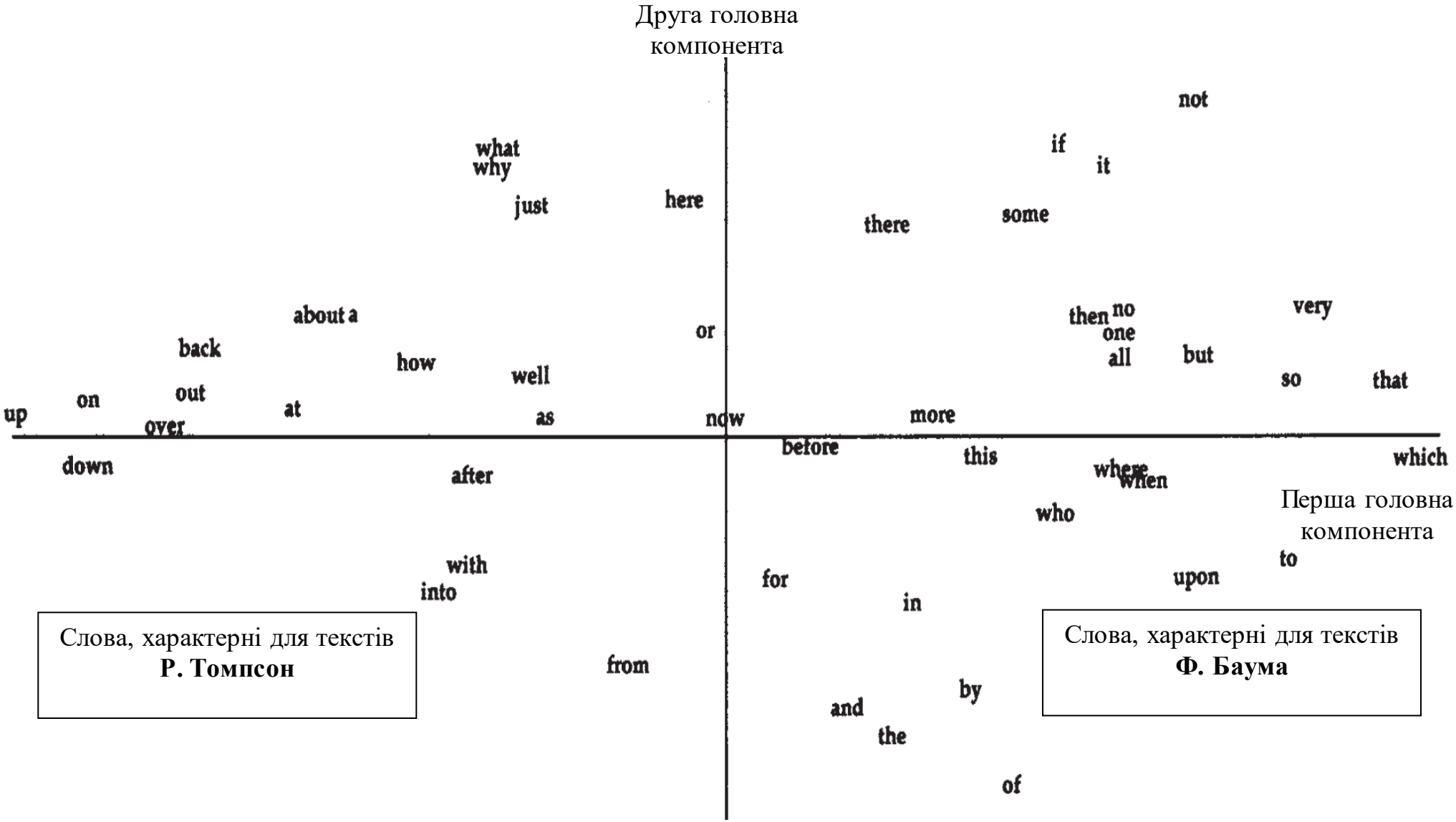


Рис. 2.3. Розподіл слів у творах Р. Томпсон та Ф. Баума у просторі першої та другої головних компонент.

У 1913 році дружина американського солдата Джорджа Піккета видала книгу “Серце солдата”, у якій зібрала листи свого чоловіка. З часом почали з’являтись сумніви, кому ж насправді належать ці листи – солдату громадянської війни чи його дружині, оскільки стиль викладу та зміст був занадто пишним та сентиментальним як для військового генерала. У дослідженні Девіда Холмса методом одночасного моніторингу групування текстів та відповідних їм слів, словосполучень показано, що ці листи належать перу його дружини [172, p. 21]. Це було виявлено шляхом аналізу 60 найчастіше вживаних службових слів.

Збірник політичних статей “Federalist Papers” використано як базу даних для проведення авторської атрибуції із залученням методу одночасного моніторингу групування текстів та відповідних їм слів, словосполучень [169; 214]. Відомо, що 5 статей написав Дж. Джем, 51 – належить А. Гамільтону, 14 – Дж. Меддісону, 3 – написані у співавторстві Гамільтона і Медісона, а 12 – підписані псевдонімом “Publius”. Аналізуючи спочатку тексти кожного автора окремо, потім порівнюючи їх між собою та з анонімними статтями, застосовуючи набори із 49 (*a, all, an, and, are, as, at, be, been, but, by, can, every, for, from, has, have, if, in, into, is, it, its, may, more, must, no, not, of, on, one, only, or, so, some, such, than, that, the, their, this, to, was, were, which, who, will, with, would*) та 30 (*according, also, although, always, an, apt, both, by, commonly, consequently, considerably, direction, enough, innovation, kind, language, matter, of, on, particularly, probability, there, this, though, to, upon, vigorous, while, whilst, work*) слів, Д. Холмс, так само як і Ф. Мостелер за допомогою методу одночасного моніторингу групування текстів та відповідних їм слів показав, що під псевдонімом “Publius” писав Дж. Медісон [169, p. 118–121].

Аналіз одночасного моніторингу групування текстів та відповідних їм слів, словосполучень апробовано також для дослідження авторського стилю американського письменника-журналіста Стефена Крейна на 17 статтях, авторство яких було невідомим, але їх приписували цьому автору [171,

p. 316–320]. Статті були написані в період з 1889 по 1892 рік у журналі “New York Tribune”. Матеріалом аналізу були журнальні статті та художні твори С. Крейна. Це дослідження проводили паралельно з використанням традиційних і нетрадиційних методів, тобто із залученням думки експертів і за допомогою методів математичного опрацювання даних. Метод одночасного моніторингу групування текстів та відповідних їм слів, словосполучень використано для кластеризації робіт на дві окремі групи за авторами: С. Крейном і Дж. Конрадом [там само, p. 323]. Статті із сумнівним авторством попали в групу статей С. Крейна. У процесі аналізу виявлено: найбільш характерні слова для кожного з цих авторів; відмінності стилю С. Крейна у написанні ним статей до журналів і створенні художніх творів; групування статей С.Крейна відповідно до видання, де вони були опубліковані.

Методом одночасного моніторингу групування текстів та відповідних їм слів, словосполучень досліджено 72 тексти на чітко задану тематику, що були написані не професійними письменниками, а 8 студентами, 4 з яких навчалися на першому курсі, решта – на четвертому. Кожен із них повинен був написати по три тексти у вигляді казки, переказу та твору на вільну тему з обсягом до 1000 слів [136]. Щоб дослідити розмежування статей за тематикою, автором і віком, аналізували вживання 50 найчастотніших службових слів, використовуючи метод одночасного моніторингу групування текстів та відповідних їм слів, словосполучень. Було показано, що цей метод є ефективним для розмежування текстів за тематикою та віком автора [там само, p. 32].

Існують різні погляди щодо придатності методу одночасного моніторингу групування текстів та відповідних їм слів, словосполучень для авторської атрибуції текстів. Так, Д. Хувер вказує, що точність цього методу при одночасному аналізі великої кількості авторів є недостатньо високою [174]. Він досліджував творчість 27 авторів на основі розгляду їх 50 новел та проаналізував 50 найчастіше вживаних службових слів. Ефективність розпізнавання авторів становила лише 25%, збільшення кількості найчастіше

вживаних слів не сприяло значному поліпшенню рівня ефективності: 100 – 22%, 200 – 45%. Проте, якщо зменшити кількість досліджуваних авторів до 10, рівень ефективності авторської атрибуції збільшується і досягає 70%.

Погіршення ефективності методу одночасного моніторингу групування текстів та відповідних їм слів, словосполучень можуть бути пояснені: 1) обмеженням кількості лінгвістичних параметрів (наприклад, аналізом лише службових слів або ж використанням тільки послідовності сполучення зі слова); 2) відсутністю нормування елементів частотної матриці; 3) розглядом результатів аналізу лише у двовимірному просторі перших двох компонент і не врахуванням можливості побудови моделі у системі координат інших головних компонент, відмінних від першої та другої головної компоненти. Саме на способи вдосконалення ефективності методу одночасного моніторингу групування текстів та відповідних їм слів, словосполучень для атрибуції текстів і звернена увага в цій дисертаційній праці. У дослідженні вперше для аналізу наукових та художніх текстів використовується поєднання лінгвістичного параметра послідовності вживання слів із такими методами, як метод одночасного моніторингу текстів та відповідних їм слів та метод ентропії.

2.3 Аналіз стильової та авторської атрибуції художніх текстів різних мов за допомогою програмних систем

Значна кількість програмних систем дозволяє проводити різного типу атрибуцію текстів. Нижче представлено короткий огляд найважливіших програм. Програму “Авторовед”, розроблену А. С. Романовим, використовують для ідентифікації авторів невідомих текстів, написаних російською мовою [99, с. 107]. Для аналізу програмі достатньо працювати з обсягом тексту в 20000 символів. Основні параметри, з якими працює програма, є частота окремих букв, знаків пунктуації, високочастотних триграмів букв [101].

Програма “Атрибутор” є лінгвістичним процесором для автоматичного порівняння і класифікації текстів відповідно до параметрів індивідуального авторського стилю прозових творів. Щоб провести атрибуцію тексту у програму потрібно ввести текст розміром не менше 20 Кб (близько 20 сторінок тексту). Програма проводить аналіз частоти три-грам букв. Під час аналізу до уваги не беруться власні назви й імена [43].

Експертна система “ВААЛ” проводить класифікацію документів, написаних російською та англійською мовами, проводить оцінку емоційного впливу фонетичної структури текстів та окремих слів на свідомість людини. Передбачена можливість генерування слів із заданими фоносемантичними характеристиками, а також задання характеристики бажаного впливу та цілеспрямованої корекції тексту з метою досягнення бажаного ефекту впливу. Програма дозволяє проводити аналіз словника текстів та контент-аналіз тексту [44].

Програма “Лингвоанализатор” Д. В. Хмельова, створена на основі бази 1357 творів 128 письменників, може встановити як приналежність введеного тексту перу одного з авторів зі списку бази даних, так і віднести довільний текст до стилю одного з авторів [40]. Програма працює з ланцюгами Маркова та ентропією текстових документів [189, p. 202]. Ця процедура реалізується на основі підрахунку таких характеристик тексту, як кількість службових слів, морфем та їх послідовності, складність граматичних конструкцій з урахуванням словника, який використовував автор.

“Лингвистический анализатор”, над яким працював А. Львов, аналізує тексти російських письменників-фантастів, проводячи підрахунок середньої довжини речення, 2-грам букв, частин мови, унікальних слів, активний словниковий запас та вживання вигаданих слів, а також працює з середньоквадратичним відхиленням від досліджуваних параметрів [37].

Інформаційна система “СМАЛТ” (Статистичні Методи Аналізу Літературного Тексту), розроблена в Петрозаводському державному університеті, проводить аналіз літературних творів на основі їх

морфологічного та синтаксичного аналізу, використовуючи при цьому метод Колмогорова-Смирнова, кластерний аналіз, критерій Стьюдента [97; 98; 42].

“СтилеАнализатор”, розроблений О. Шевельовим, удосконалення якого зараз відбувається, працює з таким комплексом методів, як метод хі-квадрат та мір подібності Кульбака, ієрархічний кластерний аналіз, дерева рішень, ентропійний метод, нейронні мережі, факторний лінійний та нелінійний аналіз [121].

Програму “Delta”, розроблену Й. Бероузом, застосовують для авторської та стильової атрибуції англомовних текстів [148]. Міра подібності текстів визначається на основі оцінки вкладу груп слів, характерних для попередньо визначених опорних текстів [134, p. 132; 177, p. 455].

Програма “JGAAP” (Java Graphical Authorship Attribute Program), розроблена під керівництвом П. Юоли для проведення атрибуції текстів, працює з такими методами, як метод найближчих сусідів, метод аналізу головних компонент, лінійний дискримінантний аналіз, машина опорних векторів, байєсівський аналіз [185, p. 272–286; 38]. Спершу цю програму розробляли для проведення авторської атрибуції тексту, проте вона може успішно виконувати такі завдання, як класифікація документів за періодом написання та жанром.

“Signature Stylometric System” була розроблена П. Міліканом, щоб полегшити аналіз та порівняння текстів за авторським стилем [39]. Ця програма проводить підрахунок букв, знаків пунктуації та довжин слова, речення, параграфа, використовуючи для подальшого обчислення метод аналізу головних компонент та алгоритми програми “Delta”.

Об’єднує наведені вище програми те, що майже всі вони працюють з методами, що потребують попереднього навчання системи. Також можна зауважити, що серед цих розробок відсутнє програмне забезпечення для авторської атрибуції текстів, написаних українською мовою. Більшість з розглянутих стилеметричних програм є реалізацією конкретних методів і не дозволяють користувачеві робити зміни параметрів у ході опрацювання

даних. Розглянуті програми працюють з текстами художньої літератури. Можливість атрибуції вузькоспеціалізованих наукових текстів у них не розглядається. Цей факт засвідчує актуальність і новизну дисертаційного дослідження у проведенні атрибуції англомовних наукових текстів. Передумовою створення власного програмного забезпечення для атрибуції наукових текстів є те, що жодна з розглянутих програм не дозволяє користувачеві вводити до наявних програм зміни, зокрема, додавання набору нових параметрів або нових методів. Досить часто, працюючи з готовим програмним продуктом, ми не знаємо, які додаткові функції додані до нього, що може також мати вплив на ефективність атрибуції текстів.

2.4 Комплексна методика аналізу стильової, тематичної та авторської атрибуції текстів

Для проведення атрибуції тексту можна запропонувати **схему етапів опрацювання текстів,** наведену на рис. 2.4. Залежно від виду атрибуції буде визначатись відповідна вибірка текстів і подальше їх опрацювання.

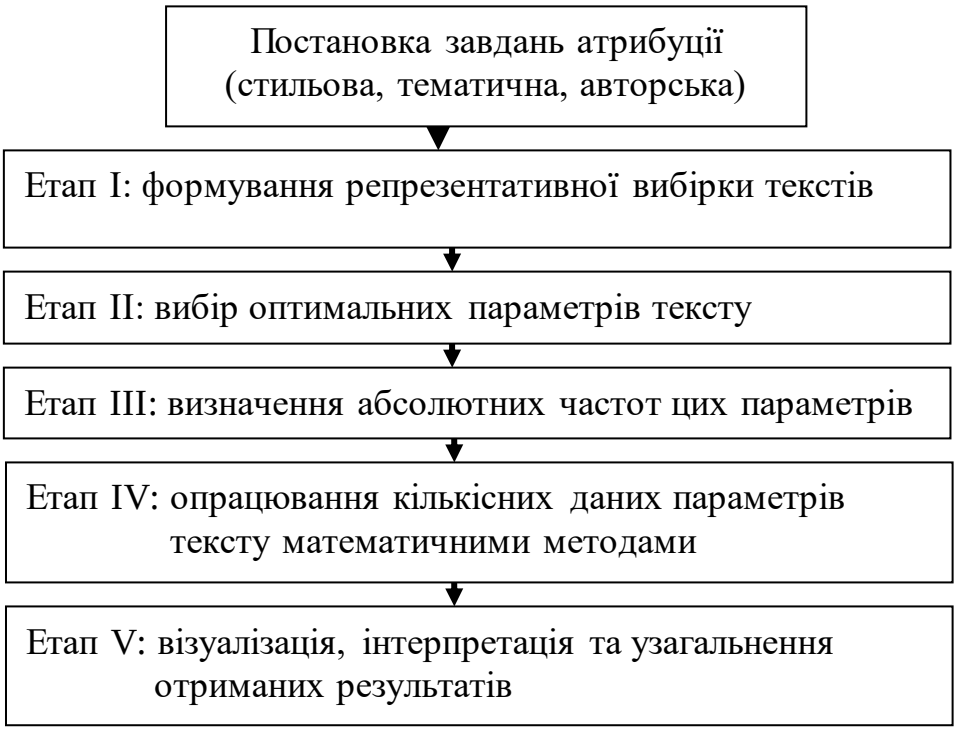


Рис. 2.4. Основні етапи методики аналізу текстів для їх атрибуції.

Етап І. У випадку стильової атрибуції текстів методом частотного розподілу слів (закон Ципфа) сформовано суцільні вибірки текстів наукової та художньої літератури. Для атрибуції текстів наукового стилю підібрано англо-, німецько- та украномовні наукові тексти різних жанрів (монографія, стаття, дисертація) з галузі фізики, щоб простежити за поведінкою апроксимаційної кривої рангово-частотного розподілу слів залежно від тематики та жанру наукових текстів. Для проведення тематичної атрибуції текстів вибрано праці наукової конференції. Відомо, що тематика конференційних секцій є заздалегідь визначеною. Ця обставина дозволяє полегшити процедуру визначення ефективності кластеризації текстів у відповідні групи. Серед наукової спільноти, яка досліджує люмінесцентні процеси, однією з авторитетних конференцій є Міжнародна конференція LUMDETR "Luminescent Detectors and Transformers of Ionizing Radiation" (Люмінесцентні детектори та перетворювачі іонізуючого випромінювання), матеріали якої друкують у збірнику тез конференції та статей у журналі "Radiation Measurements", імпакт-фактор якого – 0,97.

З метою авторської атрибуції текстів розглянуто вузькоспеціалізовані наукові статті в галузі люмінесцентної спектроскопії та люмінесцентного матеріалознавства. З даної галузі знань найбільш цитованими є роботи європейських вчених: проф. д-р. Pieter Dorenbos, проф. д-р. Andries Meijerink, д-р. Gregory Stryganyuk, проф. д-р. Georg Zimmerer. Так, за даними інформаційної системи Sciencedirect [41] видавництва Elsvier кількість посилань на роботи P. Dorenbos сягає 1635, A. Meijerink – 3427, G. Zimmerer – 1677. Для аналізу вибрано також роботи д-ра Г. Стриганюка та д-ра Ю. Зоренка, які працюють у цій же галузі знань, однак, кількість цитувань їх робіт сягає 146 та 106 відповідно. Критерієм вибору згаданих авторів була не лише подібна тематика досліджень, але і їх спільна праця в Міжнародному науковому центрі HASYLAB (Синхротронна лабораторія, м. Гамбург, Німеччина). Підбір текстів з однаковою тематикою та експериментальними методами дослідження передбачає, що розмежування цих

вузькоспеціалізованих текстів буде швидше авторським, ніж тематичним. Вибір статей, які аналізувалися, та їх особливості були обговорені з авторами. Це дозволило перевірити ефективність авторської атрибуції, проведеної в цій дисертації, навіть для статей, написаних у співавторстві.

До текстів художньої літератури увійшли твори класиків світової художньої літератури XIX-XXI століття, мова оригіналу яких англійська, німецька, українська, так і твори, перекладені англійською мовою. З текстів цих творів були створені вибірки, що містили:

1) тексти, написані носіями англійської, німецької, української мов;

2) тексти, перекладені з російської мови англійською;

3) об'єднання текстів, написаних носіями англійської мови, та текстів, перекладених англійською мовою;

4) тексти XIX століття або тексти XXI ст.

Таке розмаїття вибірок створено, щоб дослідити залежність нахилу апроксимаційної кривої рангово-частотного розподілу слів (закон Ципфа) від згаданих відмінностей художніх текстів. Щоб дослідити залежність параметрів рангового розподілу слів від мови тексту, було створено вибірки англо-, німецько- та україномовних наукових і художніх текстів.

Етап II. Серед багатоманіття параметрів для проведення атрибуції текстів (табл. 1.1.) у цій дисертаційній праці вибрано лінгвістичний параметр послідовності сполучуваності слів, словосполучень (n-грам слів). Як згадувалося, саме цей параметр є ефективним для авторської атрибуції англомовних художніх текстів [163].

Етап III – визначення абсолютних частот цих параметрів. У більшості випадків формувалась матриця параметрів частот, де кожен елемент матриці відповідав частоті певного слова в певному тексті. Визначення абсолютних частот та формування матриці здійснено за допомогою програми "Lexical Content Searcher", розробленої у цій дисертаційній праці. Програма "Lexical Content Searcher" (Додаток А, рис. 1) дозволяє шукати всі варіанти

послідовності вживання одного і більше слів (n-грам слів) та здійснювати підрахунок їх частот у масиві текстів.

Ця програма аналізує варіанти послідовності вживання одного і більше слів як “графічні слова”. Це означає, що програма забезпечує статистичну інформацію про слово, але відсутня інформація про нього як граматичну одиницю мови, не проводиться лематизація слів, не враховуються синонімічні зв’язки [119, с. 123]. Саме за принципом “графічних слів” укладено словники Ф. Достоєвського “Статистический словарь языка Ф. Достоевского” [120]. Виділення “графічних слів” у тексті задається правилами орфографії і не залежить від фонологічної, морфологічної, граматичної та семантичної системи мови. У подальшому замість терміна “графічне слово” використано термін “слово”. Програма дозволяє отримати частотні словники n-грам слів для кожного тексту окремо, а також узагальнений частотний словник n-грам слів для усього масиву текстів. Результатом роботи програми є таблиця частот усіх слів, що зустрічалися у заданому наборі текстів, яка містить інформацію про кількість входжень кожного слова у кожний текст, а також кількість текстів, в яких зустрілося дане слово. Окрім певних слів, програма може визначати частоти послідовності вживання одного і більше слів. Кількість слів у *n*-грамі користувач може задати до початку роботи програми. Для простоти опису надалі будемо розглядати процедури розрахунку статистики окремих слів, оскільки технологія підрахунку частот входження для *n*-грам є ідентичною.

Робота з групою текстів технічно реалізована так. Користувач до початку розрахунків повинен вказати ім’я (із повним шляхом доступу на диску) одного із тестових файлів аналізованої групи. Програма аналізуватиме всі тексти (всі файли з розширенням “txt”), які знаходяться в тій же директорії, що й файл, заданий користувачем. Результат роботи програма виводить у текстовий файл “table.txt” табличного виду, перший рядок якого містить назви аналізованих файлів, а перша колонка – знайдені в них слова, наступні колонки містять інформацію про частоти слів у текстах. Файл із

отриманими результатами створюється самою програмою та записується у директорію “!Processed”, яка також створюється автоматично.

Схема роботи програми така. Після запуску програма проводить аналіз вмісту заданої директорії та складає список назв усіх текстових файлів у ній і зберігає його у вигляді одновимірної таблиці **A**. Тип даних кожного її елемента – стрічковий. Далі починається збір статистичних даних для кожного із файлів. Для цього почергово кожен із текстових файлів відкривається, а його дані зчитуються в оперативну пам’ять ЕОМ у вигляді послідовності символів та аналізуються. Робота програми продовжується, доки не будуть опрацьовані всі файли, назви яких наявні у таблиці **A**.

Аналіз кожного текстового файлу починається з розбиття його на речення. Реченням програма вважає набір символів, який завершується одним зі знаків, – крапкою, знаком оклику, знаком запитання тощо. Виявлені в тексті речення зберігаються програмою у вигляді одновимірної динамічної таблиці, кожен елемент якої є стрічковим. Після заповнення таблиці у кожному реченні проводиться пошук слів (чи *n*-грам). Словом програма вважає неперервний набір букв, який не може містити жодного із символів пунктуації, пробілів, символу табуляції, лапок, дужок, керуючих символів, таких, як символ закінчення стрічки та символ переводу каретки й інші.

Знайдені слова програма заносить у динамічну таблицю **B**. Одночасно створюється динамічна таблиця **C** (елементи якої мають цілочисельний тип даних), яка міститиме частоти знайдених слів в аналізованих текстах. По горизонталі у таблиці **C** вказуються назви текстів, по вертикалі – слова, таким чином, що елемент, наприклад, **C**[3,5] буде містити частоту 5-го слова таблиці **B** у третьому тексті згідно з таблицею **A.** У процесі роботи програми розмір таблиці **C** збільшується: при знаходженні нового слова до неї додається рядок даних (частоти цього слова для кожного тексту), а для опрацювання кожного нового тексту додається колонка (частоти всіх виявлених слів для цього тексту). При додаванні рядків та колонок усі нові елементи таблиці **C** ініціалізуються нулями. Процес підрахунку частот та

створення таблиці слів **B** є наступним. Кожне знайдене слово програма зіставляє з усіма словами, занесеними у таблицю **B** раніше. Якщо не виявлено збігу, то слово додається у таблицю **B**, а відповідний цьому слову елемент таблиці **C** встановлюється рівним одиниці. Якщо ж збіг виявлено, то відповідний елемент таблиці **C** збільшується на одиницю. Треба зауважити, що для уникнення повторів одного й того ж слова, які можуть виникнути через різний регістр букв, усі букви слів програма переводить у нижній регістр. Із параметрів таблиць **A**, **B** та **C** програма формує файл результату.

Програма була апробована для порівняння завдань і цілей прикладної лінгвістики в Україні та за кордоном [20]. Проведено кількісний аналіз найбільш уживаних слів у журналі “The Journal of Applied Linguistics” [245] та монографії М. М. Пещак “Нариси з комп’ютерної лінгвістики” [90]. Незважаючи на однакове фахове спрямування досліджуваних матеріалів, ключові слова текстів виявились різними. Аналіз перших 100 ключових слів (за винятком службових слів) показує, що науковці в Україні у своїх роботах працюють над різними проблемами лінгвістики. Наукові статті журналу “The Journal of Applied Linguistics” присвячені вивченню та розвитку лінгвістичних методів для полегшення вивчення іноземних мов. У монографії “Нариси з комп’ютерної лінгвістики” автор віддає перевагу комп’ютерним методам вивчення мов, укладанню словників.

Етап IV. Важливим етапом є підбір відповідних методів та алгоритмів, які забезпечують успішну атрибуцію текстів. Враховуючи переваги та недоліки відбору лінгвістичних параметрів тексту (див. Розділ 1.2 – 1.4) та математичних методів для атрибуції документів (див. Розділ 2.1), доступність програмного забезпечення і машинного ресурсу (див. Розділ 2.3), у цій дисертації запропоновано методи розв’язання завдань атрибуції текстів відповідно до рис. 2.5. Тексти аналізували із залученням ентропійного методу й методу одночасного моніторингу групування текстів та відповідних їм слів, які дозволяють групувати, порівнювати тексти, використовуючи як

вхідні параметри увесь масив даних тексту без попереднього виділення найчастотніших слів, що характерні певній тематиці, стилю чи автору.

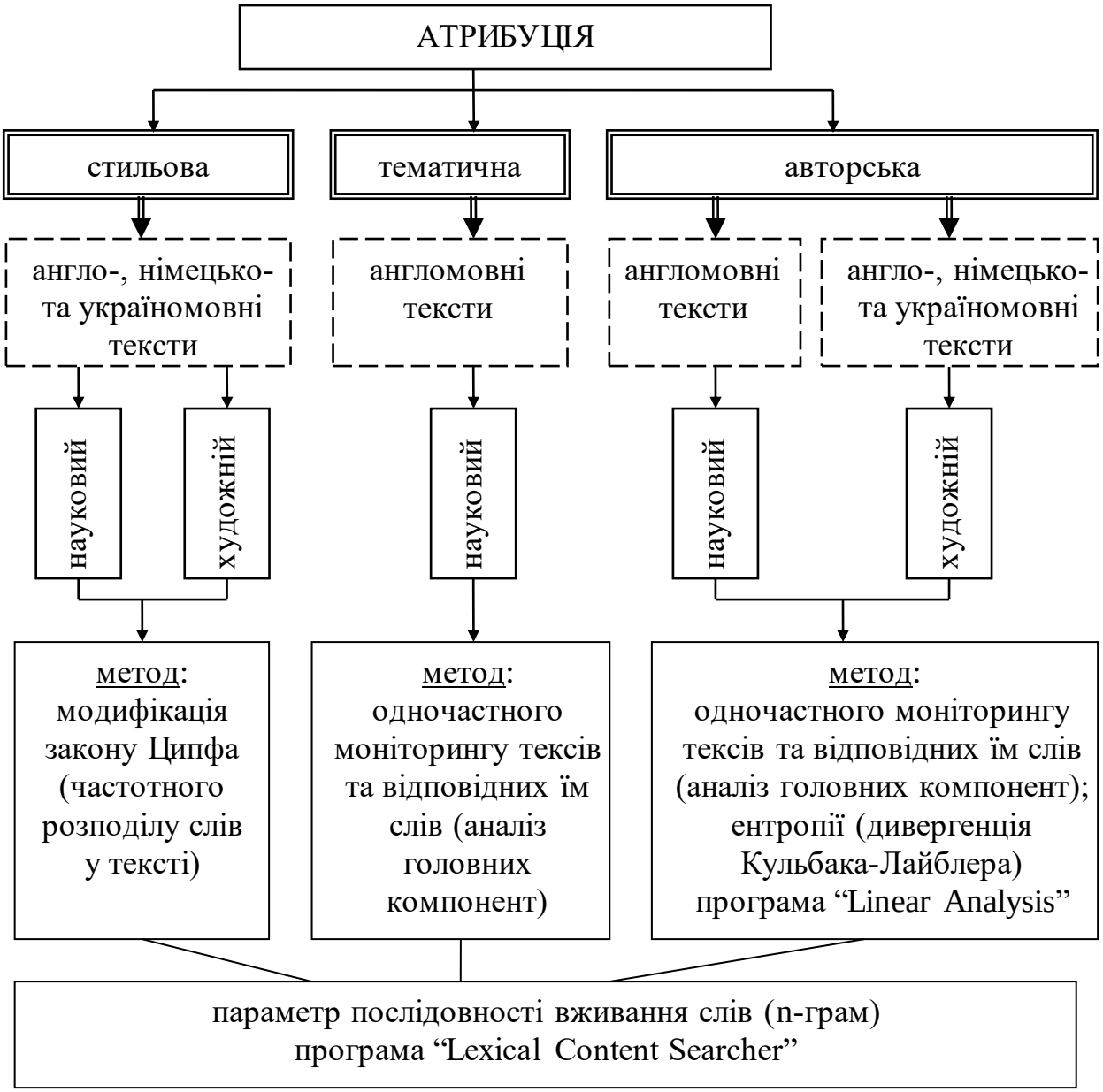


Рис. 2.5. Завдання атрибуції та методи їх розв’язання, використані у дисертаційній роботі.

Стильову атрибуцію наукових та художніх англо-, німецько та україномовних текстів проведено з використанням однієї з модифікацій рангово-частотного закону Ципфа – модифікованої у цій дисертації функції Лавалетті. Тематичну атрибуцію, яка полягала у виявленні відмінностей тематики конференційних секцій, у визначенні тематичної близькості між

заголовком, анотацією та статтею (тріади “заголовок–анотація–стаття”), здійснено методом аналізу одночасного моніторингу групування текстів та відповідних їм слів із залученням параметра послідовності вживання слів. Вибір тріади “заголовок–анотація–стаття” зумовлений тим, що концентрація ключових слів суттєво збільшується при переході від статті до заголовка. Авторська атрибуція у цій дисертаційній праці спрямована на вирізнення статей одного автора серед інших або на одночасне розпізнавання наукових текстів чотирьох авторів. Основними текстовими параметром для проведення тематичної та авторської атрибуції був параметр послідовності вживання з одного та більше слів. Опрацювання кількісних даних текстових параметрів проведена методом одночасного моніторингу групування текстів та відповідних їм слів, методом хі-квадрат та методом ентропії. Авторська атрибуція художніх текстів здійснена із залученням методу одночасного моніторингу групування текстів та відповідних їм слів.

На відміну від методу аналізу головних компонент, який функціонує в багатовимірному просторі, метод ентропії (дивергенції Кульбака-Лайблера) [187] зводить характеристики текста до його інформативності, мірою якої виступає ентропія (див. формулу 2.2).

Для розрахунку дивергенції текстів за формулою Кульбака-Лайблера (формула 2.2) використано опцію розробленої нами програми “LinearAnalisys” (Додаток А, рис. 2). Вхідні дані задавались як частотна матриця текстів, вказувалась назва опорного файлу (тексту) і, за необхідності, файл словника, за яким необхідно проводити аналіз. Це дозволяє задати словник автора, якщо авторство жодного з текстів не є відомим. Якщо ж така множина не була вказана, то програма приймала для аналізу словник опорного тексту. Опорний текст – довільний текст, з яким порівнюють інші тексти, що класифікують. Задання малого значення параметра “Small parameter value” (μ=0.0005) передбачено в програмі для уникнення ділення на нуль у випадку відсутності слова у тексті. Результат обчислень виводився як таблиця значень дивергенції текстів, обчисленої

відносно опорного тексту, або ж заданого словника. Програма “LinearAnalisys” оперує з вхідними даними у вигляді двовимірної таблиці частот слів, що відповідають досліджуваним текстам. У таблиці фігурують частоти усіх слів, які зустрічались у текстах, причому слова виведено в рядки, а назви текстів – у стовпчики.

Результатом розрахунків програми є дивергенція (“розбіжність”) даного тексту у відношенні до якогось конкретного, вибраного користувачем як опорного. “Розбіжність” розраховують за модифікованою формулою Кульбака-Лайблера (2.9):

$$KLD_i(Q_1, Q_i) = \sum_{j=1}^{k} B_{1j} \times \log_2\left(\frac{B_{1j}}{B_{ij}}\right), \text{ де} \qquad (2.9)$$

$$B_{ij} = \frac{f_{ij}}{\mu + d_i} + \frac{\mu}{\mu + d_i} P_j, \qquad (2.10)$$

$$P_j = \frac{\sum_{i=1}^{n} \frac{f_{ij}}{d_i}}{n}, \qquad (2.11)$$

f_{ij} – частота j-го слова у i-му тексті; μ – параметр згладжування; d_i – загальна кількість слів в i-му тексті; n – кількість текстів; P_j – імовірність появи слова в усіх текстах; Q_1 – множина слів опорного тексту; Q_i – множина слів i-го тексту.

Програма дозволяє розраховувати “розбіжність” текстів за ключовими словами. У ролі ключових слів можуть бути як слова, що зустрілись в опорному тексті, так і слова, задані користувачем як додаткові вхідні дані програми у вигляді текстового файлу із набором необхідних слів. Результати розрахунків програма виводить у текстовий файл (Додаток А, рис. 3).

Етап V. Візуалізація результатів атрибуції методом аналізу головних компонент проведена шляхом відображення розподілу текстів графічно у просторі головних компонент. Результати оцінки параметрів рангово-ймовірнісного розподілу Ципфа для наукової та художньої літератури представлено експериментальною та апроксимаційною кривими розподілу за

допомогою графічних засобів пакету Origin-8. Параметри апроксимації та коефіцієнт відтворення, який характеризує якість апроксимації, представлено чисельно на рисунках або в таблицях. Результати оцінки близькості текстів методом дивергенції Кульбака-Лайблера було представлено в графічному вікні Origin-8 зі значенням дивергенції, відсортованої за зростанням, що дало можливість відобразити прояв характеристик опорного тексту у наборі проаналізованих текстів.

Висновки до розділу 2

У цьому розділі здійснено аналіз широко застосовуваних математичних методів атрибуції текстів, з-поміж яких: частотний розподіл слів у тексті (закон Ципфа), штучні нейронні мережі, дерева рішень, дискримінантний аналіз, метод машини опорних векторів, ентропійні методи, кластерний аналіз, метод одночасного моніторингу групування текстів та відповідних їм слів. Попри значні зусилля, які дослідники докладають для розв'язання завдань атрибуції, на сьогодні питання ефективних методів атрибуції текстів залишається все ще відкритим.

Існує низка програмних продуктів, що використовують для атрибуції текстів: "Авторовед", "Атрибутор", "ВААЛ", "Лингвоанализатор", "Лингвистический анализатор", "СМАЛТ", "Delta", "JGAAP", "Signature Stylometric System", більшість із яких не передбачають доступу користувача до зміни параметрів опрацювання даних, що не дозволяє використовувати їх як інструмент наукового дослідження. Наведені вище програми працюють із текстами художньої літератури. Можливість атрибуції наукових текстів у них не розглядається. Це зумовлює необхідність розробки власних програм для атрибуції наукових текстів.

Атрибуція наукових і художніх англо-, німецько- та україномовних текстів у цій дисертаційній праці передбачає: постановку завдання атрибуції (стильова, тематична або авторська), формування репрезентативної вибірки

текстів (наукові та художні), вибір параметрів тексту (параметр послідовності сполучуваності з одного та більше слів), визначення абсолютних частот цих параметрів (програма “Lexical Content Searcher”), опрацювання статистичних даних параметрів тексту математичними методами (закон Ципфа, метод одночасного моніторингу групування текстів та відповідних їм слів, метод ентропії), візуалізацію отриманих результатів.

Можливість одночасного проведення атрибуції текстів та автоматичного виділення семантичних груп слів, зіставлення слів із ознаками текстів зумовила вибір методу одночасного моніторингу групування текстів та відповідних їм слів (аналізу головних компонент) для тематичної й авторської атрибуції текстів. Головна компонента об’єднує слова з високо корельованими ознаками в одну групу: кожній групі слів ставиться у відповідність своя головна компонента. Наявність вибірки слів для кожної компоненти дозволяє проводити її аналіз і робити висновок про відповідність головної компоненти певній ознаці, характеристиці тексту (у нашому випадку тематиці або автору). Метод є особливо актуальним, якщо взяти до уваги, що він не потребує попереднього навчання системи та наперед заданих критеріїв атрибуції. Поряд із методом одночасного моніторингу групування текстів та відповідних їм слів, груп слів для проведення тематичної та авторської атрибуції текстів пропонується метод ентропії (програма “LinearAnalisys”) як більш доступний для програмування порівняно з методом одночасного моніторингу групування текстів та відповідних їм слів, груп слів. Метод ентропії (дивергенції Кульбака-Лайблера) базується на порівнянні ентропії досліджуваних текстів з опорним текстом, автор якого відомий.

Розроблено програму “Lexical Content Searcher”, яка здійснює пошук усіх варіантів n-грам слів (послідовності вживання з одного і більше слів) та підрахунок їх абсолютних та відносних частот у масиві текстів. Програма дає змогу отримати частотні словники послідовності вживання одного і більше слів для кожного тексту окремо, а також узагальнений частотний словник

слів, словосполучень для усього масиву текстів. Програма аналізує слова як “графічні слова”, тобто відсутня інформація про нього як граматичну одиницю мови, не проводиться лематизація слів, не враховуються синонімічні зв’язки. Для розрахунку ентропії текстів за формулою дивергенції Кульбака-Лайблера розроблено програму “LinearAnalisys”. Для роботи програми слід вказати назву опорного файлу (тексту) і за необхідності файл словника, за яким варто проводити аналіз. Таким чином, передбачено можливість задання словника автора, коли авторство жодного з текстів не було відомим. Опорний текст – довільний текст, з яким порівнюють інші тексти, що класифікують. Результат обчислень виводився як таблиця значень дивергенції текстів, обчисленої відносно опорного тексту, або ж заданого словника.

Основні положення розділу висвітлено у працях автора [20; 251; 252].

РОЗДІЛ 3
СТИЛЬОВА АТРИБУЦІЯ НАУКОВИХ І ХУДОЖНІХ АНГЛО-, НІМЕЦЬКО- ТА УКРАЇНОМОВНИХ ТЕКСТІВ

3.1 Закономірності частотного розподілу слів у наукових та художніх текстах

На початку 1930-х рр. Джордж К. Ципф запропонував емпіричний закон розподілу слів природної мови [261]. Насправді, важко визначити авторство ідеї, яка лягла в основу емпіричного закону Ципфа, адже початки рангового розподілу Ципфа були закладені ще до 1912 року Еступом (Jean-Baptiste Estoup), який аналізував частоту появи графем у стенографії [155]. Згідно з ранговим законом Ципфа (відомим у мовознавчій літературі як закон Зіпфа) імовірність f появи слова в тексті є обернено пропорційною до рангу слова k у списку n слів, упорядкованих у порядку зменшення їх імовірності:

$$f(k;s,n) = k^{-s} \Big/ \sum_{i=1}^{n} i^{-s} = \frac{c}{k^s}, \qquad (3.1)$$

де показник ступеня s дорівнює 1 у випадку класичного рангового розподілу Ципфа, c – константа, n – обсяг словника. Ранг слова (k) є порядковим номером слова у списку слів, що зустрічаються у тексті, або ж наборі текстів.

Суттєвим є те, що ранговий розподіл Ципфа має статистичний характер, який виявляється за умови, що обсяг словника тексту є достатньо великим. Виконання закону Ципфа можна використовувати як доказ того, що розмір тексту (словник тексту) є достатнім для прояву статистичних закономірностей.

Загалом, ранговий розподіл Ципфа, побудований у логарифмічній системі координат, можна розділити на три ділянки: перша ділянка стосується слів у діапазоні малих значень рангового числа k (так званий лівий хвіст, де розташовані слова з високою частотою появи у тексті); прямолінійна ділянка з нахилом $s \approx 1$ та ділянка з високими значеннями

рангового числа *k* (так званий правий хвіст, де розташовані низькочастотні слова). Дослідження показали, що слова з максимальною частотою (лівий хвіст), як правило, – це прийменники, частки, займенники, артиклі. Прямолінійну ділянку формують найбільш вагомі слова в тексті, які можуть бути ключовими для проведення інформаційного пошуку. Ділянка з високими значеннями рангового числа *k* – це слова, що зрідка зустрічаються [113, с. 69]. Ще одна цікава особливість трактування закону Ципфа пов'язана з можливістю використання його для інтерпретації часу появи слів. Припускається, що слова правого хвоста – це нові слова, а лівого хвоста – старі слова [там само, с. 158]. Розбиваючи криву Ципфа на ділянки і досліджуючи тексти різних часових зрізів за зміною рангів виділених груп слів, можна простежити часову динаміку цих слів [там само, с. 159].

Традиційно вважалося, що кількісні показники ступеня *s* є однаковим для різних мов і не залежить від таких факторів, як автор, жанр, час написання твору тощо. Однак, більш детальні дослідження виявили, що нахил прямолінійної ділянки кривої Ципфа (закономірності рангово-частотного розподілу слів) не є однаковим для мов різних груп людей (здорових і шизофреників) [253, p. 10]. Нахил кривої Ципфа змінюється також залежно від віку дитини. Для дітей дошкільного віку вона є більш пологою, ніж у дітей шкільного віку [111, с. 106]. Певні відмінності в нахилі були знайдені для текстів з різних галузей природничих наук, наприклад, статті з математики описувались прямою із s=1,00, з екології s=0,89, з фізики s=0,94, з фізіології s=0,92 [253, p. 50]. Більшого успіху в ідентифікації текстів було досягнуто, коли дослідники звернули увагу на особливості рангово-частотного розподілу слів відповідно до закону Ципфа в діапазоні малих значення рангу $k<200$ та великих – $k>5000$. Так, для $k>5000$ у художніх текстах $s\approx2$. Ця обставина дозволила висловити припущення про можливість використання закону Ципфа для проведення класифікації текстів [213, p. 570–572].

З огляду на незначні відмінності значення параметра s для різного типу текстів становить інтерес апроксимація закону Ципфа формулою, параметри якої більш суттєво реагували б на зміни параметрів тексту.

Незважаючи на велику кількість спроб модифікувати ранговий закон Ципфа, не вдалося підібрати оптимальну функцію для опису розподілу слів у текстах. Більшість із запропонованих функцій розподілу передбачають два і більше апроксимаційних параметри [113, с. 77–84]. Однак, поліпшення якості відтворення рангового розподілу слів все ще залишається бажаним [158; 213; 216; 225]. Закономірними є спроби розвинути та вдосконалити вже існуючі підходи [3; 21; 83; 113; 225; 238; 252].

Треба відзначити, що при побудові рангового розподілу слів вигідніше працювати не з частотою слова, а з імовірністю його появи у тексті, оскільки частота слова буде залежати від кількості слів (N) у тексті. Нехтування цією особливістю може призвести до відхилення між кривими раногово-частотного розподілу слів навіть у споріднених за тематикою текстах [201, p. 92]. Щоб уникнути впливу цієї особливості на значення апроксимаційних параметрів, розрахунки в цій дисертаційній праці проведено, оперуючи імовірністю (f) появи слова у тексті, яка обчислюється як частота появи слова у тексті, розділена на загальну кількість слів (N) у тексті. Сума імовірностей появи слів у тексті дорівнює одиниці.

Відомо, що вигляд кривої рангово-частотного розподілу слів є різним для аналітичних та синтетичних мов: чим більше мова має форм слів, формотворчих афіксів, флексій (синтетична мова), тим пологішою стає крива рангового розподілу слів [225, p. 714]. У цій дисертації всі обчислення виконано для текстів, написаних англійською, німецькою та українською мовами, з метою розробки оптимальної функції для інтерпретації параметрів рангово-частотного розподілу слів. Апроксимація рангового розподілу слів у текстах відповідно до вибраного закону проведена ітераційним методом шляхом пошуку оптимальних значень параметрів розподілу, при яких досягалось мінімальне значення суми квадратів відхилень апроксимаційної

кривої від експериментальних значень. Апроксимація та графічне представлення результатів у цій праці проводились із використанням програми Microcal Origin 8.

На рис. 3.1. суцільною кривою зображена апроксимація рангово-ймовірнісного розподілу слів художнього твору Ш. Бронте “Джейн Ейр”, використовуючи функцію розподілу Ципфа $f(k; s, n) = f(k; 1.06, 12682)$. Як видно, ця апроксимація не забезпечує точного відтворення розподілу слів для всіх рангів k. Можна виділити прямолінійну ділянку з параметром s=1.06, яка узгоджується з експериментальними даними в діапазоні слів із середнім рангом k, проте значні відхилення спостерігаються для малих та великих значень рангу k. Єдиний апроксимаційний параметр (показник степеня s) у функції розподілу Ципфа (формула 3.1) описує лише кут нахилу кривої підгонки в логарифмічній системі координат.

Дотепер відсутня єдина теоретично розроблена модель, яка б одночасно забезпечила опис рангового розподілу слів у діапазонах малих, середніх та великих рангів слів. Натомість існує багато модифікацій та адаптованих до певних Ципфо-подібних функцій, які забезпечують задовільну якість апроксимації в окремих випадках, однак не забезпечують прийнятних результатів для більшості інших випадків. Слід звернути увагу, що всі Ципфо-подібні формули, які не були виведені теоретично, можна використовувати з метою опису або ж певного роду зіставного аналізу лише для окремих випадків. Відсутність теоретичного підґрунтя ускладнює виявлення кореляції характеристик тексту з параметрами емпірично запропонованої моделі.

Теоретично обґрунтовані моделі Ципфо-подібного розподілу можна розділити на два загальні класи [225, p. 714]:

1. Моделі типу “Мандельброта” – використовують підхід, в якому припускають, що статистична структура тексту визначається статистичною організацією мови, зокрема, принципами оптимізації затрат для продукування слів.

2. Моделі типу “Саймона” – використовують підхід, згідно з яким припускають, що спостережуваний розподіл слів є результатом когнітивних процесів, що впливають на генерацію тексту.

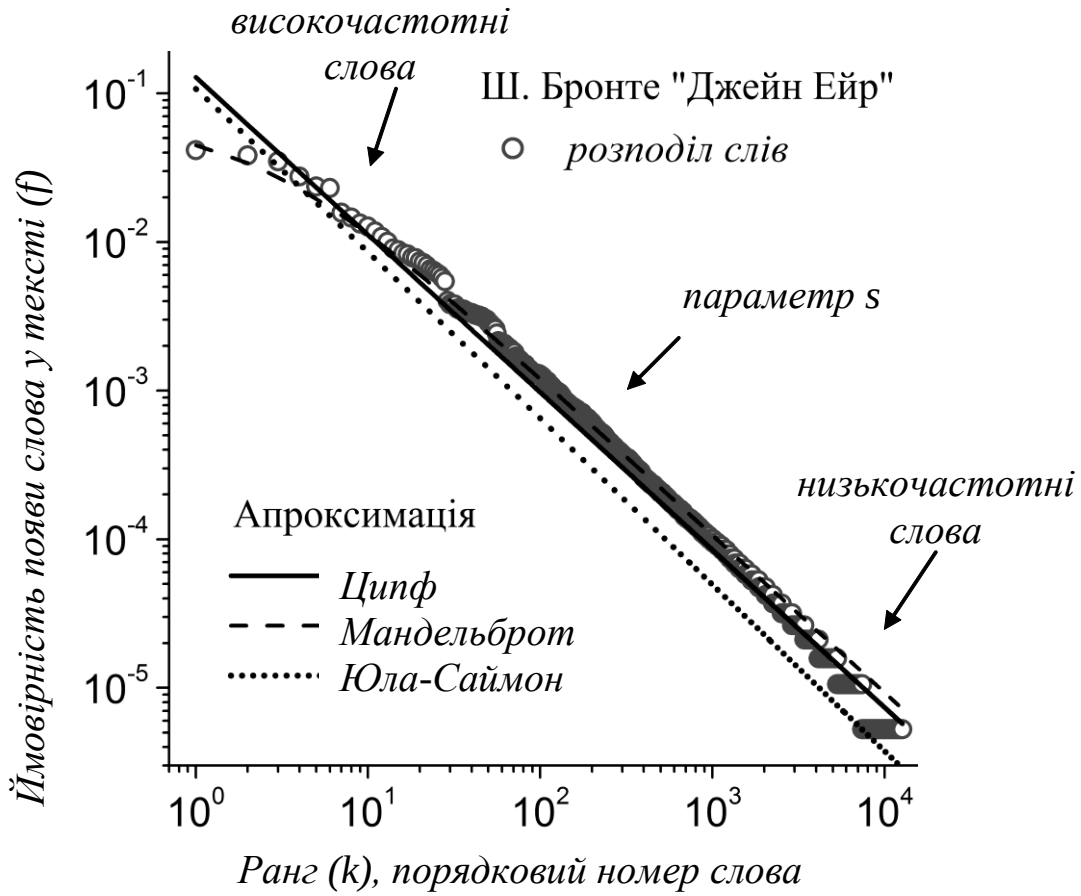


Рис. 3.1. Результати апроксимації рангово-ймовірнісного розподілу слів (Ш. Бронте “Джейн Ейр”) з використанням закону Ципфа та його модифікацій, запропонованих Мандельбротом, Юлом та Саймоном.

Математик Бенуа Мандельброт узагальнив закон Ципфа та вдосконалив функцію опису рангового розподілу слів, ввівши додатковий параметр q, що покращує результати апроксимації в діапазоні малих значень рангу k (слова з високою частотою вживання у тексті) [206]:

$$f(k;q,s,n) = (k+q)^{-s} \Big/ \sum_{i=1}^{n} (i+q)^{-s}, \qquad (3.2)$$

де f – ймовірність появи слова; k – ранг слова в списку слів; q, s – параметри апроксимації; n – обсяг словника. Слід зауважити, що параметр q та показник степеня s (s=1 у випадку класичного закону Ципфа) були теоретично обґрунтованими [там само]. Штрихова лінія на рис. 3.1

представляє результат апроксимації $f(k;\ q,\ s,\ n) = f(k;\ 2.35,\ 1.06,\ 12682)$ у випадку q=2.35 для рангово-ймовірнісного розподілу слів у творі Ш. Бронте “Джейн Ейр” з використанням розподілу Мандельброта. Функція Мандельброта (формула 3.2) описує розподіл слів у діапазоні малих та середніх значень рангу k, однак не забезпечує відтворення розподілу слів з високим рангом k (рис. 3.1, штрихова лінія).

Адні Юла та Герберт Саймон запропонували функцію розподілу [239, p. 426]

$$f(k;\rho) = \rho\,\mathrm{B}(k, \rho+1) \tag{3.3}$$

(де f – ймовірність появи слова; k – ранг слова в списку; ρ – параметр апроксимації; B – бета функція), що передбачає параметр апроксимації ρ (ρ >0) і дозволяє поліпшити результати відтворення в області великих значень рангу k (рис. 3.1, пунктирна лінія). Для діапазону малих значень рангу запропоновано двопараметричне узагальнення розподілу Юла-Саймона, де бета-функція замінена неповною бета-функцією $\mathrm{B}_{1-\alpha}$ ($0 \le \alpha < 1$):

$$f(k;\rho,\alpha) = \frac{\rho}{1-\alpha^{\rho}}\,\mathrm{B}_{1-\alpha}(k, \rho+1)\,. \tag{3.4}$$

Ця функція розподілу використана для аналізу роботи програм електронної пошти та частоти нот у музичних творах [198; 255].

Як видно з рис. 3.1, розглянуті функції Ципфа (формула 3.1), Мадельброта (формула 3.2), Юла-Саймона (формули 3.3 та 3.4) не дозволяють відтворити рангово-частотний розподіл слів у текстовому корпусі природної мови. Кожен із попередньо розглянутих розподілів, які були виведені емпірично або теоретично, забезпечує задовільний результат апроксимації лише в обмеженому діапазоні значень рангу слова (k). У цій дисертації здійснено детальний аналіз апроксимаційних показникових функцій з метою пошуку оптимальної функції розподілу для опису ймовірнісних характеристик слів у текстовому корпусі природної мови та інтерпретації змісту параметрів апроксимаційної функції.

У 1996 році французький біофізик Даніель Лавалетті запропонував закон рангування наукових журналів відповідно до їх імпакт-факторів [200]. Імпакт-фактор – показник цитованості журналів, що визначає інформаційну значимість наукових журналів. Відповідно до закону рангування Д. Лавалетті, розподіл імпакт-фактора f для наукових журналів визначається загальною кількістю журналів n, показником степеня s та ранговим номером k зі списку журналів, проаналізованих у порядку зниження їх імпакт-факторів. Для відтворення особливостей рангування журналів Д. Лавалетті запропонував таку функцію (базова функція):

$$f(k;s,n) = C\left[nk/\left(n-k+1\right)\right]^{-s}, \qquad (3.5)$$

де f – імпакт-фактор журналу; k – ранг журналу у списку, відсортованому в порядку зменшення f; s – показник ступеня; n – кількість журналів; C – масштабувальний множник.

Детально дослідивши особливості закону рангування та функції Лавалетті, І. Попеску дійшов висновку, що межа застосування функції Лавалетті є значно ширшою, ніж початково було запропоновано [223]. І. Попеску зазначив, зокрема, що, на противагу функції Ципфа та існуючим її модифікаціям, функція Лавалетті є більш придатною для апроксимації рангово-частотного розподілу слів у тексті [224, p. 85].

Підбір масштабувальний множника C та показника ступеня s базової функції Лавалетті (формула 3.5) дозволяє описати рангово-ймовірнісний розподіл слів у діапазоні середніх та великих значень рангу слова k (рис. 3.2, пунктир). Складності з використанням базової функції Лавалетті виникають при спробі застосувати її для опису частотного розподілу слів у діапазоні малих значень рангового числа k. Слід зауважити, що, на відміну від єдиного апроксимаційного параметра s у класичному законі Ципфа (формула 3.1), функція Лавалетті (формула 3.5) оперує співвідношенням $kn/(n\text{-}k+1)$, в якому присутній додатковий параметр – обсяг словника n.

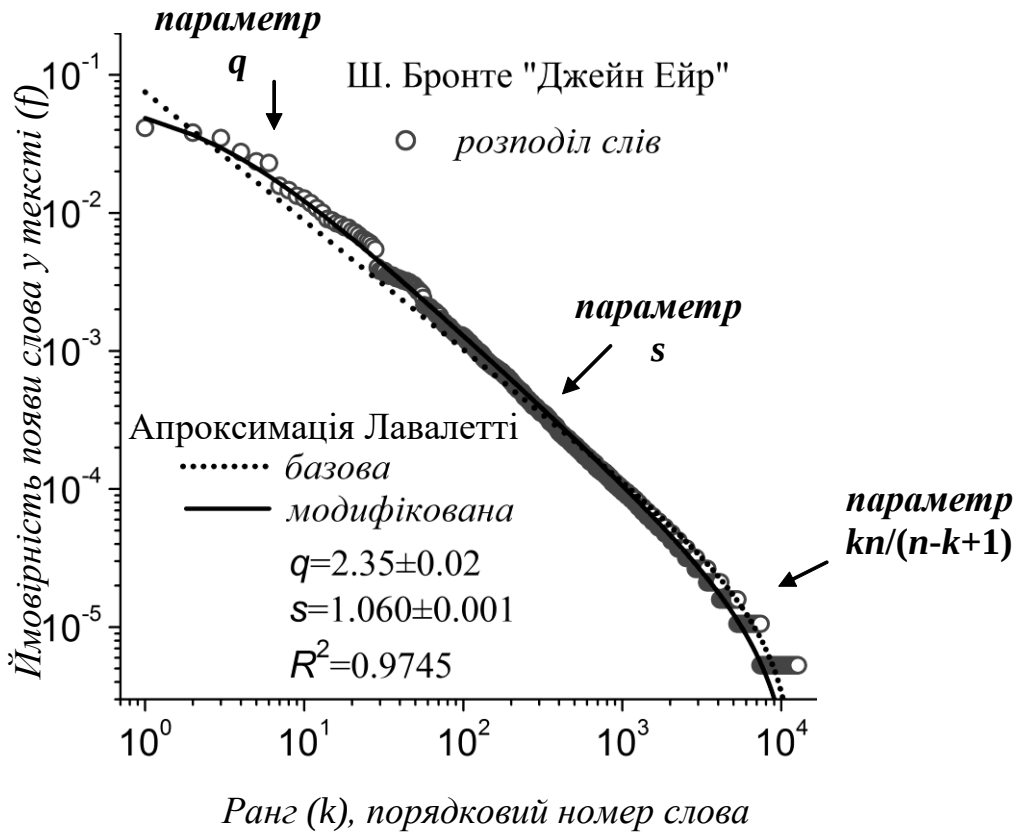


Рис. 3.2. Результати апроксимації рангово-ймовірнісного розподілу слів (Ш. Бронте “Джейн Ейр”) з використанням базової та модифікованої функцій Лавалетті.

Незважаючи на довгу історію вивчення та використання закону Ципфа, ряд дослідників уважає, що закон Ципфа не вичерпав своїх пізнавальних можливостей щодо його застосування в мовознавстві. Нещодавно запропоновано новий підхід для апроксимації Ципфо-подібних розподілів [225, p. 718]. Зроблено припущення, що текст складається з класів слів, комбінація яких формує загальний профіль кривої розподілу. Запропонована модель уже при розгляді лише двох компонент (m=1, 2) була успішно застосована до розподілів синсемантичних та автосемантичних слів для 100 текстів 20 мовами [там само, p. 718]. Інтерпретація змісту параметрів цієї моделі вважається складним, однак перспективним завданням, яке дозволить виявити імпліцитні характеристики тексту, пов’язані з його автором, тематикою, стилем, мовою тощо.

Використання комбінованих функцій із багатьма апроксимаційними параметрами дає можливість досягти якості відтворення рангового розподілу слів. Так, для поліпшення якості апроксимації запропоновано використання розподілу у вигляді комбінації показникових функцій з різними показниками ступеня [157, p. 168]. Проте, велика кількість параметрів ускладнює інтерпретацію їх змісту, а також зумовлює значну взаємозалежність параметрів. Завданням цієї дисертаційної праці є пошук апроксимаційної функції, зміст та величина параметрів якої корелюють з реальними характеристиками текстів.

Базова функція Лавалетті передбачає підгонку лише показника ступеня *s*, який визначає опис рангово-частотного розподілу слів у діапазоні малих та середніх значень рангового числа *k*. Апроксимація розподілу слів із великим значенням рангу (рис. 3.2) не потребує додаткового параметра, а забезпечується у випадку функції Лавалетті (3.5) наявністю співвідношення $nk/(n-k+1)$, де n – обсяг словника. Пунктирна лінія на рис. 3.2 представляє відтворення рангового розподілу слів (Ш. Бронте “Джейн Ейр”) функцією Лавалетті $f(k; s, n)=f(k; 1.06, 12682)$. Базова функція Лаваллетті (3.5) все-таки не описує задовільно розподілу в діапазоні малих значень рангового числа *k*. Наведені вище характеристики показникової функції розподілу Лавалетті стали вирішальними у виборі її як базової для подальших модифікацій.

Результати аналізу етапів модифікації показникових функцій рангового розподілу дозволили вибрати шлях вдосконалення базової функції Лавалетті (формула 3.5) з метою її адаптації до апроксимації рангового розподілу слів у текстах. Єдиною моделлю, яка забезпечує прийнятну апроксимацію рангового розподілу слів у діапазоні малих значень рангового числа *k*, є теоретична модель Б. Мандельброта (формула 3.2), згідно з якою функція розподілу оперує параметром q, який забезпечує опис рангово-частотного розподілу слів у діапазоні малих значень *k*. Параметр q з функції Мандельброта можна адаптувати до функції Лавалетті шляхом заміни

рангового числа k на суму (k+q). Отже, запропонувавши заміну рангового числа k на (k+q), функція Лавалетті може бути записана так [21; 252]:

$$f(k;q,s,n) = \left[n(k+q)/\left(n-(k+q)+1\right)\right]^{-s} \Big/ \sum_{i=1}^{n}(i+q)^{-s}, \qquad (3.6)$$

де f – ймовірність появи слова; k – ранг слова в списку; q – апроксимаційний параметр, що описує розподіл високочастотних слів; s – апроксимаційний параметр (показник ступеня), що описує розподіл слів із середньою частотою вживання; n – обсяг словника.

Слід зауважити, що параметр q, адаптований у модифікованій функції Лавалетті (3.6), запозичений із функції Б. Мандельброта (3.2), і визначає профіль кривої розподілу так, як і у випадку теоретично обґрунтованої моделі Мандельброта. Таким чином, два незалежні параметри (q та s), запропоновані в модифікованій функції Лавалетті (3.6), описують розподіл у різних діапазонах, а саме: в області малих та середніх значень рангового числа k, відповідно. Завдяки співвідношенню $n(k+q)/(n-(k+q)+1)$, використаному у модифікованій функції Лавалетті, описано ранговий розподіл слів у діапазоні великих значень рангового числа k без залучення додаткових параметрів.

Суцільна лінія на рис. 3.2 відображає результат апроксимації модифікованою функцією розподілу Лавалетті (3.6) $f(k; q, s, n) = f(k; 2.35, 1.06, 12682)$ з q=2,35 для “Джейн Ейр” (Ш. Бронте). Для визначення ступеня якості апроксимації між розрахунковими $f_i^{розр.}$ та експериментальними значеннями ймовірності f_i появи i-го слова у досліджуваному тексті використовують коефіцієнт детермінації R^2 [187, p. 58; 225, p. 718]. Коефіцієнт детермінації може набувати значень у таких межах: $0 \le R^2 \le 1$. Коефіцієнт детермінації R^2 набуває нульового значення у випадку апроксимації експериментальних даних їх середнім значенням. Максимального значення 1 коефіцієнт детермінації R^2 досягає у випадку повного відтворення експериментальних даних апроксимаційною функцією.

Близькі до 1 значення коефіцієнта детермінації R^2 відповідають високій якості відтворення експериментальних даних апроксимаційною функцією.

Додатковий параметр q у модифікованій функції Лавалетті (3.6) дозволяє отримати задовільну підгонку рангово-ймовірнісного розподілу в області малих значень рангового числа k. Наявність параметра q забезпечує однозначне відтворення параметра s, який визначає стрімкість спаду ймовірності появи слова у тексті із збільшенням його порядкового номера, представленої у логарифмічній системі координат.

3.2 Рангово-частотний розподіл слів в англо-, німецько- та україномовних наукових і художніх текстах

3.2.1. Зіставлення рангово-частотного розподілу слів в англомовних наукових і художніх текстах. Тестування запропонованої апроксимаційної функції (3.7) проведено для англо-, німецько- та україномовних текстів наукового та художнього стилів.

Корпус текстів наукового стилю сформовано з наукових праць з фізики та математики. До корпусу наукової літератури увійшли монографії (“Crystal Design: Structure and Function” by Gautam R. Desiraju, John Wiley & Sons, Ltd. 2003; “Lecture notes in Statistics: Bayesian spectrum analysis and parameter estimation” by Bretthorst, G. Larry, Springer-Verlag 1988; “Mathematical models for speech technology” by Stephen E. Levinson, John Wiley & Sons Ltd. 2005; “PLS Toolbox 3.5 for use with MATLAB” by Barry M. Wise et al., Eigenvector Research, Inc. 2005;), дисертаційні роботи (Rene T. Wegh 1999 “Vacuum ultraviolet spectroscopy and quantum cutting for trivalent lanthanides”; Marcus True 2004 “Fine structure in d-f and f-f transitions of Tm^{3+} and systematic investigation of $3d^5$-$3d^4$4s absorption of Mn^{2+} doped fluorides”; Dmitri V. Talapin 2002 “Experimental and theoretical studies on the formation of highly luminescent II-VI, III-V and core-shell semiconductor nanocrystals”; Lisabeth van Pieterson 2001 “Charge transfer and $4f^n$-$4f^{n-1}$5d luminescence of lanthanide ions”; Yury

Kuzminykh 2006 “Crystalline rare-earth-doped sesquioxide and YAG PLD-films”; Christoph Bostedt 2002 “Electronic structure of germanium nanocrystal films probed with synchrotron radiation”), реферовані статті з міжнародного журналу Physical Rewiew B [227], наукові статті 4 авторів (представники голландської наукової школи P. Dorenbos та A. Meijerink, представник української наукової школи G. Stryganyuk, представник німецької наукової школи G. Zimmerer) з галузі люмінесцентного матеріалознавства. Однорідність даної вибірки забезпечена вибором текстів, що належать науковому стилю.

Текстовий корпус художньої літератури включає твори різного жанру англійською мовою: “War and peace”, “Anna Karenina” та “Childhood” Leo Tolstoy; “Villette” та “Jane Eyre” Charlotte Bronte; “The adventures of Sherlock Holmes” Arthur Conan Doyle; “Dracula” Bram Stoker; “Dombey and son” та “A tale of two cities” Charles Dickens; “Robinson Crusoe” Daniel Defoe; “Sense and Sensibility” та “Emma” Jane Austen; “Alice's adventures in Wonderland” та “The hunting of the Snark” Lewis Carroll; “The prince and the pauper” та “The adventures of Tom Sawyer” Mark Twain; “Don Quixote” Miguel de Cervantes; “Robin Hood” J. Walker McSpadden; “The tragedy of King Lear” William Shakespeare; “Bridget Joness” та “Diary of Bridget Jones: The Edge of Reason” Helen Fielding; “Chocolat” та “Five Quarters of the Orange” Joanne Harries; “Harry Potter and the Goblet of Fier” та “Harrry Potter and the Sorcerers Stone” Joanne Rowling; “Duma Key” та “The Stand” Stephen King; “Tuesday with Morrie” та “Five People you meet in Heaven” Mitch Albom; “Murder Mysteries” та “The Graveyard” Neil Gaiman. Із цього текстового корпусу сформовано вибірки: 1) текстів художньої літератури 16 – 19 – 21 століття; 2) текстів художньої літератури 19 століття, написаних носіями англійської мови; 3) текстів художньої літератури 19 століття, перекладених англійською мовою; 4) текстів художньої літератури 19 століття, написаних носіями англійської мови, та текстів, перекладених англійською мовою. Серед запропонованих вибірок тільки вибірка 2) відповідає вимогам однорідності. Вибірка 1) не є однорідною за часом написання художніх творів, вибірка 4)

містить твори носіїв англійської мови та переклади англійською мовою. Такі вибірки створені для того, щоб перевірити, чи є вони однорідними з точки зору виконання закону Ципфа.

Зіставлялися тексти наукового та художнього стилів, оскільки вони відрізняються стильовою контрастністю, а саме: наявністю образності у художньому стилі та її відсутністю – у науковому, логічним викладом матеріалу в науковій літературі та емоційним висловленням думок, експресивним розвитком подій сюжету в художньому творі [96]. У випадку зіставлення стилів оперують поняттям *нульового стилю*, введеного В. І. Перебийніс, який є незалежним від решти аналізованих стилів і з яким можна проводити зіставлення інших стилів [85, p. 18]. У цій дисертації текст нульового стилю не формувався, оскільки зіставлялися лише два стилі між собою. У таких випадках створення нульового стилю є недоцільним [там само, p. 17].

Для детального дослідження якості апроксимації рангового розподілу слів текстів наукової та художньої літератури модифікованою функцією Лавалетті (3.6) проведено аналіз залежності параметрів q та s від обсягу словника. Простежується різний рангово-ймовірнісний розподіл слів у науковій та художній літератури. Причому цей нахил майже не залежить від обсягу словника n. Пор.: криві 1, 1’та 2, 2’ між собою (Додаток Б, рис. 1). На рис. 3.3 показано, що якість апроксимації (коефіцієнт детермінації R^2) залежить від обсягу словника (n) текстів художньої літератури: зі збільшенням обсягу словника коефіцієнт детермінації прямує до одиниці. Однак, така залежність не була виявлена для текстів наукової літератури.

Як показано на рис. 3.4.а, на противагу науковим текстам, параметр q у художній літературі виявляє суттєву варіацію та прямує до більших значень. У жодному з розглянутих випадків не виявлено коваріації параметра q з коефіцієнтом детермінації R^2. У художній літературі чітко виявляється залежність значення параметра s від коефіцієнта детермінації R^2 (рис. 3.4.б, штрихова лінія). Оскільки для художньої літератури виявлена подібна

залежність R^2 від n, то можна було б припустити, що параметр s зростає зі збільшенням обсягу словника. Параметр s набуває відносно менших значень у науковій літературі та не демонструє чіткої залежності від коефіцієнта детермінації R^2. Отже, особливості текстів, які описує параметр q, характерні більше загалом для художньої літератури, а особливості, відтворювані параметром s, залежать від обсягу словника художнього твору.

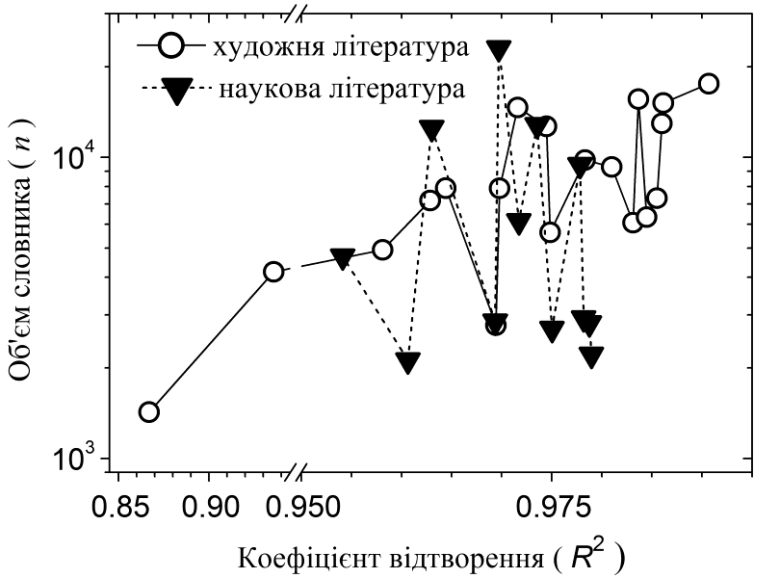


Рис. 3.3. Вияв залежності якості детермінації (R^2) від обсягу словника для текстів наукової та художньої літератури.

У розглянутих у цій дисертації апроксимаціях рангового розподілу слів (відповідно до запропонованої модифікації функції Лавалетті) досягнуто середнього значення коефіцієнта детермінації $R^2 = 0.970 \pm 0.015$, що забезпечило високу якість апроксимації усіх особливостей експериментальних кривих розподілу. Таке значення коефіцієнта детермінації означає, що 97% вихідних даних апроксимуються запропонованою модифікованою функцією Лавалетті. Запропонована апроксимація з високою точністю описує всі ділянки рангового розподілу слів у тексті, а саме:

а) помірний розподіл слів на початковому етапі спаду ймовірності появи слова в області високочастотних слів до ~10-100, що визначається параметром q;

б) основний розподілу слів із середньою частотою появи у тексті, що задається параметром s;

в) зростаючий спад ймовірності появи слова у тексті в області низькочастотних слів, який відтворюється завдяки наявності співвідношення $n[k+q]/[n-(k+q)+1]$.

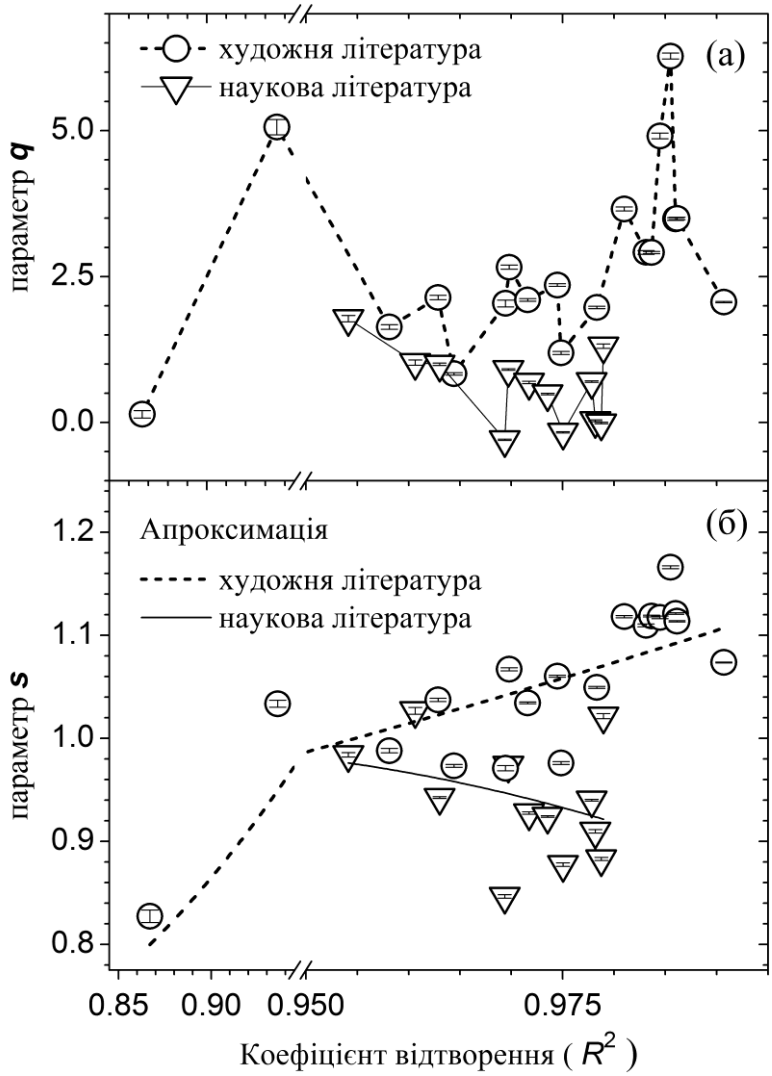


Рис. 3.4. Апроксимаційні параметри модифікованої функції Лавалетті q (а) та s (б) для наукової (кружечки) та художньої (трикутники) літератури, впорядковані у порядку зростання коефіцієнта детермінації R^2. Абсолютні похибки представлені щілинами всередині відповідних символів.

Усі параметри в сукупності визначають точність опису розподілу слів у різних діапазонах рангового числа *k*. Якщо, наприклад, забрати з розгляду параметр *q*, опустивши його як доданок у формулі (3.6), тоді апроксимація відбуватиметься за рахунок відхилення параметра *s*, однак при цьому погіршується опис прямолінійної ділянки залежності. Параметр *s* визначає лише розподіл слів із середньою частотою появи у тексті, однак не призначений для відтворення високочастотних слів на початковому етапі спаду ймовірності появи слова, де розподіл слів повинен визначатись параметром *q*. Наявність у формулі (3.6) такого параметра, як обсяг словника *n* також створює додаткові можливості для опису рангового розподілу слів, змінюючи *n*, увівши додатковий параметр *n'* і розглядаючи замість *n* величину *n* + *n'*. Варіація параметра *n'* забезпечить ефективний опис розподілу слів в області великих значень рангового числа *k*. У вище розглянутих прикладах значення *n* було фіксованим відповідно до обсягу словника і не змінювалося для оптимізації опису кривих розподілу слів для наукових та художніх текстів.

Результати, представлені на рис. 3.4.б та Додатку Б (рис. 1), показують, що рангово-ймовірнісний розподіл слів (параметр *s*) є різним для наукової та художньої літератури. Така відмінність виявляється краще при збільшенні обсягу словника (*n*) творів художньої літератури. Для перевірки достовірності виявлених відмінностей рангово-ймовірнісного розподілу слів (параметрів *s* та *q*) проаналізовано тексти наукової та художньої літератури, які мали приблизно однаковий обсяг словника (≈40000 слів).

На рис. 3.5 представлені рангові розподіли слів у сумарних текстових корпусах наукової (трикутники) та художньої (кружечки) літератури. Рангово-ймовірнісні розподіли слів були апроксимовані з використанням модифікованої функції Лавалетті $f(k;q,s,n)$ (формула 3.6), що дозволило знайти оптимальні функції для наукової *f*(*k*; 0.962, 1.0018, 43688) та для художньої *f*(*k*; 3.548, 1.1203, 42653) літератури з коефіцієнтом детермінації R^2 = 0,96745 та 0,99378, відповідно.

У наукових текстах показник ступеня *s* (*s*=1.0018) набуває менших значень у зв'язку з менш стрімким спадом ймовірності появи слова у тексті (пунктирна лінія на рис. 3.5), порівняно з художньою літературою, для якої більш стрімкий спад ймовірності появи слова у тексті (суцільна лінія на рис. 3.5) відтворюється більшим значенням показника ступеня *s* (*s*=1.1203).

Параметр *q* забезпечує опис рангового розподілу слів у діапазоні малих значень рангового числа *k*. Параметр *q* є різним для текстів наукової та художньої літератури. У текстах наукової літератури область початкового спаду ймовірності слова виявляється меншою мірою, і тому значення параметра *q* для наукових текстів (0.962) є суттєво меншим, порівняно з текстами художньої літератури. У текстах художньої літератури початковий етап спаду ймовірності появи слова має більш помірний нахил в області значень рангового числа *k* ~100 і параметр *q* сягає значення 3.548.

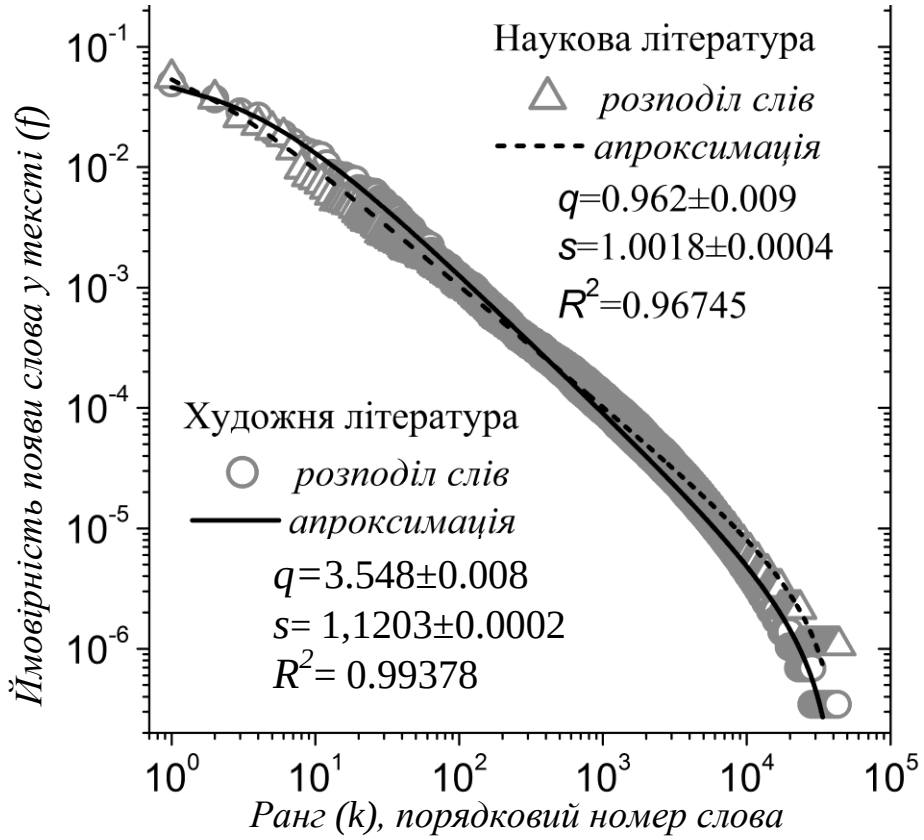


Рис. 3.5. Результати апроксимації рангового розподілу слів для текстів наукової та художньої літератури з використанням модифікованої функції Лавалетті.

У вище розглянутому прикладі для побудови рангового розподілу слів для наукових текстів було обрано дисертації, монографії, наукові статті. Розглянемо, як змінюються апроксимаційні параметри у випадку одного жанру наукової літератури – наукової статті. Для цього проведемо аналіз масиву текстів, сформованого з реферованих журнальних статей P. Dorenbos, A. Meijerink, G. Stryganyuk та G. Zimmerer. У цьому випадку ранговий розподіл слів описується параметрами q=0,664 та s=0,984. Параметр s=0.984 для праць чотирьох авторів є близьким до параметру s=1.001 для наукових текстів різних жанрів. Отже, як і було попередньо висловлено, параметр s може відповідати за науковий стиль. Однак, параметр q для цих наукових текстів відрізняється. Крім того збереглась і попередня тенденція, параметр q значно менший, ніж цей параметр для художньої літератури, де q=2.35 (див. рис. 3.2), або ж q=3.55 (рис. 3.5). Отже, можна ще раз підтвердити, що параметр s відповідає за розмежування текстів за стилем.

Запропонована модифікація функції Лавалетті $f(k; q, s, n)$ передбачає два параметри (q та s) для апроксимації рангово-ймовірнісного розподілу слів. Оскільки ці апроксимаційні параметри набувають різних кількісних показників для розподілу слів у текстах наукової та художньої літератури, то вони можуть бути використані для розмежування цих двох стилів.

Параметр q визначає характер рангового розподілу слів у діапазоні слів з високою частотою вживання. Більші кількісні показники параметра q відповідають більш пологому спаду рангового розподілу слів. Згідно з даними для текстів художньої літератури переважаючим для цих значень k є службові слова. З огляду на це, оптимальні кількісні показники параметра q є меншими для наукової літератури, порівняно з художньою літературою. Можна припустити, що функція службових слів у текстах наукової літератури є менш валідною, порівняно з текстами художньої літератури.

Параметр s визначає основний нахил рангово-ймовірнісного розподілу слів у логарифмічній системі координат. Стрімкість спаду ймовірності появи слова у тексті виявилась меншою в науковій літературі (s=1,0018), що

виявляється в менших значеннях параметра *s*, порівняно з художньою літературою (*s*=1,1203). Помірна стрімкість спаду рангово-ймовірнісного розподілу слів може бути зумовлена широким використанням наукової термінології. У художній літературі визначальним може бути частотне вживання імен, явищ, ознак тощо. Частка цих слів зростає у художніх текстах із малим обсягом словника. Тому рангово-ймовірнісний розподіл слів у художній літературі (параметр *s*) виявляє залежність від розміру тексту.

Звичайно, можливою є ситуація, коли запропонована модифікація функції Лавалетті не буде оптимальною для апроксимації рангово-частотного розподілу слів. У таких випадках досить перспективним є застосування моделі кумулятивних класів "*cumulative classes*" [225, p. 218; 252, p. 288], згідно з якою текст розглядають як комбінацію кількох класів, для кожного з яких характерним є Ципфо-подібний рангово-ймовірнісний розподіл властивих класові лексичних одиниць. Однак, чітка схема методики розмежування класів лексичних одиниць для моделі кумулятивних класів є відсутньою.

У роботі В. П. Маслова згадується, що розподіл Ципфа дозволяє провести розмежування текстів за жанром, однак не за авторським стилем [78, с. 719]. М. Монтемуро ж указує на можливість проведення авторської атрибуції художніх текстів, аналізуючи рангово-частотний розподіл слів та використовуючи різницю в нахилах кривої розподілу для ділянок з великим та малим значенням рангового числа *k* [213, p. 570]. У цій дисертації проаналізовано можливість використання рангового розподілу послідовності вживання одного і більше слів (n-грам) слів для стильової та авторської атрибуції наукових текстів. До розгляду взято послідовність слів, розміром n=1, ... 4. Для такого дослідження вибрано 40 англомовних праць з галузі фізики твердого тіла, опублікованих у реферованих журналах. Реферований журнал – це журнал, статті до якого відібрані за критеріями відповідності сучасним стандартам наукових публікацій. Праці розділені на 4 групи (по 10 у кожній) відповідно до авторів (співавторів): 1) проф. д-р. P. Dorenbos; 2)

проф. д-р. A. Meijerink); 3) д-р. G. Stryganyuk; 4) проф. д-р. G. Zimmerer. Загальний словник проаналізованих робіт налічує 11385 словоформ.

Програма "Lexical Content Searcher" використана для опрацювання наукових статей P. Dorenbos, A. Meijerink, G. Stryganyuk, G. Zimmerer, щоб перевірити можливість використання рангового розподілу послідовності вживання одного і більше слів (n-грам) для розмежування наукових текстів. Отримано статистичні дані щодо ймовірностей 1, ... 4-грам слів у текстах цих авторів, побудовано ранговий розподіл для розглянутих послідовностей слів. Рангові розподіли слів для робіт різних груп авторів є близькими у випадку 1-грам у діапазоні середніх значень рангового числа k. Такий результат може підтверджувати спорідненість текстів за стилем. Це узгоджується з попереднім висновком про можливість використання рангового розподілу слів для розмежування текстів за стилем.

Зі збільшенням розміру послідовності вживання одного і більше слів виявляються відмінності у стрімкості спаду ймовірності появи слова у тексті. Можна припустити, що це свідчить про тематичні або ж авторські особливості текстів. У випадку послідовності з 2 слів відмінності рангово-ймовірнісного розподілу є незначними. Для послідовності з 3 та 4 слів ці відмінності є наочними в діапазоні всіх значень рангового числа k. Така поведінка стрімкості спаду ймовірності появи слова у тексті дає перспективу для атрибуції текстів, однак аналіз 2-, 3- та 4-грам ускладнений, оскільки ці розподіли не описуються задовільно функцією Лавалетті з параметрами q, та s. Винятком є лише 1-грам розподіл, який можна апроксимувати модифікованою функцією Лавалетті зі змінними параметрами q, s та n.

Апроксимація рангового розподілу слів модифікованою функцією Лавалетті з параметрами q, s та n для кожного з авторів у випадку послідовності вживання з 1 слова представлена у Додатку Б (рис. 2). Застосована апроксимація з високою точністю описує вихідні дані. Так, для наукових статей G. Zimmerer коефіцієнт детермінації складає $R^2 = 0.97053$. Це означає, що 97% даних описується використаною апроксимацією. Параметр s

є близьким для текстів різних авторів і змінюється в межах 0.8950 , ... 1.0145. Відхилення, які наводяться в табл. 3.1, наприклад, s=1.0145±0.0025 відображають точність визначення параметру відповідно до вибраної апроксимації розподілу слів модифікованою функцією Лавалетті, а не межі зміни цього параметру для досліджуваного масиву текстів чотирьох авторів.

Щоб знайти $s_{середнє}$ необхідно провести усереднення за s кожного з авторів. У той же час параметр апроксимації q та n різні для кожного з авторів. Так, q=1.461 для статей A. Meijerink і q=0.207 для статей G. Stryganyuk. Значення параметра n відображає обсяг словника автора, а параметр q набуває більших значень за умови інтенсивного використання автором службових частин мови. Отож, значення цих обидвох параметрів можуть залежати від авторського стилю, який у багатьох випадках може також залежати від рівня володіння мовою автором.

Таблиця 3.1

Параметри апроксимації рангово- ймовірнісного розподілу слів наукових текстів різних авторів модифікованою формулою Лавалетті

Автор	q	s	n	R^2
A. Meijerink	**1.461**±0.043	**1.0145**±0.0025	**5327**±62	0.96128
P. Dorenbos	**0.349**±0.019	**0.9320**±0.0016	**6813**±58	0.96462
G. Zimmerer	**0.303**±0.022	**0.8936**±0.0019	**4950**±53	0.97053
G. Stryganyuk	**0.207**±0.026	**0.8950**±0.0025	**3220**±33	0.96212
$s_{середнє}$		**0.93**		

Отже, відмінність числових значень параметрів q та n для кожного з авторів припускає можливість використання цих параметрів для авторської атрибуції наукових текстів методом аналізу рангового розподілу слів.

Подібні закономірності щодо залежності параметрів q, s та n від автора тексту знайдено і для авторів художньої літератури. У табл. 3.2 наведено параметри апроксимації для текстів художньої літератури 19 століття,

написаних носіями англійської мови. Високий рівень апроксимації досягнуто для Ch. Dickens, коефіцієнт детермінації R^2=0.996. Параметр s для художніх текстів лежить у межах 1.105, ..., 1.182, і близькість цих параметрів підтверджує, що ці твори належать одному стилю. Для параметра q можна відзначити тенденцію, що він значно більший, ніж у науковій літературі. Для певних авторів (L. Carroll, A. Conan Doyle, B. Stoker,) значення параметра q близькі між собою, у той час як параметр n відрізняється суттєвіше. Відтак, для розмежування текстів за автором важливо не тільки враховувати параметр q, але також і параметр n.

Таблиця 3.2

Параметри апроксимації рангово-ймовірнісного розподілу текстів художньої літератури різних авторів

Автор	q	s	n	R^2
Ch. Bronte	**3.0792**±0.0124	**1.10543**±0.00048	**24084**	0.99177
Ch. Dickens	**4.2432**±0.0099	**1.16344**±0.00032	**26647**	0.99623
A. Conan Doyle	**3.6989**±0.0219	**1.13742**±0.00084	**15483**	0.98845
J. Austen	**5.8293**±0.0257	**1.18244**±0.00077	**13083**	0.99316
L. Carroll	**3.6944**±0.0302	**1.10503**±0.00119	**8332**	0.98972
B. Stoker	**3.8425**±0.0183	**1.15028**±0.00067	**17196**	0.99170
$s_{середнє}$		**1.14**		

Для оцінки статистичної значимості розкидів параметрів апроксимації використаємо метод довірчих інтервалів. Зупинімося на аналізі параметру s. Для оцінки розкиду параметра s кожного автора запишемо межі $\overline{S}$ ($s_{середнє}$) у рамках 95% інтерквантильного інтервалу. Під 95% інтерквантильним інтервалом розуміють інтервал між квантиліями рівнів 0.975 та 0.025 розподілів частот. Він задається: $s_{середнє} - 2\sigma$, $s_{середнє} + 2\sigma$, де σ – стандартне відхилення. Стандартне відхилення визначається за формулою:

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}, \qquad (3.7)$$

У табл. 3.3 представлено межі 95% інтерквантильного інтервалу для параметра *s*, що описує апроксимацію рангового розподілу слів для текстів наукової літератури (P. Dorenbos, A. Meijerink, G. Stryganyuk, G. Zimmerer) та для текстів художньої літератури, написаних носіями англійської мови (Додаток Б, рис. 2). $s_{середнє}$ та σ, отримані на основі аналізу результатів приведених у табл. 3.1 та 3.2.

Таблиця 3.3

Смуги коливань параметра *s*

	$\bar{s}$	$\sigma_{\bar{s}}$	$\bar{s} \pm 2\sigma_{\bar{s}}$
Наукові тексти: (*P. Dorenbos, A. Meijerink, G. Stryganyuk, G. Zimmerer*)	0.93	0.05	0.83 – 1.03
Художні тексти 19 століття: (*J. Austen, Ch. Bronte, L. Carroll, A. Conan Doyle, Ch. Dickens, B. Stoker*)	1.14	0.04	1.06 – 1.22
Художні тексти 21 століття: (*J. Harries, J. Rowling, S. King, M. Albom, N. Gaiman*)	1.08	0.02	1.04 – 1.12

Обчислені смуги коливань параметра $s_{середнє}$ (0.83 – 1.03 та 1.06 – 1.22) не перетинаються, що вказує на те, що $s_{середнє}$ значимо відрізняється для наукових та художніх текстів. Варто відзначити, що $s_{середнє}$=1.14, обчислене для текстів художньої літератури, написаних носіями англійської мови 19 століття, практично збігається зі значенням *s*=1.12, знайденим у випадку аналізу масиву всіх досліджуваних художніх текстів (рис. 3.5). Параметр $s_{середнє}$=0.93, знайдений на основі аналізу наукових статей, дещо відрізняється від *s*=1.00, отриманого на основі аналізу всіх наукових текстів. Незначне зменшення $s_{середнього}$ для наукових статей, порівняно із цим параметром для масиву всієї проаналізованої наукової літератури, може бути зумовлено

малим обсягом статей, що приводить до флуктуації частоти ознак і зумовлює спостережуване відхилення *s*.

Можна відзначити, що значення параметра *q* для художньої літератури у декілька разів більше, ніж значення цього параметра для наукової літератури, наприклад, *q*=3.69 для художніх текстів (A. Conan Doyle) і *q*=0.349 для наукових текстів (G. Zimmerer). Це підтверджує загальну тенденцію, виявлену раніше за аналізу всіх масивів наукових та художніх текстів (рис. 3.5), де *q*=0.962 для наукової літератури і *q*=3.548 для художньої літератури.

Визначимо, якою мірою вибірка тексту впливає на значення параметру *s* для текстів художньої літератури. Для цього проаналізуємо зміну параметра *s* для таких масивів текстів:

1) англомовні тексти художньої літератури 16-19-21 століття;

2) англомовні тексти 19 століття (J. Austen, Ch. Bronte, L. Carroll, A. Conan Doyle, Ch. Dickens, B. Stoker);

3) об'єднання текстів, що увійшли в масиви текстів 1) та 4);

4) тексти, перекладені англійською мовою (Л. Толстой).

На рис. 3.6 представлені апроксимовані криві рангового розподілу слів для цих масивів текстів. Щоб уникнути накладання, апроксимаційні криві зміщені один відносно одного вздовж осі *y*. У результаті попереднього аналізу англомовних текстів 19 століття (масив 2) було встановлено, що 95% інтерквантильний інтервал має межі 1.14±0.08 [1.06 – 1.22] (табл. 3.3). Вибірка англомовних текстів 19 століття – це твори англомовних авторів, написані в межах одного століття, що дозволяє розглядати вибірку як однорідну. Кількісні показники параметру *s* для масивів текстів 1), . . 4) є близькими між собою і знаходяться в інтервалі 1.12 – 1.17 (див. табл. 3.4), не виходячи за межі 95% інтерквантильного інтервалу [1.06 – 1.22], знайденого для вибірки однорідної щодо носія мови та часу написання. Це дозволяє говорити, що відмінності між параметром *s* для цих текстів не є суттєвими, а отже, вибірки текстів 1), ..., 4) можна вважати однорідними відносно

параметра *s*. Вони не мають статистично значимих розбіжностей, а отже, належать одному стилю. Спостерігається збіг параметра $s_{середнє}$ =1.14 для художніх текстів 19 століття зі значенням *s*=1.1406 для англомовних текстів, об'єднаних в одну вибірку.

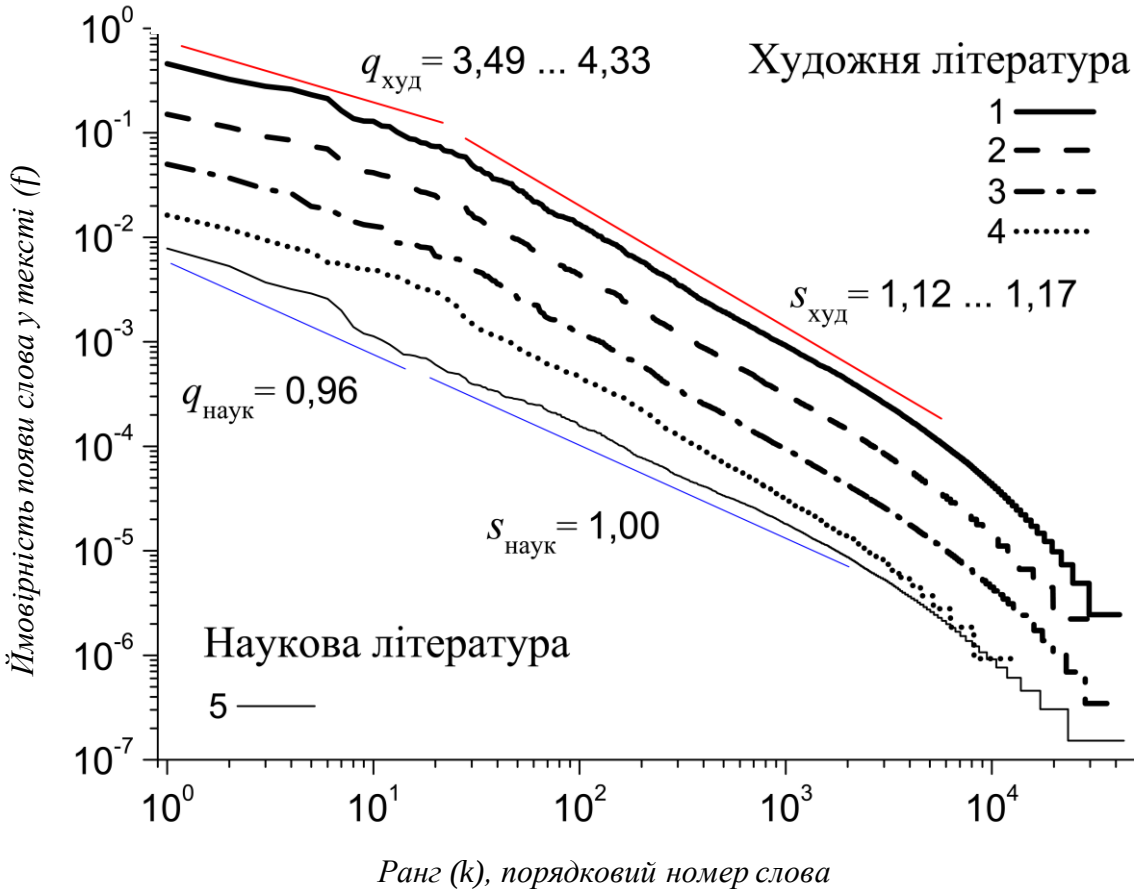


Рис. 3.6. Криві рангово-ймовірнісного розподілу слів для текстів художньої літератури: 1 – англомовні тексти художньої літератури 16-19-21 століття ; 2 – англомовні тексти 19 століття (Ch. Bronte, A. Conan Doyle, Ch. Dickens, J. Austen, L. Carroll, B. Stoker); 3 – англомовні тексти художньої літератури 16-19 століття та тексти, перекладені англійською мовою, 4 – тексти, перекладені англійською мовою (Л. Толстой). Крива 5 відображає рангово-ймовірнісний розподіл слів у текстах наукової літератури.

Враховуючи цей факт та обставину, що параметр *s* для текстів художнього та наукового стилів англійської мови має статистично значимі розбіжності, то:

1) параметр s=0.93±0.10 модифікованої функції Лавалетті, знайдений на основі аналізу англомовних текстів наукової літератури, може бути характерним параметром для визначення англомовних текстів наукового стилю;

2) параметр s=1.08±0.04 модифікованої функції Лавалетті, знайдений на основі аналізу англомовних текстів художньої літератури, може бути параметром приналежності тексту до художнього стилю.

Таблиця 3.4

Параметри апроксимації кривих рангово-ймовірнісного розподілу текстів наукової та художньої англомовної літератури

Текст	q	s	R^2
1 - Англомовні 16-19-21 ст	**4.335**±0.009	**1.1704**±0.0002	0.99465
2 - Англомовні 19 ст	**3.827**±0.009	**1.1406**±0.0003	0.99516
3 - Об'єднання (16-19 ст. та перекладені)	**3.548**±0.008	**1.1503**±0.0002	0.99378
4 - Перекладені	**3.487**±0.023	**1.12116**±0.0008	0.98600
5 - Наукові	**0.964**±0.009	**1.0018**±0.0004	0.96745

Примітка: назви текстів та їх нумерація наведені відповідно до позначення кривих на рисунку 2.14.

3.2.2. Зіставлення рангово-частотного розподілу слів у німецькомовних наукових і художніх текстах. Проаналізовано: німецькомовні наукові тексти XXI століття (монографії: *Wolfgang W. Osterhage* Studium Generale Physik. Ein Rundflug von der klassischen bis zur modernen Physik, *Michael Komma* Moderne Physik mit Maple: von Newton zu Feynman, *Rainer Scharf* Ausgezeichnete Physik; дисертаційні праці: *A. Guesmann* Selbstorganisation zwischen Mannigfaltigkeiten euklidischer und nichteuklidischer Geometrie durch Kooperation und Kompetition 2006, *T. Latz* Spektroskopie im Laser-Resonator mit höchster Nachweisempfindlichkeit und spektraler Auflösung 2000, *C. Granzow* Quanten-Trajektorien von

zusammengesetzten Systemen 1999, *Ch. Rotsch* Elastizitätsmessungen an Lebenden Zellen mit dem Rasterkraftmikroskop), німецькомовні художні тексти XIX століття (*W. Raabe:* Die Chronik der Sperlingsgasse, Deutscher Mondschein, Zum wielden Mann, Meister Autor; *T. Mann*: Buddenbrooks, Der Tod in Venedig, Tonio Kroger, Der kleine Herr Friedermann; *T. Fontane*: Effi Briest, Schach von Wuthenow, Unterm Birnbaum; *T. Storm*: Pole Puppenspoler, Immensee, Aquis Submersus, Die Regentrude) та XXI століття (*A. Friedrich*: Süden, Totsein verjährt nicht, Verzeihen; *K. Gier*: Der Braut sagt leider nein, Lüge die von Herzen kommen, Rubinrot; *J. Rudiger*: Himmelslichter; *F. Shätzing*: Keine Angst, Der Schwarm, Tod und Teufel; *P. Süskind*: Das Parfurm, Die Taube, Ein Kampf) .

Апроксимаційні параметри розподілу слів для наукових та художніх німецькомовних текстів відповідно до модифікованої формули Лавалетті представлено в табл. 3.5 та табл. 3.6 відповідно та на рис. 3.7. Ранговий розподіл слів у наукових текстах зображено трикутниками, а у художніх – кружечками. Пунктирною лінією позначено апроксимацію рангового розподілу слів для німецькомовних наукових текстів, суцільна лінія відображає апроксимацію рангового розподілу слів для німецькомовних художніх текстів. Коефіцієнт відтворення R^2=0,96 прямує до одиниці як для наукового, так і для художнього стилів німецької мови, що підтверджує високу точність опису розподілу слів у текстах та отриманих результатів.

Таблиця 3.5

Параметри апроксимації рангового розподілу слів для німецькомовних наукових текстів (жанр – дисертація)

Автор	s	q	R^2
A. Guesmann	0.898±0.0016	0.917±0.033	0.95593
T. Latz	0.909±0.0016	1.105±0.033	0.96575
C. Granzow	0.915±0.0022	3.345±0.090	0.94689
C. Rotsch	0.870±0.0018	1.072±0.041	0.94545
$s_{середнє} \pm 2\sigma$	0.90±0.04, [0.86 – 0.94]		

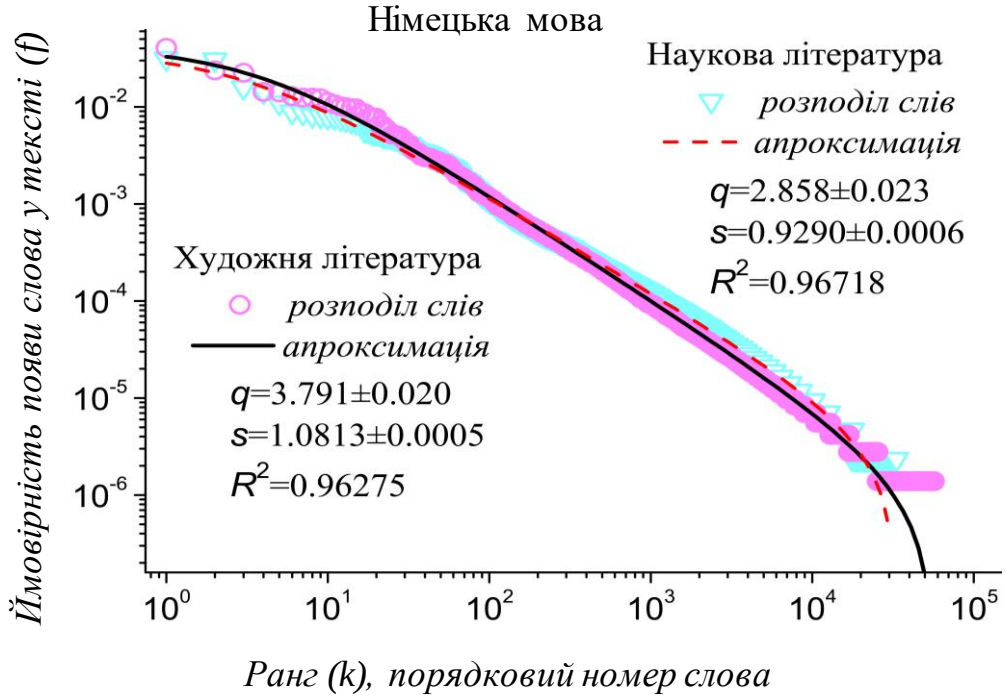


Рис. 3.7. Зіставлення апроксимованих рангових розподілів слів для наукового та художнього німецькомовного тексту.

Таблиця 3.6

Параметри апроксимації рангового розподілу слів для текстів художньої літератури німецькомовних авторів

Автор		s	q	R^2
XIX століття	W. Raabe	1.021±0.0011	2.634±0.035	0.944
	T. Fontane	1.096±0.0013	5.214±0.055	0.942
	T. Mann	1.028±0.0007	2.476±0.021	0.957
	T. Storm	1.033±0.0020	4.010±0.072	0.931
	$s_{середнє}\pm 2\sigma$	1.05±0.06, [0.99 – 1.11]		
XXI століття	A. Friedrich	1.09	2.317±0.024	0.957
	K. Gier	1.07	2.664±0.083	0.949
	J. Rudiger	1.04	4.132±0.035	0.963
	F. Shätzing	1.06	3.512±0.024	0.971
	P. Süskind	1.01	4.132±0.041	0.957
	$s_{середнє}\pm 2\sigma$	1.05±0.08, [0.97 – 1.13]		

Середнє значення параметра *s* для німецькомовної наукової (*s* =0.90) та художньої (*s*=1.05) літератури відрізняються. Ця відмінність є суттєвою, оскільки їх 2σ інтерквантильні інтервали для наукових текстів [0.86 – 0.94] та художніх текстів [0.99 – 1.11] (XIX століття) і [0.97 – 1.13] (XXI століття) не перекриваються. А це означає, що тексти належать до різних функціональних стилів. Межі 2σ інтерквантильного інтервалу для художніх текстів XIX та XXI століття перетинаються, [0.99 – 1.11] і [0.97 – 1.13] а отже, вони статистично однорідні щодо апроксимаційного параметра *s,* і їх можна об'єднувати в одну вибірку. Для німецькомовних текстів, так само як і для англомовних текстів, значення апроксимаційного параметру *s* у науковому стилі є меншим, ніж у художньому стилі.

3.2.3. Зіставлення рангово-частотного розподілу слів в україномовних наукових і художніх текстах. Статистичний аналіз текстів наукової літератури здійснено для таких жанрів наукової літератури: наукова стаття, дисертація, підручник. Наукові статті взято із журналів: "Український фізичний журнал", "Вісник ЛНУ, серія Фізична", "Фізика конденсованих високомолекулярних систем". 2σ довірчий інтервал параметра *s* для наукових статей має межі [0.830 – 0.838] (табл 3.7). Виявлені межі інтерквантильного інтервалу параметра *s* для україномовних статей є досить малими. У дисертаційних працях діапазон зміни параметра *s* є досить великим [0.78 – 0.92] (В. Вістовський *Механізми перетворення високоенергетичних електронних збуджень в багатокомпонентних йодистих кристалах,* 2004; А. Пушак *Люмінесцентна спектроскопія агрегатів лужно- та рідкісноземельних іонів у галоїдних матрицях*, 2011; П. Савчин *Люмінесцентні властивості мікро- та нанофаз, вкраплених у галоїдні кристали*, 2008; Г. Стриганюк *Випромінювальні остовно-валентні та міжконфігураційні переходи в галоїдних сцинтиляційних матеріалах,* 2003) (табл. 3.8). Параметр *s* в україномовних підручниках у межах 2σ довірчого інтервалу набуває значень [0.81 – 0.89] (табл. 3.9).

Таблиця 3.7

Україномовні тексти наукової літератури (жанр – наукова стаття)

п/н	Наукові статті в журналах	*параметр* ***s***	*параметр* ***q***	R^2
1	Український фізичний журнал	0.836±0.0008	2.023±0.039	0.946
2	Вісник ЛНУ, серія Фізична	0.835±0.0008	2.337±0.038	0.956
3	Фізика конденсованих високомолекулярних систем	0.832±0.0016	2.744±0.077	0.916
	$s_{середнє}$ ±2σ	0.834±0.004, [0.830 – 0.838]		

Таблиця 3.8

Україномовні тексти наукової літератури (жанр – дисертація)

п/н	Автори	***s***	***q***	***N***	***n***	R^2
1	В. Вістовський	0.813±0.0021	1.953±0.076	24290	5556	0.926
2	А. Пушак	0.906±0.0024	4.450±0.119	29338	4969	0.945
3	Г. Стриганюк	0.848±0.0020	2.403±0.078	27737	5880	0.934
4	П. Савчин	0.816±0.0022	2.825±0.097	22604	5110	0.932
	$s_{середнє}$±2σ	0.85 ±0.07, [0.78 – 0.92]				

Таблиця 3.9

Україномовні тексти текстах наукової літератури (жанр – підручник)

п/н	Назва книжки	***s***	***q***	***N***	***n***	R^2
1	Електрика і магнетизм Т. Г. Січкар	0.843±0.0015	2.153±0.057	35536	5933	0.960
2	Конспект лекцій з фізики	0.833±0.0016	2.497±0.068	33458	6247	0.953
3	Оптика .О Романюк	0.878±0.0009	3.021±0.046	111975	14093	0.960
	$s_{середнє}$±2σ	0.85±0.04, [0.81 – 0.89]				

Проаналізовано україномовні художні тексти XIX століття (*В. Винниченко*: Сонячна машина, Чорна пантера і білий медвідь, Між двох сил; *О. Кобилянська*: Земля, В неділю рано зілля копала, Царівна, Людина;

Б. Лепкий: Мазепа, Мотря, Бтурин; *П. Мирний*: Хіба ревуть воли, як ясла повні, Лихо давнє і сьогочасне, Казка про правду та кривду, Повія; *І. Нечуй-Левицький*: Кайдашева сім'я, Гетьман Іван Виговський, Хмари, Микола Джеря; *І. Франко:* Борислав сміється, Захар Беркут, Украдене щастя, Іван Вишенський, Мойсей) та XXI століття (*Л. Дереш*: Культ, Поклоніння ящірці, Трохи пітьми; *О. Забужко*: Музей покинутих секретів, Дівчатка, Сестро сестро, Польові дослідження з українського сексу; *Л. Костенко*: Берестечко, Сніг у Флоренції, Записки українського самашедшого, Скіфська Одіссея; *Ю.Покальчук*: Озерний вітер, Заборонені ігри, Хлопці від Катеринки; *В. Шкляр*: Ключ, Елементал, Залишенець, Чорний ворон).

Апроксимуючи ранговий розподіл слів україномовних художніх текстів, отримано параметри розподілу, представлені в табл. 3.10. Середнє значення параметра *s* для україномовних художніх текстів – 0.95, смуга коливань параметра *s*: [0.91 – 0.99].

Таблиця 3.10

Параметри апроксимованих рангових розподілів слів для текстів художньої літератури різних україномовних авторів

століття	Автор	***s***	***q***	R^2	$s_{середнє} \pm 2\sigma$
XIX	В. Винниченко	0.940	2.79	0.947	0.95 ±0.04 [0.91–0.99]
	О. Кобилянська	0.980	2.35	0,95	
	Б. Лепкий	0.937	1.33	0.95	
	П. Мирний	0.968	2.38	0.94	
	І. Нечуй-Левицький	0.927	1.54	0.94	
	І. Франко	0.934	1.67	0.93	
XXI	Л. Дереш	0.908	1.98	0.96	0.93±0.04 [0.89 – 0.97]
	О. Забужко	0.950	2.07	0.98	
	Л. Костенко	0.915	1.59	0.97	
	Ю. Покальчук	0.937	2.46	0.94	
	В. Шкляр	0.924	1.89	0.94	

Таблиця 3.11

Зіставлення інтерквантильних інтервалів параметра *s* україномовних наукових та художніх текстів

п/н	україномовні	$s_{середнє}$	$s_{середнє}\pm 2\sigma$
1.	Наукові статті	0.834	0.830 – 0.838
2.	Дисертації	0.85	0.78 – 0.92
3.	Підручники	0.85	0.81 – 0.89
4.	Художні твори	0.95	0.91 – 0.99

Зіставляючи інтерквантильні інтервали україномовних текстів наукової та художньої літератури різних наукових жанрів (табл. 3.11), можна відзначити, що між текстами наукових статей, підручників та текстами художніх творів статистичні відмінності параметра *s* є суттєвими, оскільки межі їх інтерквантильних інтервалів не перекриваються. У випадку дисертацій та художніх текстів відмінності не є суттєвими.

3.2.4. Зіставлення апроксимаційного математичного параметра *s* для стильової атрибуції англо-, німецько- та україномовних наукових і художніх текстів. Оскільки досліджувані масиви текстів не охоплюють весь обсяг наукової і художньої літератури відповідної мови, то доцільно проаналізувати рангові розподіли слів для більших масивів текстів. Зі збільшенням обсягу текстів (словника текстів) кількісні показники параметра *s* мають тенденцію до збільшення як для художніх, так і для наукових текстів. Чим більший масив текстів, тим ближчим буде кількісний показник параметра *s* до *s*, характерного для певної мови. Збільшення масиву текстів мало б зумовити зменшення відхилення параметра *s* від *s*, характерного для кожної з досліджуваних мов. Наголосимо, що обсяг досліджуваних текстів для кожної з мов становить біля 2 млн. словоформ. Зіставлення рангового-імовірнісного розподілу слів в англо-, німецько- та україномовних наукових та художніх текстах для

загального масиву всіх аналізованих текстів показало, що параметр *s* є різним для аналізованих мов. Тут зберігається тенденція для залежності параметра *s* від мови твору, знайдена для текстів менших обсягів. Параметр *s* є найбільший для англомовних наукових текстів (*s*=1.00) та найменший – для україномовних наукових текстів (*s*=0.96). Ця ж тенденція спостерігається щодо зменшення параметра *s* у напрямку англійська (*s*=1.15) – німецька (*s*=1.08) – українська (*s*=0.97) мови і для художніх текстів великого обсягу. У табл. 3.12 наведено узагальнюючі дані діапазону змін параметра *s* для англо-, німецько- та україномовних наукових та художніх текстів, які можна використовувати для стильової атрибуції текстів.

Таблиця 3.12

Межі зміни параметра *s* для наукових та художніх текстів

тексти	англомовні	німецькомовні	україномовні
НАУКОВІ XXI ст.	[0.85, ... , 1.01]	[0.86, ... , 0.94]	[0.82, ... , 0.86]
ХУДОЖНІ XXI ст.	[1.04, ... , 1.12]	[0.97, ... , 1.13]	[0.89, ... , 0.97]

Середні значення параметра *s* і довірчі інтервали ±2σ параметра *s* художніх творів XIX та XXI століть, написаних англійською, німецькою та українською мовами, перекриваються для кожної з мов (таблиця 3.12). Спостерігається загальна тенденція зменшення параметра *s* у послідовності: англійська мова (1.14) – німецька мова (1.05) – українська мова (0.95) – тексти XIX століття та англійська мова (1.08) – німецька мова (1.05) – українська мова (0.93) – тексти XXI століття. Оскільки смуги коливань параметра *s* для а) англійської [1.06 – 1.22] та німецької [0.99 – 1.11] мов (XIX століття) перекриваються; для б) англійської [1.04 – 1.12] та німецької [0.97 – 1.13] мов (XX століття) також перекриваються, то це свідчить про статистичну однорідність цих мов щодо параметра *s*. Статистичні відмінності є суттєвими у випадку україномовних та англомовних текстів, смуги коливань їх параметрів *s* не перекриваються. Смуги коливань параметрів *s*

україно- та німецькомовних текстів художньої літератури перекриваються тільки на межі цих смуг у точці дотику s=0.99 (XIX століття) та s=0.97 (XXI століття). Наявність точки дотику інтервалів дозволяє говорити про статистичну однорідність цих текстів відносно параметра s.

Таблиця 3.13

Смуги коливань параметра s для художнього стилю різних мов

Художні твори	$\bar{s}$	$2\sigma_{\bar{s}}$	$\bar{s} \pm 2\sigma_{\bar{s}}$
англомовні тексти XIX ст	1.14	0.04	1.06 – 1.22
англомовні тексти XXI ст	1.08	0.04	1.04 – 1.12
німецькомовні тексти XIX ст	1.05	0.06	0.99 – 1.11
німецькомовні тексти XXI ст	1.05	0.08	0.97 – 1.13
україномовні тексти XIX ст	0.95	0.04	0.91 – 0.99
україномовні тексти XXI ст	0.93	0.04	0.89 – 0.97

Наявність великого масиву англо-, німецько- та україномовних наукових та художніх текстів дозволила виділити перших 30 найчастіше вживаних слів у науковому (табл. 3.14) та художньому (табл. 3.15) стилях для кожної з аналізованих мов. Відповідно до результатів аналізу, спільним у науковому тексті для трьох мов є наявність у першій тридцятці лише службових частин мови та дієслова *бути*, відсутність займенників, іменників, прикметників. Спільним у художніх англо-, німецько- та україномовних текстах є той факт, що поруч зі службовими частинами мови високою частотою характеризуються займенники. Займенник *“I”, “Ich”, “ Я”* у трьох мовах в художньому стилі має подібну частоту вживання. Найчастотнішим сполучником у художніх текстах для досліджуваних мов є *“and”, “und”,“i”*.

Отримані результати узгоджуються значною мірою з опублікованими частотними словниками: A Frequency Dictionary of German by R. L. Jones and E. Tschirner, Word Frequency in Written and Spoken English by G. Leech, P. Rayson and A. Wilson. Так, для англійської мови 22 із 30 (73%) слів першої тридцятки збігається *(the, of, and, a, in, to, it, was, I, for, that, you, he, with, on,*

at, but, had, they, his, she, that); для німецької мови – 20 із 30 (66%) *(der/die/das, und, sein, in, im, ein, zu, haben, ich, sie, von, nicht, mit, es, sich, auf, an, er, als, wie).*

Таблиця 3.14

Список перших 30 найчастотніших слів наукового стилю

№	НАУКОВИЙ ТЕКСТ					
	англійська		*німецька*		*українська*	
1.	the	0,0548	die	0.0623	та	0,0217
2.	of	0,0370	der	0.0605	в	0,0185
3.	and	0,0259	und	0.0314	у	0,0158
4.	in	0,0223	in	0.0285	з	0,0142
5.	a	0,0203	von	0.0218	при	0,0116
6.	to	0,0180	mit	0.0176	для	0,0115
7.	is	0,0117	ist	0.0171	на	0,0110
8.	for	0,0105	des	0.0162	що	0,0098
9.	that	0,0095	den	0.0159	до	0,0097
10.	are	0,0087	das	0.0155	і	0,0092
11.	with	0,0081	auf	0.0153	є	0,0085
12.	as	0,0075	eine	0.0148	від	0,0083
13.	by	0,0070	werden	0.0145	а	0,0082
14.	this	0,0065	im	0.0140	за	0,0070
15.	be	0,0061	zu	0.0137	із	0,0070
16.	at	0,0058	sich	0.0128	як	0,0068
17.	on	0,0055	wird	0.0112	не	0,0065
18.	we	0,0052	bei	0.0110	між	0,0057
19.	from	0,0050	durch	0.0105	також	0,0052
20.	which	0,0047	ein	0.0104	не	0,0050
21.	an	0,0045	als	0.0104	це	0,0048
22.	can	0,0043	einer	0.0098	але	0,0047
23.	not	0,0041	nicht	0.0097	які	0,0046
24.	have	0,0040	sind	0.0097	а	0,0046
25.	or	0,0038	dem	0.0095	може	0,0046
26.	has	0,0037	aus	0.0094	бути	0,0045
27.	was	0,0036	an	0.0092	й	0,0043
28.	between	0,0034	auch	0.0089	про	0,0040
29.	these	0,0033	so	0.0084	де	0,0039
30.	also	0,0032	es	0.0083	по	0,0038

Таблиця 3.15

Список перших 30 найчастотніших слів художнього стилю

№	ХУДОЖНІЙ ТЕКСТ					
	англійська		*німецька*		*українська*	
1.	the	0,0466	und	0,0256	і	0,0175
2.	and	0,0263	die	0,0227	не	0,0158
3.	to	0,0233	der	0,0174	на	0,0135
4.	a	0,0219	er	0,0168	в	0,0110
5.	of	0,0195	sie	0,0157	що	0,0107
6.	I	0,0190	in	0,0153	я	0,0106
7.	he	0,0159	Ich	0,0123	з	0,0097
8.	was	0,0149	das	0,0119	у	0,0088
9.	in	0,0137	zu	0,0114	а	0,0077
10.	it	0,0134	nicht	0,0107	до	0,0063
11.	you	0,0107	den	0,0105	й	0,0062
12.	that	0,0100	sich	0,0103	як	0,0061
13.	his	0,0096	es	0,0096	він	0,0054
14.	had	0,0088	war	0,0083	це	0,0054
15.	on	0,0075	auf	0,0082	за	0,0046
16.	she	0,0075	ein	0,0078	але	0,0045
17.	said	0,0073	mit	0,0076	його	0,0040
18.	her	0,0070	von	0,0076	так	0,0040
19.	with	0,0069	ist	0,0058	ти	0,0039
20.	at	0,0068	hatte	0,0054	вона	0,0037
21.	but	0,0060	an	0,0053	вже	0,0037
22.	for	0,0060	dem	0,0051	ще	0,0036
23.	as	0,0059	eine	0,0050	мене	0,0036
24.	my	0,0057	was	0,0050	та	0,0036
25.	him	0,0053	wie	0,0048	то	0,0035
26.	they	0,0049	sagte	0,0048	ж	0,0033
27.	me	0,0049	als	0,0045	казав	0,0033
28.	up	0,0044	du	0,0040	було	0,0032
29.	be	0,0044	im	0,0040	все	0,0032
30.	all	0,0042	aber	0,0040	від	0,0031

Зіставлення перших 30 найчастіше вживаних слів у досліджуваних мовах дозволило виявити:

1) 11 спільних для англо-, німецько- та україномовних наукових текстів найчастіше вживаних слів: *in – in – в, also – auch – також, and – und – і (та),*

from – von – від, with – mit – з, as – als – як, is – ist – є, on – an (auf) – на, not – nicht – не, be – werden – бути, by (at) –bei – при;

2) 13 спільних для англо-, німецько- та україномовних художніх текстів найчастіше вживаних слів: *and – und – і (та), I – Ich – я, he – er – він, was – war – було, in – in – в (у), it – das – це, you – du – ти, on – auf – на, she – sie – вона, said – sagte – казав, with – mit – з, but – aber – але, as – als – як;*

3) 6 спільних найчастіше вживаних слів для наукового та художнього текстів: *and, in, on, with, as – und, in, auf, als – і, в, на, з, як.*

Якщо до числа перших 30 найчастіше вживаних слів у досліджуваних англомовних наукових текстах належать службові частини мови, то перші іменники серед найчастіше вживаних слів зустрічаються, починаючи з 50 порядкового номера, а дієслова (не враховуючи дієслова *бути*) – після 100 порядкового номера. Так, серед перших 300 найчастіше вживаних слів у науковому тексті виділено такі:

1) 37 іменників: *data, figure, probability, frequency, analysis, research, example, function, model, set, case, information, crystal, form, study, state, system, problem, methods, approach, results, test, development, evidence, factors, values, view, signal, standard, recognition, power, specific, parameters, processing, molecules, materials, sample;*

2) 12 прикметників: *new, different, same, critical, particular, possible, important, general, complex, single, larg, similar;*

3) 7 дієслів: *use, base, give, show, change, might, plot*.

Якщо до числа перших 50 найчастіше вживаних слів у досліджуваних англомовних художніх текстах належать службові частини мови та займенники, то перші дієслова (не враховуючи дієслова *бути*) серед найчастіше вживаних слів зустрічаються, починаючи з 33 порядкового номера, іменники – після 68 порядкового номера, прикметники – 72 порядкового номера. Так, серед перших 300 найчастіше вживаних слів у художньому тексті виділено такі:

1) 43 іменники: *Mr, time, man, day, Mrs, way, hand, eyes, face, room, Miss, head, back, life, house, night, people, place, door, father, heart, men, thing, moment, sir, God, lady, friend, voice, word, world, side, home, woman, mother, morning, mind, name, Andrew, day, Tom, child, wife;*

2) 15 прикметників: *little, such, good, great, old, long, still, last, quite, young, clear, whole, poor, new, kind;*

3) 47 дієслів: *say, know, see, like, come, go, make, think, hear, take, look, let, seem, tell, find, begin, give, have, ask, put, mind, know, feel, do, get, turn, stay, reply, bring, call, leave, return, want, believe, speak, love, lay, cry, hear, hope, wish, count, present, pass, keep, open, answer.*

У німецькомовних наукових та художніх текстах постежується подібна тенденція щодо розподілу перших 30 найчастіше вживаних слів. У числі перших 300 найчастіше вживаних слів у німецькомовних наукових текстах виявлено:

1) 31 іменник: *die Abbildung, das System, die Zeit, die Gleichung, die Funktion, die Physik, der Parameter, die Dynamik, die Form, die Eigenschaft, die Seite, der Abschnitt, die Berechnung, der Werte, die Darstellung, die Methode, das Beispiel, der Teil, das Ergebnis, das Verhalten, die Anzahl, das Problem, der Faktor, die Phase, die Umgebung, die Anwendung, die Bedeutung, die Frage, die Bestimmung, der Gegensatz, die Komponente;*

2) 14 прикметників: *anderen, folgenden, verschiedene, allgemeine, kleiner, direkt, stark, theoretische, gerade, gleich, gut, linear, einfache, weiter;*

3) 26 дієслів: *zeigen, ergeben, liegen, folgen, geben, bestimmen, verwenden, berechnen, entsprechen, darstellen, bestehen, lassen, erhalten, beschreiben, betrachten, bezeichnen, untersuchen, sehen, stellen, bedeuten, handeln, bilden, liefern, erreichen, annehmen, reduzieren*.

У німецькомовних художніх текстах серед перших 300 найчастіше вживаних слів вживаються такі:

1) 30 іменників: *die Frau, der Herr, das Auge, die Hand, die Mutter, der Kopf, der Gott, das Haus, die Stadt, die Liebe, das Gesicht, die Welt, das Kind, der*

Tag, der Vater, das Fenster, das Ende, die Menschen, die Stimme, der Junge, Leute, der Morgen, das Jahr, der Abend, Thomas, die Sache, das Fräulein, Christian, die Nacht, das Zimmer;

2) 4 прикметники: *gut, klein, alt, groß;*

3) 23 дієслова: *sagen, haben, sein, gehen, sehen, kommen, können, stehen, leben, wollen, wissen, nehmen, rufen, liegen, scheinen, halten, sitzen, sprechen, tretten, fragen, machen, geben, begienn*.

У проаналізованих україномовних наукових текстах перші 300 найчастіше вживаних слів представлені такими частинами мови:

1) 48 іменниками: *значення, випадок, результат, допомога, тип, метод, значень, величина, вплив, систеаи, параметр, властивість, температура, дослідження, структура, межі, умова, залежність, наприклад, визначення, вигляд, наука, поля, умова, вплив, дані, зразок, зменшення, утворення, залежність, структура, використання, зростання, характеристика, область, статті, вивчення, густина, поверхня, спектр, положення, зміна, час, урахування, точка, стан, енергія, поглинання*;

2) 2 прикметниками: *різних, експериментальних*;

3) 17 дієсловами: *показано, мають, видно, залежить, дає, визначає, наведено, становить, досліджено, отримано, свідчить, вважати, визначають, впливає, зменшується, одержано, зростає*.

Серед перших 300 найчастіше вживаних слів в україномовних художніх текстах виявлено такі найчастіше вживані слова:

1) 28 іменників: *час, життя, раз, рука, обличчя, мати, сон, нога, людина, слово, очі, світ, голос, кров, правда, жінка, рік, речі, дні, голова, чоловік, ніч, земля, сльози, дім, Бог, тиша, душа;*

2) 1 прикметник: *цілий*;

3) 17 дієслів: *казати, спати, знати, бачити, стояти, дивитися, робити, мати, могти, хотіти, піти, видно, стати, жити, чути, любити, говорити*.

Висновки до розділу 3

Залежність частоти вживання слів у тексті (розподіл Ципфа) описує взаємозв'язок частоти слова в тексті та його рангу у списку слів. Ранговий розподіл Ципфа, побудований у логарифмічній системі координат, має три ділянки: перша стосується слів у діапазоні малих значень рангового числа *k;* прямолінійна ділянка з нахилом *s≈1*; ділянка з високими значеннями рангового числа *k*. Слова з максимальною частотою, як правило, – це прийменники, частки, займенники, артиклі. Прямолінійну ділянку формують найбільш вагомі слова в тексті. Ділянка з високими значеннями рангового числа k – це слова, що зрідка зустрічаються. Існують різні модифікації закону Ципфа, які задовільно описують рангово-частотний розподіл слів у текстах лише в обмеженому діапазоні значень рангового числа *k*. Запропонована у цій дисертації модифікація функції Лавалетті $f(k;q,s,n)$ передбачає два параметри (q та s) для апроксимації рангово-ймовірнісного розподілу слів:

$$f(k;q,s,n) = \left[n(k+q)/\left(n-(k+q)+1\right)\right]^{-s} \Big/ \sum_{i=1}^{n}(i+q)^{-s} .$$

Параметр q для модифікованої функції Лавалетті запозичено із теоретично виведеної функції розподілу Б. Мандельброта. Параметр q та s описують розподіл в області слів з високою та середньою частотою вживання відповідно. Співвідношення *n(k+q)/(n-(k+q)+1)* описує ранговий розподіл низькочастотних слів без залучення додаткових параметрів. Увівши додатковий параметр n′, розглядаючи замість n величину $n + n'$, можна додатково впливати на опис рангового розподілу слів в області низькочастотних слів.

Тестування запропонованої модифікованої функції Лавалетті проведено для текстів наукового та художнього стилів з метою визначення їх стильових особливостей. Ранговий розподіл слів для англо-, німецько- та україномовних наукових і художніх текстів є різним, а кількісні показники параметри s та q є відмінними. Параметр s визначає основний нахил стрімкості спаду ймовірності появи слова у тексті. Стрімкість спаду

ймовірності появи слова виявилась меншою у наукової літературі у менших кількісних показниках параметра *s*, порівняно з художньою літературою. Виявлено тенденцію, що параметр *q* для художньої літератури більший, ніж значення цього параметра для наукової літератури. Комплексний аналіз англо-, німецько- та україномовних наукових та художніх текстів показав, що математичний параметр *s* відповідає за функціональний стиль.

Зіставлено рангово-частотні розподіли слів англо-, німецько- та україномовних наукових і художніх текстів. Середнє значення параметра *s* збільшується в напрямку українська – німецька – англійська мови. Однак, різниця між розподілами щодо параметра *s* є суттєвою тільки у випадку україномовних та англомовних художніх текстів. Їх довірчі інтервали [0.89 – 0.97] та [1.04 – 1.12] не перекриваються. Виявлено тенденцію зростання параметра *s* у випадку зростання обсягу словника для наукових та художніх текстів аналізованих мов. Параметри *s* текстів наукових статей у досліджуваних мовах суттєво відрізняються від параметру *s* художніх творів відповідних мов.

Якщо аналізувати кожен із жанрів наукового стилю, то значення їх параметра *s* є близькими між собою і знаходяться в межах 2σ інтерквантильного інтервалу, визначеного для наукового стилю англійської, німецької та української мов. Близькими між собою виявились також значення параметра *s* для художніх текстів різних століть (XIX та XXI). Ці результати в межах 2σ інтерквантильного інтервалу перетинаються, що дозволяє об'єднувати тексти різних століть в один масив і розглядати таку вибірку як однорідну.

Для кожної з досліджуваних мов визначено межі інтерквантильних інтервалів параметра *s* рангово-ймовірнісного розподілу слів для стильової атрибуції наукових і художніх текстів. Так, межі інтерквантильних інтервалів не перетинаються для англомовних наукових текстів [0.85 – 1.01] та художні текстів [1.04 – 1.12], а отже, між текстами існує значима статистична різниця: вони належать до різних функціональних стилів. Аналогічні результати

отримано і для німецькомовних (наукові [0.86 – 0.94]; художні [0.97 – 1.13]) та україномовних (наукові [0.82 – 0.86]; художні [0.89 – 0.97] текстів.

Загальна тенденція для проаналізованих наукових англо-, німецько- та україномовних текстів виявляється у наявності серед перших 300 найчастіше вживаних слів загальнонаукових термінів (*data, figure, function, model, methods, parametrs, model, function, result, values; das System, der Parametr, die Methode, das Ergebnise, die Berechnung, die Gleichung, die Berechnung; структура, метод, величина, система, параметр, результат*).

У художньому тексті для трьох досліджуваних мов серед виділених перших 300 найчастіше вживаних слів переважають слова на позначення частин тіла (наприклад, *hand, eyes, head, face, die Hand, die Augen, der Kopf, das Gesicht, обличчя, рука, нога, голова*) та періодів дня (*day, night, der Tag, die Nacht, день, ніч*), іменники *мати, батько, людина, Бог*, дієслова зору та сприйняття (*бачити, дивитись, чути*).

Основні положення цього розділу висвітлено у працях автора [21; 252].

РОЗДІЛ 4
ТЕМАТИЧНА Й АВТОРСЬКА АТРИБУЦІЯ АНГЛОМОВНИХ НАУКОВИХ ТЕКСТІВ

4.1 Тематична атрибуція англомовних наукових текстів

Метод одночасного моніторингу групування текстів та відповідних їм слів використано для тематичної атрибуції праць VI-ої Міжнародної конференції LUMDETR-2006 (Luminescent Detectors and Transformers of Ionizing Radiation – Люмінесцентні Детектори та Перетворювачі Іонізуючого Випромінювання), опублікованих англійською мовою в журналі Radiation Measurements [228]. Проаналізовано 96 статей із галузі фізики, які поділено авторами та організаторами конференції на 9 тематичних розділів (Додаток В). Роботи позначено латинськими літерами відповідно до розділів та пронумеровано для поліпшення візуалізації розподілу робіт у просторі головних компонент. Тематична атрибуція робіт у межах вузькоспеціалізованої наукової тематики конференції є ускладненою близькими характеристиками об’єктів та подібними методами дослідження. Це зумовлює високі вимоги до вибору оптимальних параметрів атрибуції та математичних методів. Для встановлення внутрішніх зв’язків між елементами текстового корпусу та словоформами сформовано матрицю даних **A**, яка містить інформацію про частоту появи слова в тексті. Матриця сформована за допомогою програми “Lexical Content Searcher” (див. розділ 2.4). Не взято до уваги слова, що вжиті лише в одній роботі, а також службові частини мови. Слід відзначити, що метод дозволив визначити службові частини мови в аналізованих текстах, роль яких у формування першої та другої компонент становить 97% (Додаток Г, рис. 1). Якщо аналізувати тексти, беручи до уваги службові слова, то їхня роль є важливою для характеристики текстів щодо виявлення закономірностей уживання цих службових слів. За таких умов прояв інших характеристик нівелюється, і для

їх детального аналізу необхідно виключити з розгляду службові слова. Інші слова виконують другорядну функцію і згруповані в околі початку координат (у Додатку Е, рис. 1 обведено пунктирним колом). Міра внеску слова в опис характеристики відображається відхиленням розташування слова вздовж осі, що відповідає характеристиці – головній компоненті. Так, слова *on, for, be, is, are, as* є найбільш вагомими для першої характеристики, виявленої в проаналізованих масивах тексту. Вздовж осі другої головної компоненти така група слів, як *and, in, at* розташована в додатньому напрямку, на відміну від від'ємного напрямку розташування групи слів *with, of, the, by*. Розташування груп слів у протилежних напрямках осі вказує на взаємообернену залежність частоти їх вживання в текстах. Слід зауважити, що одне і те ж слово може бути значимим для різних компонент. Так, слово *be* є вагомим для PC-1 (значне відхилення в додатньому напрямку вздовж осі PC-1), тоді як для PC-2 воно нівелюється (проекція положення слова на вісь PC-2 є малою). А слово *and* є валідним для обидвох компонент.

У повторному аналізі текстів, за умови неврахування службових слів, матриця даних **A** мала розмірність 96×4847. Сума частот слів у кожному тексті нормовано до одиниці з метою врахування різної кількості слів у текстах. Розподіл текстів та відповідних їм словоформ побудовано з урахуванням 20 головних компонент, на які припадає 76,37% опису даних. Спроба розмежування текстів за першою головною компонентою не дала задовільних результатів. Це дає підстави припустити, що словоформи, які формують дану компоненту, притаманні усім текстам.

Метод одночасного моніторингу групування текстів та відповідних їм слів розташовує в просторі тексти відповідно до їх значення для головних компонент. Завдання дослідника – серед виділених груп виокремити такі області у просторі головних компонент, що репрезентують тексти з однаковою тематикою. Така процедура може бути зроблена, якщо наперед відомо групування текстів за певною характеристикою, в нашому випадку за тематикою. Ці області виділяються з умовою, щоб у них зосереджувалась

максимальна кількість текстів однієї тематики (Додаток Г, рис. 2). Знаючи такі області, можна встановити, за якими критеріями відбувається об'єднання текстів, оскільки метод дозволяє одночасно відстежувати розподіл словоформ і текстів у просторі головних компонент. Аналізуючи виділені області, можна встановити, чи правомірне розміщення тексту в цій області, та виділити приховані ознаки об'єднання текстів. Маючи розбиття відомих текстів на групи у просторі головних компонент, можна здійснити класифікацію невідомих текстів, застосувавши цей метод. У цьому випадку до масивів відомих текстів додають невідомі тексти і до сукупності цих текстів застосовують аналіз головних компонент. Тоді попадання невідомих текстів у відому область дозволить провести їх тематичну ідентифікацію.

Найкраще тексти розділяються на групи, якщо їх представити у просторі другої, третьої та четвертої головних компонент (Додаток Г, рис. 2) [22]. Тут можна чітко виділити три області B, C та G, в яких групуються тексти, що відповідають тематичним секціям конференції: "B. Дозиметричні матеріали", "C. Запасаючі та інші фосфори" і "G. Домішки, дефекти, пастки". Так, в область B із 14 представлених у ній статей потрапляє 10 статей, в область C із 10 – 8, в область G із 16 – 6 статей. Крім згаданих трьох областей, можна виділити досить компактну область, у якій зосереджено тексти, що належать тематичним секціям D, E, F, H, I. Виділяється також область Z, де представлено поодинокі роботи з різних тематичних секцій.

У тому ж самому масштабі побудовано розподіл словоформ у просторі головних компонент (Додаток Г, рис. 3). У цих координатах виділяються словоформи, які є найбільш значимими для формування цих головних компонент. Аналізуючи вагомість текстів (Додаток Г, рис. 2) та словоформ (Додаток Г, рис. 3) для головних компонент, можна відзначити таке. Словоформи другої головної компоненти найбільш притаманні для групи робіт B "Дозиметричні матеріали". Тематичну секцію "Домішки, дефекти, пастки" (група G) формують словоформи третьої головної компоненти. Четверта головна компонента відповідає характеристикам, які найбільше

притаманні роботам групи C (секція “Запасаючі та інші фосфори”). Накладання просторів (D, E, F, H, I), що належать працям із різних тематичних секцій, попадання окремих статей в області, їм не притаманні (наприклад, C36 потрапляє в область B), виділення нової області Z, зумовлені все ж таки певним перекриттям тематик секцій або описом нових експериментальних методів, матеріалів або підходів для пояснення фізичних процесів. Використовуючи метод головних компонент, виявлено спорідненість статті C 36 (T. Rivera “Thermoluminescence properties of copper doped zirconium oxide for UVR dosimetry”) із групою робіт B. Автор та організатори конференції віднесли дану роботу до групи C, тоді як навіть сама назва роботи C 36 свідчить про її тематичну належність до секції B “Дозиметричні матеріали”. Роботи груп B та C справді подібні за тематикою, однак їх відрізняють словоформи четвертої головної компоненти (Додаток Г, рис. 3).

Отже, із 9 тематичних секцій конференції у просторі головних компонент чітко виділяються три групи: B, C та G. Тексти груп B та C виділяються найбільшою мірою, що пояснюється новизною напрямів досліджень відповідних тематичних секцій і, відповідно, введенням нової термінології, яка не притаманна для інших тематичних напрямків. Групи D, E, F, H, та I виявились найменш рознесеними у просторі головних компонент, оскільки тексти у відповідних тематичних напрямах характеризуються спільними, усталеними та широко розповсюдженими для даної тематики дослідженнь словами. Неочікуваним виявився розподіл 16 праць (область Z) із різних тематичних секцій у додатному напрямку осі другої головної компоненти (Додаток Г, рис. 2). Для з’ясування причин такого розподілу проаналізовано значення словоформ для другої головної компоненти (Додаток Г, рис. 3). Додатній напрямок осі другої головної компоненти характеризується словами, що вказують на досліджувані матеріали – плівки та монокристали (*films, SC, single crystals*) сполук гранатів (*YAG, YAP, LuAG*), активованих іонами церію (*Ce, Ce^{3+}*) (Додаток Г, рис. 3). Ця група слів

відображає об’єкти дослідження, характерні для текстів з області Z. РС-2 притаманна також група слів, яка описує основні параметри люмінесцентної спектроскопії *(emission, luminescence, excitation, spectra, decay*).

Прізвище Ю. Зоренка (Zorenko) виявилось єдиним, що виокремлюється з групи прізвищ для області Z. Роботи саме цього автора (A2, A3, A4, A10, C32, C33, C38, G78, H89, I95) присвячені люмінесцентній спектроскопії кристалів та плівок сполук гранатів, активованих іонами церію (Додаток Г, рис. 3). Таким чином, в область Z потрапляє 10 з 10-ти праць, у яких співавтором є Ю. Зоренко. Ці статті не були одноосібними, і тому не варто в даному випадку робити передчасні висновки щодо прояву авторського стилю. Швидше – це відображення тематичних особливостей текстів, а саме: об’єктів дослідження – кристалів гранатів. Щоб переконатись у тому, що основою для об’єднання цих статей є сполуки гранатів до масиву текстів додана ще одна стаття Ю. Зоренка (**Z-X**) “Single-crystalline Films of Ce-doped YAG and LuAG Phosphors: Advantages Over Bulk Crystals Analogues”, присвячена люменісценції кристалів гранатів [262]. Результат застосування методу аналізу головних компонент є точка **Z-X** (Додаток Г, рис. 2), яка також потрапляє у попередньо виділену тематичну область **Z.** В області Z поряд із виділеними статтями Ю. Зоренка (див. Додаток Г, рис. 2) виділяються також: 1) роботи наукової групи академіка Б. Гриньова (A14, G72, G74, G75), присвячені люмінесценції домішкових іонів церію (Ce, Ce^{3+}); 2) роботи В. Панкратова (D40) та М. Грінберга (F59), у яких досліджують кристали гранатів, активовані іонами Ce^{3+}. Отже, можемо зробити висновок, що ця компонента об’єднує тексти та термінологію, що відповідають люмінесцентній спектроскопії іонів церію у кристалах.

Спільне представлення у просторі головних компонент областей, що відповідають за групування текстів в ту чи іншу тематичну секцію та словоформ, розташованих відповідно до їх внеску в головні компоненти, дозволяє виділити слова, характерні для текстів тематичних секцій конференції. Характерними словами для групи B є: *radiation, irradiation,*

heating, glow, peak, dose. Дійсно, саме ці терміни описують особливості параметрів дозиметричних матеріалів (*glow, peak, dose*) чи методів їх дослідження (*radiation, irradiation, heating*). Терміни *storage, X-ray, PSL (photostimulated luminescence), phosphors* та формули хімічних сполук чи елементів *CsBr, $CsEuBr_3$, $EuAl_2O_4$, Cu* є найбільш вживаними для текстів групи С. Ці слова та формули описують об'єкти та методи дослідження запасаючих фосфорів. Для групи G характерними є: *TSL (thermoluminescence), temperature, deffects, $PbWO_4$*. Об'єднавчим для текстів цієї групи виявились кристали $PbWO_4$. На час проведення конференції цей кристал був у центрі уваги багатьох дослідників, оскільки він був обраний як основний матеріал для експериментів з пошуку бозонів Хігса, що проводились у Центрі Європейських Ядерних досліджень на прискорювачі в Женеві (Швейцарія). Для групи Z характерні: гранати (*yag, luag*), домішкові центри (Ce^{3+}), плівки (*film*); їх внесок у другу компоненту є найбільшим.

4.2 Виділення тематичних напрямків за заголовками, анотаціями та тезами наукових доповідей

Дослідник не завжди має доступ до наукових праць, опублікованих у тому чи іншому журналі. Більш доступними є менші за розмірами від статей їх заголовки, анотації або тези доповідей. Оскільки вони зберігають ключові слова статей, то існує ймовірність, що і у випадку меншого обсягу тексту, можна провести тематичну атрибуцію. Тому важливо дослідити ефективність методу аналізу головних компонент, використовуючи тексти меншого обсягу. Для проведення експерименту сформовано текстову вибірку праць конференції LUMDETR-2006, у яку увійшли 36 пар "тези-стаття" із різних тематичних секцій конференції [23; 141]. Тези доповідей повинні бути близькими за змістом до опублікованих за матеріалами конференції статей, і тому слід очікувати близького розташування елементів пари "теза-стаття" у просторі головних компонент. Хоча, звичайно, є відмінності, які полягають у

різній композиції текстів цих двох жанрів [125]. Текст статті структурований на “вступ”, “огляд літератури”, “методика дослідження” тощо, тоді як тези доповіді не мають згаданих структурних частин і є значно меншими за обсягом. Тези чутливі до таких чинників, як вид конференції, специфіка наукової дисципліни, критерії відбору тез тощо. Є. Візе зазначає, що мова тез доповідей вирізняється високим рівнем деперсоналізації, ознаками якого є часте вживання конструкцій із модальними дієсловами та інфінітивами [16].

Щоб дослідити міру подібності між тезами доповіді та статтею створено матрицю даних **A** із розміром (72×4830), де 72 – це кількість текстів, а 4830 – кількість словоформ (не враховано словоформи, які зустрічались у тексті тільки один раз). Найбільш інформативним виявилося представлення результатів розрахунку для пар “теза-стаття” у системі координат головних компонент PC-2, PC-3, PC-5 (Додаток Г, рис. 4). Дійсно, виявляється, що деякі статті та тези групуються поряд. Це можна використати для того, щоб дослідити, якою мірою опублікована стаття відрізняється від заявлених тез. Виходячи з розподілу робіт уздовж координатних осей PC-2, PC-3, PC-5 (Додаток Г, рис. 4), можна зробити такі висновки щодо принципу їх розподілу на тематичні групи: 1) найбільший внесок у формування головної компоненти PC-2 вносить термінологія праць групи G “Домішки, дефекти, пастки”; 2) тематичні особливості праць груп H “SUPERLUMI Experiment” та B “Дозиметричні матеріали” характеризує компонента PC-3; 3) працям групи C “Запасаючі та інші фосфори” притаманна термінологія, що формує компоненту PC-5. Просторовий розподіл груп B, C та H зумовлений різними об’єктами та методами, що притаманні цим тематичним напрямкам. І, навпаки, перекриття груп текстів F та G – результат використання спільної термінології, що свідчить про подібність розглянутих об’єктів та методик дослідження.

Велика відстань між текстами деяких із розглянутих пар “теза-стаття” може відображати зміни, внесені у статтю, порівняно з текстом тез (Додаток

Г, рис. 4). Цілком природною є можливість правки авторами початкової ідеї досліджень та перегляду зроблених ними висновків у період між реєстрацією тез доповіді та редагуванням остаточної версії статті. Слід врахувати також редакторські правки вже опублікованих статей.

Незважаючи на те, що обсяг тез є меншим від обсягу статті, отримано якісно близькі тематичні розподіли текстів тез та статей. У такій ситуації виникає запитання щодо впливу обсягу вибірки на якість тематичного розмежування текстів. Щоб виявити залежність якості тематичної атрибуції від обсягу вибірки, було сформовано вибірку, до якої увійшли тексти статей, їх анотації та заголовки. Окремо взяті заголовок, анотація або абзац із тексту мають структурну завершеність, є структурно-смисловими частинами тексту і можуть становити повністю завершений текст [50, с. 57]. Заголовок та анотацію І. Колегаєва визначає так: 1) анотація – допоміжний компонент мегатексту, в якій зміст є згорнутим, скомпресованим варіантом змістової структури основного тексту, її функція – надати читачеві можливість зробити висновок щодо подальшого ознайомлення зі статтею; 2) заголовок – гранично згорнута форма презентації надлінійної фрагментної структури твору, він у більш стиснутій формі, ніж анотація, повідомляє про тематику основного документа, його функція – повідомити про тематику основного тексту [62, с. 74-82]. Заголовок може виконувати нормативну, організуючу, інформативну, сигнальну, рекламну, експресивно-апелятивну функцію [9, с. 151]. Так, для наукових текстів заголовок виконує інформативну та сигнальну функції. Такі особливості наукових заголовків, як інформативність, точність, лаконічність, узгодженість із змістом наукового твору зазначали Васильєв Ю. А. [14], Медведэв А. Р. [79], Яхонтова Т. В. [124]. Заголовок статті повинен відповідати таким вимогам, як висока інформативність і чітка та стисла форма викладу [103, с. 145–147]. Коваленко А. М показав, що поєднання структурних, семантичних та прагматичних особливостей заголовка дозволяє визначити його роль у репрезентації текстової інформації, а також розкрити роль заголовка у

смисловій організації всього мікротексту [59]. Анотація повідомляє тему та отриманий результат наукового дослідження, подає основний зміст статті, готує читача до сприйняття основного тексту праці [52, с. 55; 126].

Для аналізу розподілу тріад "заголовок–анотація–стаття" у просторі головних компонент випадково зі збірника LUMDETR 2006 вибрано 5 статей із різних тематичних груп і відповідні до них анотації та заголовки. Представлення результатів розрахунку найбільш інформативне у координатній системі головних компонент PC-4, PC-5, PC-7 (Додаток Г, рис. 5). Слова у заголовку статті є близькими до ключових слів анотації. Так, у статті групи B (*B23*) назва статті і перше речення з анотації практично збігаються. 9 із 13 слів заголовку зустрічаються в анотації. Нижче подано заголовок та анотацію цієї статті, де підкреслено слова, спільні для заголовку та анотації. "A model for distinguishing between static and dynamic exposure of personal thermoluminescence dosemeters" (Kopec, 2006, Додаток В). *To distinguish between static and dynamic (normal) exposure of personal TL dosemeters, a model of radiation deposition and of TL light transport in the TLD dosemeter is proposed. The RADOS dosemeter badge using MCP-N (LiF: Mg, Cu, P) TL detectors with standard filters replaced by special Pb and Cu filters with a pattern of holes or inserts, was modeled. The photon radiation transport in the dosemeter and energy deposition in the TL detector, were simulated by the Penelope Monte Carlo transport code. The model of TL light transport within the TL pellet takes into account the distribution of energy deposition in the TL detector, light self-absorption in the detector and reflection of TL light off the heating planchet. The shape and hole pattern of the filters were optimized with respect to best distinction between static and dynamic exposures. The results of calculations were verified experimentally by exposing RADOS badges with modified filters to beams of low energy X-rays directed at various angles.*

Для тріади "заголовок–анотація–стаття" у статтях із секцій C та G характерним є багаторазове повторення термінів заголовка в реченнях анотації. Так, формула сполуки $EuAl_2O_4$ із заголовка статті *C33* була вжита 6

разів у 7 реченнях анотації. “High-pressure luminescence spectroscopy of $EuAl_2O_4$ phosphor” (Zorenko, 2006, Додаток В). *$EuAl_2O_4$ powder phosphor was prepared by solid-state reaction of EuO and Al_2O_3 oxides in vacuum. The influence of conditions of preparation on spectral lineshape of Eu^{2+} emission was analyzed. It was found that the fluorescence spectra of vacuum-prepared $EuAl_2O_4$ samples at 300 K present the superposition of three bands peaked at 430, 500 and 528 nm, corresponding to the 4f65d1>4f7(8S7/2) transition of Eu2+ ions in the different sites of $EuAl_2O_4$ lattice. The luminescence of Eu^{2+} centers in $EuAl_2O_4$ host was also studied using the high-pressure spectroscopy up to 67 kbar. It was found that the bright green-yellow fluorescence of $EuAl_2O_4$ at 300 K in the band peaked at 520-530 nm range can be presented by superposition of two Gaussian sub-bands. The different pressure shifts – 23 cm-1/kbar and – 27 cm-1/kbar for two sub-bands were found. Such a structure of the emission spectrum was attributed to the existence of two different Eu^{2+} centers in the $EuII^{2+}$ sites of $EuAl_2O_4$ lattice with higher coordination number.*

Позначення хімічного елемента Ln^{2+} із заголовку статті *G70* зустрічається у кожному реченні анотації. “Influence of the crystal structure on the stability of Ln^{2+} in strontium borates” (Dotsenko, 2006, Додаток В). *The results of luminescence measurements on Ln (Eu, Yb) doped alkaline earth (M=Ca, Sr) borates $M_3(BO_3)_2$, MB_2O_4, $M_2B_5O_9X$ (X=Cl, Br), MB_6O_{10}, MB_4O_7 after high-temperature annealing in various atmospheres are reported and discussed. The stability of Ln^{2+} (Ln=Eu, Yb) is found to increase in the sequence $Sr_3(BO_3)_2$ <SrB_2O_4, SrB_6O_{10}, $Sr_2B_5O_9X$ < SrB_4O_7. This observation is explained based on local balance considerations and the differences in charge-compensating mechanism for Ln^{2+}. A simple criterion is proposed to predict the stability of Ln^{2+} in alkaline earth borates.*

Водночас для статті Н87 не притаманні повторення термінології заголовка в анотації. Із 9 слів заголовку лише 3 слова зустрічаються у двох із п’яти речень. “Luminescent properties of Yb-doped $LaSc_3(BO_3)_4$ under VUV excitation” (Guerassimova, 2006, Додаток В). *Ytterbium doped borate crystals*

are promising laser media, e.g. in $LaSc_3(BO_3)_4$ (LSB) matrices large distance between ytterbium ions results in reduced concentration quenching of the ytterbium f-f luminescence (Petermann et al., 2005). Yb^{3+} ions in complex oxides in addition to the 4f - 4f transitions often manifest fast charge transfer luminescence (CTL) in the UV-visible range. In some borates it was not observed at all, like in orthoborates of Sc, Y and La (Van Pieterson et al., 2000). In haloborates $Sr_2B_5O_9X$, where X = Cl, Br, the UV/visible luminescence was attributed to ytterbium CTL though it looked substantially different from other matrices (Dotsenko et al., 2002); while in oxyborate $Li_2Lu_5O_4(BO_3)_3$ "classical" CTL was observed (Jubera et al., in press). In this work the luminescence properties of another borate, namely LSB doped by Yb are presented.

Статті, анотації та заголовки розмістились уздовж чітко сформованих напрямів (Додаток Г, рис. 5). Найбільш віддаленими від центру координатної системи виявились заголовки статей (t A, t B, t C, t H, t G), що може свідчити про більшу міру прояву тематичних характеристик у заголовках. Дані характеристики виявляються меншою мірою в анотаціях і ще меншою – у текстах статей. Подібна тенденція властива також просторовому розподілу пар "теза–стаття" (Додаток Г, рис. 4), де більші внески у головні компоненти спостерігаються для тез. Так, тріада "artB – absB – tB" витягнута вздовж осі PC-4, і це може бути зумовлено наявністю в заголовку слів, що притаманні компоненті PC-4. Текст B23 належить секції "Дозиметричні матеріали", для якої ключовим словом є *dosimeter*. Внесок слова *dosimeter* у заголовку tB становить 0.08, тоді як частка цього слова у тексті тези absB є 0.018. Оскільки частка слова *dosimeter* в заголовку тексту є більшою, ніж в анотації, то і внесок *dosimeter* у компоненту PC-4 для заголовку є більшим, ніж для анотації. Така контрастна побудова, коли два тексти однакової тематики мають різний обсяг, може бути використана для визначення прихованого змісту головних компонент. Якщо ж перейти до аналізу самих назв, то їх інформативність щодо групування текстів втрачається, оскільки внески слів у різних текстах стають співвимірними.

У послідовності “стаття-тези-анотація-заголовок” частка ключових тематичних термінів зростає. У статтях поряд із термінами, визначальними для даної тематики, наявні й інші слова, притаманні опису інших характеристик. Отже, заголовки та анотації можна розглядати як елементи з високим умістом ключових слів, які відіграють визначну роль у тематичній атрибуції. Розгляд статей, анотацій до них та заголовків сприяє виявленню лексичних множин, що окреслюють тематику, і тим самим покращує розуміння прихованого значення головних компонент.

Висновок про необхідність зіставлення текстів, анотацій та заголовків у просторі головних компонент узгоджується із роллю назви, анотації, тексту у передачі інформації. З. Тураєва зазначає важливе місце заголовку у формуванні єдності тексту, адже заголовок має виражати основну мету повідомлення, викликати зацікавленість до тексту, актуалізувати найбільш важливу інформацію тексту [114, с. 52]. Підраховано, що середня кількість слів у заголовку наукової статті становить 6 слів, а середнє число значимих слів – 4,92 [105]. Отже, заголовок, анотацію, тези можна вважати ступенями компресії тексту.

4.3 Інтегрована методика аналізу авторської атрибуції наукових текстів методом одночасного моніторингу групування текстів та відповідних їм слів із залученням послідовності вживання слів

Процес авторської атрибуції ускладнюється, як це наголошувалось у попередніх розділах, у випадку аналізу наукових текстів, де автор наукової статті обмежений у стилі та засобах вираження своїх думок – він змушений використовувати вже усталену термінологію, висловлювати свої ідеї та результати точно, стисло, однозначно. З метою встановлення особливостей авторської атрибуції англомовних наукових текстів вперше використано інтегрований підхід із залученням методу одночасного моніторингу групування текстів та відповідних їм слів, а також лінгвістичного параметра

послідовності вживання слів [24; 250]. Проаналізовано 16 статей, присвячених тематиці люмінесцентної спектроскопії, що є результатом колективних експериментальних досліджень зі спільним співавтором Г. Стриганюком, опублікованих англійською мовою у реферованих журналах. Статті поділено на дві групи (Додаток Д): група А представлена статтями, в яких даний співавтор брав участь в обговоренні статті, але не проводив кінцевої правки тексту статті; група В поєднує статті, остаточне редагування яких проводив Г. Стриганюк. Для визначення ефективності застосування методу головних компонент до авторської атрибуції наукових текстів, співавтор цих статей Г. Стриганюк попередньо провів класифікацію статей на дві групи (А і В). Аналізовані статті позначено латинськими літерами відповідно до груп та пронумеровано (Додаток Д). Провівши статистичне опрацювання вибраного масиву з 16 текстів, створено словник, що налічує 5000 словоформ, сформовано частотну матрицю **A**(16×5000) (кількість текстів × кількість словоформ). Для виявлення умов покращення ідентифікації авторського стилю використано метод одночасного моніторингу групування текстів у поєднанні з параметром послідовності вживання слів (послідовність з 1, 2, 3, 4, 5 та 6 слів) У табл. 4.1 представлено кількість виявлених та прийнятих до розгляду послідовностей вживання слів (n-грам) частотної матриці **A** (*n*×16). До розгляду бралась послідовність вживання з 1, 2, 3, 4, 5 та 6 слів у 2-х та більше текстах. Для послідовності з одного слова до розгляду не брались службові слова.

Таблиця 4.1

Послідовності вживання слів, прийняті до розгляду

Послідовність	одного слова	двох слів	трьох слів	чотирьох слів	п’яти слів	шести слів
виявлено	5000	20565	30694	33287	32007	29301
прийнято до розгляду	1897	4532	3300	1733	878	492

У випадку аналізу головних компонент з послідовністю з одного слова використано 10 головних компонент, що описують 92,13% усіх змінних. Внесок перших чотирьох компонент у модель становить 73,18%. Поєднання другої, третьої та четвертої головних компонент виявилося найбільш інформативним для побудови просторового розподілу досліджуваних текстів (Додаток Г, рис. 6). Перша головна компонента не розглядалась, оскільки словоформи, які є найбільш значимими у її формуванні, характерні для опису властивостей, що притаманні всім текстам. У просторі головних компонент PC-2, PC-3, PC - 4 чітко розмежовуються дві області, в яких зосереджені тексти групи А та групи В. Основний внесок у розподіл статей дає компонента PC-2. З огляду на це можна сказати, що друга головна компонента PC-2 відповідає характеристикам, які найбільшою мірою розділяють статті. Виняток складає стаття В8, яка розміщена поблизу статті А2. Оскільки статті А2 та В8 присвячені одній тематиці, то їх близьке розташування може вказувати на суттєве значення тематичної термінології у головних компонентах PC-2, PC-3, PC-4. Отже, праці, редаговані Г. Стриганюком (група В), вдалось чітко виділити з-поміж усіх праць за його співавторством. За таких підходів ефективність класифікації текстів складає 90%. Розглянуту класифікацію можна розглядати навчальну вибірку. Класифікація на навчальній вибірці дозволяє окреслити просторову область головних компонент, в якій локалізуються праці Г. Стриганюка. Потрапляння статей тестової вибірки в цю область свідчило б про їх приналежність до статей Г. Стриганюка.

Метод одночасного моніторингу групування текстів та відповідних їм слів дозволяє побудувати розподіл слів у просторі головних компонент і проаналізувати слова, що є значимими для формування тих чи інших компонент (характеристик тексту). Зокрема, виявити роль внеску термінів, формул хімічних елементів та сполук у головні компоненти. Дійсно, тут можна відзначити значний внесок тематичної термінології у формування головних компонент (Додаток Г, рис. 7). Для компоненти PC-2 характерні

формули хімічних сполук та елементів і хімічна та фізична термінологія (*K_2LaCl_5, Ce_3, Ce, $LaCl_3$, KCl, NaCl, host, mol, microphase*). Саме сюди потрапляють статті Г. Стриганюка, в яких він досліджує згадані кристали та домішкові центри (Додаток Г, рис. 6). Наприклад, *B1* (Luminescent characteristics of pure and Ce doped K_2LaCl_5 phase in KCl host), *B2* (Luminescence of Pr^{3+} doped K_2LaCl_5 microcrystals encapsulated in KCl host), *B7* (Luminescence of Ce doped $LaCl_3$ microcrystals incorporated into a single-crystalline NaCl host). Для напрямку РС-4 характерні терміни (*crystals, valence, core, band, cvl, energy*) (Додаток Г, рис. 7), де групуються роботи А2, А3, В8 (Додаток Г, рис. 6), ключовою особливістю яких є дослідження остовно-валентної люмінесценції (*core valence luminescence – CVL*). Вздовж осі РС-3 виділяються *4f, 5d, Pr^3, Lu^3, Gd^3, ions*. Ці слова характерні для текстів, що групуються вздовж РС-3. Отже, розподіл статей певною мірою може бути зумовлений спорідненістю об'єктів та тематики досліджень.

Щоб уникнути групування робіт за об'єктом та тематикою досліджень, вилучено згадані терміни та формули хімічних сполук та елементів (усього 251 словоформа) із частотної матриці **A**. Це обґрунтовано, бо терміни – це слова, притаманні певній науковій галузі і позбавлені емоційного забарвлення. Вони менше впливають на авторський стиль. Розмір матриці був зменшений від 16×1897 до 16×1646. Перерахована після цього модель в межах 10-ти головних компонент описує 91,94% змінних, причому 75,52% припадає на перші чотири компоненти. Після вилучення з розгляду назв сполук та згаданих термінів модель головних компонент як і в попередньому випадку забезпечує просторовий розподіл статей Г. Стриганюка на групу А і групу В. Стаття В8 залишилась поміж статей групи А у просторі головних компонент, незважаючи на виключення з розгляду тематичної термінології. В8 є однією з перших робіт Г. Стриганюка у співавторстві, тому відокремлення роботи В8 від групи В можна пояснити становленням його авторського стилю. Фактор становлення та зміни стилю є особливо важливим у випадку, якщо автор пише статтю мовою, яка є для нього іноземною.

Найбільш чітко групування статей досягнуто для послідовності з чотирьох слів, що представлено у координатах PC-1, PC-2, PC-3 (Додаток Г, рис. 8). Слід зауважити, що кількість виявлених елементів n-грам слів сягає максимуму саме для послідовності вживання із 4 слів (табл. 4.1). Усі статті (статті групи А), автором яких не є Г. Стриганюк, згруповані у просторі головних компонент в області початку координат. Таке групування можливе, коли досліджувані тексти не мають спільних авторських характеристик стосовно цього автора. Саме група А і сформована із робіт, написаних різними авторами. Їх внесок у формування згаданих компонент є найменшим, у той час як статті групи В відрізняються значимістю для першої компоненти.

Аналіз послідовності з чотирьох слів, що входять у формування PC-1, PC-2, PC-3, дає підстави припустити, чому статті групи В виокремлюються. Послідовність з 4 слів, яка характерна для PC-1, виражає припущення (*is expected to be, may correspond to the, may be caused by, may be considered as, can be attributed to*), PC-2 – описує спостережувані об'єктів (*have been found to, is revealed for the, is more pronounced for, appears due to the, upon the excitation in*), а PC-3 стосується представлення та порівняння результатів (*it is concluded that, is estimated to be, is evaluated to be, a good agreement with, have been revealed for*). Таке вживання припущень властиве автору, про що свідчить поширення модальних слів *may* та *can* у виявлених 4-грам слів. Можна вважати, що для Г. Стриганюка властиве вираження припущень при обговоренні результатів та уникнення представлення результатів у категоричній формі. Саме ці ознаки є притаманними статтям групи В та характеризують авторський стиль Г. Стриганюка. Нижче подано приклади вживання характерних послідовностей з чотирьох слів, які формують першу, другу та третю головні компоненти. Цей параметр дозволив виокремити з масиву текстів статті Г. Стриганюка (Додаток Д), які автор написав особисто:

а) опис об'єктів: *For most Yb-doped materials the FWHM of CTL and CT absorption band have been found to be comparable, and the top of the valence*

band (VB) is consequently considered as the initial state for the charge transfer (стаття В6, Додаток Ж). *Efficient excitation is revealed for the STE emission of K2LaCl5 microphase up to 7.7 eV (Fig. 2b, curve 1) within the transparency range of the KCl host* (стаття В1, Додаток Д). *The low-energy 338 nm band is more pronounced for the nanosized samples (curve 2)* (стаття В4, Додаток Д). *The excitation of Ce3+ 5d-4f emission appears due to the recombinational mechanisms of energy transfer from LaPO4 host to the impurity Ce3+ ions upon the excitation within 8-20 eV range (figure 3(b, c))* (стаття В4, Додаток Д). *The influence of the surface defects is expected to be most essential upon the excitation in the range of host absorption when the depth of excitation quanta penetration decreases* (стаття В4, Додаток Д).

б) вираження припущень: *The increased strength of crystal field in microcrystalline LaCl3 may be caused by the reduced distance between Ce3+ion and Cl- ligands* (стаття В7, Додаток Д). *This fact may be considered as the evidence for the preferable incorporation of Pr3+ impurity ions into K2LaCl5 microphase* (стаття В2, Додаток Д). *The peak at 8.36 eV can be attributed to the near-activator exciton, and the rise component, revealed in the decay kinetics of Yb3+ CTL upon excitation at 8.36 eV, may correspond to the formation of a Yb2+ CT cluster due to the nonradiative decay of the near-activator exciton on the impurity Yb3+ ion* (стаття В1, Додаток Д).

в) представлення результатів: *The quenching temperature of Yb3+ CTL has been estimated to be about 190 K for YbP_3O_9 metaphosphate* (стаття В9, Додаток Д). *Spectral position of the second broad band (5.7-7.4 eV) in the excitation spectrum of Ce^{3+} luminescence (fig. 3 a) is in a good agreement with the range of Pr^{3+} 4f 2 > 4f 5d absorption* (стаття В7, Додаток Д). *Changes in the C parameter have been revealed for the YbP_3O_9 structure* (стаття В2, Додаток Д).

4.4 Ентропія як метод авторської атрибуції наукових текстів

У дисертації вперше запропоновано поєднати метод ентропії, зокрема її різновид – дивергенцію Кульбака-Лайблера, з таким лінгвістичним параметром, як послідовність вживання слів (n-грам слів) для здійснення авторської атрибуції англомовних наукових текстів [25; 251]. Для апробації алгоритму порівняння текстів та вибору оптимального опорного тексту використана навчальна вибірка, яка налічує 40 праць англійською мовою (Додаток Е) з галузі люмінесцентної спектроскопії, опублікованих різними науковими групами в реферованих журналах. Праці відібрані після попереднього вивчення сфер діяльності наукових груп. Вони розділені на 4 групи (по 10 статей у кожній групі) відповідно до їх авторів (керівників груп): проф., д-р. P. Dorenbos, проф., д-р. A. Meijerink, д-р. G. Stryganyuk, проф., д-р. G. Zimmerer. Тексти (D*i*, M*i*, S*i*, Z*i*; *i*=1..10) позначено відповідно до першої літери прізвища автора та пронумеровано в межах кожної групи відповідно до їх розміру в порядку його спадання, наприклад, обсяг тексту D1 є більшим за обсяг тексту D2. Загальний обсяг словника проаналізованих робіт налічує 11385 словоформ. Тестова вибірка містила дванадцять авторських статей P. Dorenbos, тексти якої позначенні DA1, ... , DA12.

Дивергенцію Кульбака-Лайблера (формула 2.2) обчислено для 40 досліджуваних текстів з метою виявлення їх спорідненості в межах наперед окреслених 4-х груп авторів. До розгляду взято 360 “функціональних слів” (Додаток Є), які були апробовані у роботі [256, p. 182], де показано, що вони є успішними параметрами для проведення авторської атрибуції художніх текстів. Дивергенцію Кульбака-Лайблера обчислено за словником “функціональних слів” для 4-х опорних текстів (D1, M1, S1, Z1), вибраних з огляду на їх максимальний обсяг у межах відповідної групи. У Додатку Ж (рис. 1) відображено результати оцінки спорідненості (розбіжності) текстів, отримані методом обчислення дивергенції (2.2) порівняно зі словником “функціональних слів”.

Якість (ефективність) класифікації (q) визначали за частотою правильно класифікованих текстів у навчальній вибірці в ході кожного експерименту (або за частотою, вираженою у %). Наприклад, навчальна вибірка складається із 40 текстів 4 авторів (по 10 статей кожного автора). Для визначення ефективності класифікації за автором кожну статтю автора (наприклад, P. Dorenbos) порівнювали з усіма іншими, використовуючи ентропійну процедуру класифікації. У випадку q =1,0 (q =100%) всі 10 статей P. Dorenbos мали би бути розміщені в позиціях 1, ... 10 відповідно до зростання дивергенції Кульбака-Лайблера. Якщо, наприклад, серед перших 10 статей знаходиться тільки 9 статей P. Dorenbos, то такій класифікації приписується ефективність q = 0,9 (90%).

Групуванням текстів за результатами обчислення дивергенції Кульбака-Лайблера зі словником "функціональних слів" виявилось найбільш вдалим для праць групи D (P. Dorenbos). 7 праць із 10-ти праць групи D отримали ранг k <11. Як згадувалось, якість атрибуції q = 100 % буде відповідати випадку, коли всі 10 текстів розміщено у діапазоні k <11. Отже, якість атрибуції текстів P. Dorenbos, беручи за опорний текст словник "функціональних слів", становить q = 70 %. Для статей A. Meijerink лише 6 текстів групи M відсортовані в межах діапазону 1-10, тобто q = 60%. Низька ефективність розпізнавання текстів за "функціональними словами" виявлена для текстів G. Stryganyk (група S -10%) та G. Zimmerer (група Z -40%).

Ефективність атрибуції наукових текстів на основі порівняння зі словником "функціональних слів" художніх текстів виявилась низькою для G. Stryganyuk та G. Zimmerer. Високу ефективність атрибуції q = 70 % для праць P. Dorenbos можна пояснити тим, що значна частина його текстів була одноосібними. Дійсно, розпізнаними були його одноосібні статті D1, D2, D3, D4, D5, D6. Частково і для G. Zimmerer була розпізнана частина статей, які були одноосібними Z1, Z4, Z5. Отже, підхід до визначення дивергенції Кульбака-Лайблера на основі використання словника "функціональних слів" потребує подальшого вивчення з метою використання для атрибуції

одноосібних наукових статей. Звичайно, результати обчислень могли б бути дещо іншими у випадку вибору інших опорних текстів. Вибором опорного тексту найбільшого обсягу передбачалось охоплення максимальної кількості характерних ознак, за якими можна було б об'єднати тексти відповідної групи.

Розрахунок дивергенції Кульбака-Лайблера був проведений для *n*-грам слів навчальної вибірки текстів, яка містить одноосібні статті та статті у співавторстві. Розраховувалась міра розходження для текстів залежно від розміру послідовності вживання слів (1, 2, 3, 4) та вибору опорного тексту. У табл. 4.2 показано ефективність атрибуції текстів на прикладі статей P. Dorenbos у випадку вибору різних опорних текстів D1, ... , D10 для послідовності з 1, 2, 3 та 4 слів. Знайдені ефективності атрибуції значимо не відрізняються в межах інтервалу $q_{сер}\pm 2\sigma$. Однак, $q_{сер}$ є найбільшим для послідовності з 2 слів і досягає значення $q_{сер}$=80 %. Найкраща ефективність атрибуції для статей P. Dorenbos отримана за вибору D1 як опорного тексту. Так, для тексту D1 із двох слів ефективність атрибуції q=90 %. У подальшому цей текст D1 буде використано для визначення ефективності атрибуції у випадку апробації алгоритму на тестовій вибірці.

Таблиця 4.2

Ефективність атрибуції статей P. Dorenbosa на навчальній вибірці

послідовність	Стаття P. Dorenbosa										$q_{сер}\pm 2\sigma$
	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	
1 слова	90	80	80	70	80	80	60	60	50	60	**71**±24
2 слова	90	90	90	70	90	90	80	80	50	70	**80**±24
3 слова	80	70	80	70	70	70	70	50	50	50	**56**±22
4 слова	70	60	50	70	60	60	70	40	40	40	**66**±20

Аналогічні дослідження щодо пошуку оптимального опорного тексту та послідовності слів проведено і для статей, де є авторами або ж співавторами A. Meijerink, G. Stryganyuk, G. Zimmerer. У Додатку Ж (рис. 2) представлено приклади графічного розподілу текстів за мірою розбіжності Кульбака-Лайблера щодо оптимального опорного текста для послідовності з чотирьох слів. Такі розподіли були використані для визначення середньої

ефективності атрибуції для кожного з авторів. Ефективність розпізнавання текстів складає для P. Dorenbos – 80%, A. Meijerink – 70%, G. Stryganyuk – 100%, G. Zimmerer – 60%. Середні значення ефективності атрибуції статей для цих авторів представлені у табл. 4.3. Із даних табл. 4.3 видно, що оптимальні значення атрибуції досягнуто у випадку n=2 та n=3. У табл. 4.4 представлено результати максимальних ефективностей атрибуції текстів навчальної вибірки для різних авторів. Зауважимо, що максимальну ефективність атрибуції досягнуто для текстів, що мають найбільший обсяг словника серед текстів своєї групи. Саме ці тексти можна вибрати як опорні для дослідження ефективності атрибуції на тестовій виборці. Максимальна ефективність групування наукових текстів навчальної вибірки досягнута для послідовності з двох слів (табл. 4.4).

Таблиця 4.3

Середня ефективність атрибуції статей навчальної вибірки методом ентропії

послідовність	P.Dorenbos	A.Mejerink	G.Stryganyuk	Zimmerer	$q_{сер}\pm 2\sigma$
1 слова	71	79	63	52	66±20
2 слова	**80**	**85**	82	53	**75**±24
3 слова	66	82	**93**	59	**75**±30
4 слова	56	69	90	**60**	69±26

Таблиця 4.4

Максимальна ефективність атрибуції статей навчальної вибірки

послідовність	Dorenbos	Mejerink	Stryganyuk	Zimmerer	$q_{сер}\pm 2\sigma$
1 слова	90%	90%	70%	70%	80±20
2 слова	90%	100%	100%	80%	93±16
3 слова	80%	90%	100%	90%	90±14
4 слова	70%	90%	100%	90%	88±22

Порівнюючи ефективності групування текстів за словником “функціональних слів” та опорними текстами (табл. 4.5), можна відзначити таке. Досягнуте групування з ефективністю $q = 90\%$ для текстів групи D (праці P. Dorenbos) є на 20% вищим, ніж на основі словника “функціональних

слів" художніх текстів. На 40% покращилась атрибуція для текстів групи M і сягнула $q = 100\%$. Значна ефективність спостерігається для праць G. Stryganyuk (90%) та G. Zimmerer (50%). Такі результати вказують на недоцільність використання словника "функціональних слів" як опорних текстів для проведення авторської атрибуції англомовних наукових статей.

Таблиця 4.5

Максимальна ефективність атрибуції наукових текстів із використанням словника "функціональних слів" та за опорним текстом

q	P. Dorenbos	A. Meijerink	G. Stryganyuk	G. Zimmerer
опорний текст	90% n=1	100% n=2	100% n=2,3,4	90% n=3,4
словник "функці-ональних слів"	70%	60%	10%	40%

Якщо у попередніх випадках розглянуто розпізнавання одного автора серед 4 авторів, то в даному розділі пропонується встановлення одного автора серед 76 авторів. З цією метою проведено авторську атрибуцію за допомогою методу дивергенції Кульбака-Лайблера наукових статей Ю. Зоренка, представлених на VI міжнародній конференції LUMDETR – 2006 [251]. До вибірки із 96 статей 76 авторів входило 10 статей Ю. Зоренка (Додаток В). У розділі 3.1 статті Ю. Зоренка, використовуючи метод аналізу головних компонент, згрупувались в окрему область Z (рис. 4.2). Для цього масиву робіт проведена авторська атрибуція наукових текстів, використовуючи метод дивергенції Кульбака-Лайблера. Ефективність атрибуції є найменшою (60%) у випадку вибору як опорного тексту словника "функціональних слів" художніх текстів (Додаток Є). Ефективність атрибуції зростає, якщо за опорний текст брати один із текстів автора, аналізуючи його методом послідовності вживання слів. Ефективність атрибуції досягає максимуму для тексту A3 як опорного при послідовності вживання з чотирьох слів і становить 80%. Отже, як і у попередніх випадках авторської атрибуції наукових текстів максимальна ефективність атрибуції досягнута при послідовності більше двох слів, зокрема чотирьох. Ефективність

атрибуції зменшується при послідовності з п’яти слів і сягає 70% для послідовності з семи слів.

Тестування алгоритму групування текстів на основі дивергенції Кульбака-Лайблера та послідовності вживання слів проведено на тестовій вибірці, до якої входить 12 одноосібних статей P. Dorenbos (Додаток Е), які не входили до навчальної вибірки, та по 10 статей A. Meijerink, G. Strygynyuk і G. Zimmerer, які входили до навчальної вибірки. За опорний текст взято одноосібну статтю P. Dorenbos D1 (Додаток Е). Для цього тексту авторська атрибуція наукових текстів була оптимальною у випадку навчальної вибірки. Ефективність групування статей P. Dorenbos на тестовій вибірці для послідовності з: одного слова становить 85%, двох слів – 92%, трьох слів – 100%, чотирьох слів – 92%.

За результатами досліджень навчальної вибірки знайдено, що найкраще авторська атрибуція наукових текстів здійснюється для послідовності з 2 або 3 слів. Запропонований алгоритм (вибір оптимального опорного тексту, застосування міри порівняння Кульбака-Лайблера та послідовності вживання слів) дозволив отримати q=100% для статей P. Dorenbos у випадку послідовності з 3 слів для тестової вибірки. Отже, такий порядок послідовності вживання слів (3 слова) збігається із розміром послідовності вживання слів, отриманим на навчальній вибірці (послідовність 2 або 3 слів).

У наведеній нижче атрибуції текстів використано підрахунок статистики χ^2 для встановлення відстані між досліджуваними текстами, а не для визначення міри статистичної однорідності текстів: $\chi^2 = \sum \frac{(x_i - \bar{x}_i)^2}{\bar{x}_i}$,

де x_i – частота слова в кожному з текстів, $\bar{x}_i$ – середня частота слова у текстах. Такий підхід для порівняння текстів застосовували О. Шевельов [121], С. Ель-Харбі [131, p. 80]. Розрахунок χ^2 був проведений для послідовності вживання слів навчальної вибірки текстів, що і у випадку дивергенції Кульбака-Лайблера. Міра χ^2 розраховувалась для текстів залежно від розміру послідовності вживання слів (1, 2, 3, 4) та вибору опорного

тексту. У табл. 4.6 представлено ефективність атрибуції текстів на прикладі статей P. Dorenbos у випадку вибору різних опорних текстів D1, ... , D10 для n=1, 2, 3, 4. Для навчальної вибірки $q_{сер}$ є найбільшим для послідовності з 1 слова і досягає значення $q_{сер}$=80 %. Збільшення розміру n-грам слів призводить до погіршення ефективності групування статей. Такі ж тенденції щодо впливу розміру послідовності вживання слів на ефективність авторської атрибуції англомовних статей характерні і для атрибуції інших авторів. Для всіх авторів із використанням χ^2 середня ефективність атрибуції є максимальною $q_{сер}$=66 %. у випадку послідовності одного слова (табл. 4.7).

Ефективність групування статей за допомогою використання міри порівняння χ^2 перевірено на тестовій вибірці, аналогічно тій, що використовували для аналізу ефективності групування текстів на основі дивергенції Кульбака-Лайблера. Максимальна ефективність групування статей P. Dorenbos для тестової вибірки як і для навчальної вибірки досягнута для одного слова. Збільшення розміру параметра послідовності вживання слів спричинює погіршення ефективності атрибуції. Так, для послідовності з одного слова розпізнано ефективність групування текстів становить 54%, двох слів – 46%, трьох слів – 38, чотирьох слів – 31%.

Проведено зіставлення середньої ефективності атрибуції за використання дивергенції Кульбака-Лайблера та міри χ^2 для навчальної вибірки. У табл. 4.8 наведено результати групування текстів 4 авторів методом дивергенції Кульбака-Лайблера та міри $\boldsymbol{\chi^2}$ для навчальної вибірки. Можна відзначити, що ефективність розпізнавання текстів за використання послідовності вживання одного слова методом дивергенції Кульбака-Лайблера та мірою χ^2 має практично однакові результати $q_{сер}$=66 %. При збільшенні розміру послідовності вживання слова n≥2 ефективність атрибуції методом дивергенції Кульбака-Лайблера зростає. Для послідовності вживання з 3 слів досягаєттся ефективність у 75%, тоді як для χ^2 ці показники зменшуються, для послідовності вживання з 3 слів $q_{сер}$=36 %.

Таблиця 4.6

Ефективність групування статей P. Dorenbos методом χ^2 (навчальна вибірка)

послідовність	Стаття P. Dorenbos										$q_{сер}\pm 2\sigma$
	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	
1 слова	70	80	90	90	90	90	90	70	60	70	80±20
2 слів	70	70	100	80	80	70	90	90	50	40	73±30
3 слів	70	60	70	60	80	80	60	30	60	50	62±26
4 слів	50	60	50	50	60	70	50	30	30	70	57±28

Таблиця 4.7

Середня ефективність атрибуції статей методом χ^2 (навчальна вибірка)

послідовність	Dorenbos	Mejerink	Stryganyuk	Zimmerer	$q_{сер}\pm 2\sigma$
1 слова	73	62	74	55	66±18
2 слова	80	37	35	35	46±70
3 слова	62	43	14	26	36±30
4 слова	57	53	12	17	35±40

Таблиця 4.8

Ефективність атрибуції статей 4 авторів

послідовність	метод	навчальна вибірка		тестова вибірка
		q	2σ	q
1 слово	дивергенція К-Л	66	±15	85
	міра χ^2	*66*	*±18*	***54***
2 слова	дивергенція К-Л	75	±16	92
	міра χ^2	*46*	*±40*	*46*
3 слова	дивергенція К-Л	75	±14	**100**
	міра χ^2	*36*	*±30*	*38*
4 слова	дивергенція К-Л	69	±18	92
	міра χ^2	*35*	*±40*	*31*

Найкращі параметри розпізнавання для дивергенції Кульбака-Лайблера (послідовність вживання із трьох слів, $q_{сер}$=75 %) та міри χ^2 (послідовність вживання одного слова, $q_{сер}$=66 %) в межах 2σ інтерквантильного інтервалу не відрізняються. Однак слід мати на увазі, що середнє значення ефективності розпізнавання є більшим для дивергенції Кульбака-Лайблера. Ще кращих результатів отримано для тестової вибірки праць P. Dorenbos. Тут

q=100% для послідовності із трьох слів для методу дивергенції Кульбака-Лайблера та q=54% для міри χ^2. Однак відсутність значень інтерквантильних інтервалів через недостатню кількість текстів інших авторів не дозволяє зробити висновок про значимість цих результатів.

Класифікуючи роботи G. Stryganyuk (параграф 4.3), було встановлено, що найкраще розпізнавання текстів методом аналізу головних компонент відбувається при послідовності вживання із чотирьох слів. Ця тенденція – високий порядок послідовності вживання слів для ефективної авторської атрибуції статей – збереглася і у випадку атрибуції наукових статей за мірою дивергенції Кульбака-Лайблера (послідовність з трьох слів є оптимальною для авторської атрибуції наукових текстів для цього підходу). Найкращий результат авторської атрибуції вибірки текстів чотирьох авторів у випадку методу одночасного моніторингу групування текстів та відповідних їм слів також досягнуто для послідовності вживання із чотирьох слів. Слід зауважити, що обсяг словника сягає свого максимального значення (107349) саме у випадку послідовності вживання чотирьох слів, для якого й було досягнуто найкращого групування досліджуваних текстів. Те ж саме значення оптимального розміру послідовності вживання слів було виявлене для авторської атрибуції вузькоспеціалізованих наукових праць Г. Стриганюка (параграф 4.3.) методом аналізу головних компонент, де також спостерігається та сама кореляція між розміром послідовності вживання слів та обсягом словника послідовності вживання слів на користь послідовності з чотирьох слів [24]. У Додатку Г (рис. 9) представлено результат атрибуції статей 4 авторів (P. Dorenbos, A. Meijerink, G. Strganyuk, G. Zimmerer) методом одночасного моніторингу групування текстів та відповідних їм слів для послідовності вживання із чотирьох слів. Розподіл статей представлено в системі координат третьої, четвертої та п’ятої головних компонент, які описують 9,45% дисперсії вхідних даних. Цього виявилося достатньо для розподілу досліджуваних робіт на відповідні їм групи D, M, S та Z. Групи S, D, а також групи M та Z розділяються уздовж осі

PC-3. Групи M та Z розділяються між собою за характеристиками, що формують п’яту головну компоненту. Ці групи (M та Z) розведені одна від одної уздовж осі п’ятої головної компоненти. Роботи групи D розподілені вздовж осі четвертої головної компоненти. Як і у випадку дивергенції Кульбака-Лайблера метод одночасного моніторингу групування текстів та їх груп дозволив чітко окреслити роботи групи S та виявити перекриття груп D, M та Z, роботи яких виявляли спорідненість і у випадку застосування методу ентропії (дивергенції Кульбака-Лайблера).

Просторове окреслення області розташування робіт авторів зроблено, ґрунтуючись на відомій приналежності статей до того чи іншого автора. Без такої попередньої інформації дані області було б важко виділити. Для окреслених областей визначено ефективність атрибуції за кількістю статей автора, що потрапляють в окреслену область. Досягнуто ефективність групування статей для P. Dorenbos – 90%, A. Meijerink – 80%, G. Strganyuk – 100%, G. Zimmerer – 90%. Середня ефективність атрибуції становить 90%.

Області, виділені в просторі головних компонент, дозволяють у подальшому використовувати їх для розпізнавання невідомих текстів, які вводяться в базу даних. Ця картина може розглядатись як навчальна мапа для атрибуції статей із сумнівним або невідомим автором. Щоб перевірити це, до масиву текстів Г. Стриганюка додано ще одну статтю S10 (на рис. 4.7 виділена червоним кільцем), яка не входила до навчальної вибірки. Після групування текстів методом аналізу головних компонент стаття S10 потрапила в область статей Г. Стриганюка, виділену для навчальної вибірки.

Для статей P. Dorenbos, A. Meijerink, G. Strganyuk та G. Zimmerer (Додаток Е) укладено словник перших 70 найчастіше вживаних послідовностей чотирьох слів (табл. 4.9 – 4.12). Аналіз словника послідовностей 4 слів для кожного з авторів показав, що кількість спільних найчастіше вживаних послідовностей чотирьох слів між авторами є мінімальною. Виявлено лише одну спільну послідовність чотирьох слів *the excitation spectrum of* серед перших 70 найчастіше вживаних послідовностей.

P. Dorenbos, A. Meijerink, G. Strganyuk та G. Zimmerer вживають послідовність *the excitation spectrum of* в однаковому контексті у своїх текстах. Ця послідовність потребує дистрибуції з назвами хімічних сполук. Хоча ці автори є представниками галузі люмінесцентної спектроскопії та працюють з однаковими методами й використовують однакове експериментальне обладнання, об'єкти їх дослідження та фізичні процеси є різними. Так, до прикладу, автори вживають виділену послідовність чотирьох слів у наукових текстах: *In the case of $Lu_2Si_2O_7$ the interpretation of the 278-nm band in the excitation spectrum of Ce^{3+} luminescence is not fully certain* (P. Dorenbos); *The agreement with the intensity ratio in the excitation spectrum of $LiYF_4:Ce^3$ reported in Ref. 40 is better* (A. Meijerink); *The excitation spectrum of LiYP4O12 intrinsic emission (figure 2, curve 2) has a threshold around 8.4 eV and a maximum at 8.55 eV corresponding to the excitonic absorption* (G. Strganyuk); *The most dominant peak in the excitation spectrum of the spectrally selected Ne_3P1 resonance fluorescence (curve (2)) corresponds to the energetic position of the surface exciton marked with "s" in absorption (curve (1))* (G. Zimmerer).

Для статей P. Dorenbos, G. Stryhanyuka і G. Zimmerer спільною виявилась послідовність чотирьох слів *in the case of*, яка побудована за моделлю preposition + article + noun + preposition. Послідовності *the energy of the* та *as a function of* є спільними лише для текстів P. Dorenbos та G. Zimmerer, а *the position of the* та *the crystal field splitting* – P. Dorenbos A. Meijerink. Послідовності *the energy difference between* та *energy difference between the* є не тільки спільними для текстів P. Dorenbos і A. Meijerink, а й вживаються авторами з подібною частотою. Наприклад, *Because of the anti-correlation, the energy difference between the first 4fn-15d state and the bottom of the conduction band is relatively invariant with type of lanthanide ion* (P. Dorenbos) та *The energy difference between the levels in CaF2 can be determined exactly, because fine structure is observed* (A. Meijerink).

Варто зазначити, що і для одного автора ймовірність появи однакової послідовності чотирьох слів у декількох текстах дуже мала. Так, P. Dorenbos лише одну послідовність *the energy of the* вживає у всіх десяти текстах, а от друга та третя за порядковим номером найчастіше вживані послідовності *of the 5d levels* та *the energy difference between* вживаються лише у вісьмох з десяти текстів (табл. 4.9). Чим більший порядковий номер послідовності, тим менша ймовірність її появи у великій кількості аналізованих текстів. Послідовність *the position of the* (вісімнадцятий порядковий номер) є у п'яти з десяти текстів P. Dorenbos. Виявлено, що кожен із досліджуваних авторів має лише одну послідовність чотирьох слів, яка б уживалась у всіх десятьох текстах одного автора. A. Meijerink має найбільшу кількість можливих повторюваних послідовностей у текстах, наприклад, чотири послідовності (*rare earth ions in, the energy difference between, energy difference between the, the excitation spectrum of*) зустрічаються у дев'яти текстах, чотири послідовності (*the crystal field splitting, of rare earth ions, excitation spectrum of the, in the excitation spectrum*) – восьми текстах, шість послідовностей (*the fact that the, were corrected for the, the positions of the, is due to the, at the desy synchrotron, levels of rare earth*) – семи текстах.

Таблиця 4.9.

Послідовності вживання з 4 слів, характерні для P. Dorenbos

п\н	**Послідовність**	к-ть текстів	**п\н**	**Послідовність**	к-ть текстів
1	the energy of the	*10*	37	is related to the	*5*
2	of the 5d levels	*8*	38	crystal field splitting and	*5*
3	the energy difference between	*8*	39	the trivalent lanthanides in	*5*
4	energy differences between the	*7*	40	can be written as	*5*
5	the excitation spectrum of	*7*	41	of the crystal field	*5*
6	5d levels of ce3	*6*	42	energy differences between the	*5*
7	in the case of	*6*	43	the centroid shift of	*5*
8	the lowest 5d level	*6*	44	centroid shift of the	*5*
			45	the end of the	*5*

9	of the lanthanide ion	*6*	46	that the crystal field	*5*
10	with the type of	*6*	47	trivalent lanthanides in	
11	the size of the	*6*		inorganic	*4*
12	is attributed to the	*6*	48	5d level to the	*4*
13	of the trivalent lanthanides	*6*	49	it was shown that	*4*
14	the shape of the	*6*	50	the same for all	*4*
15	are compiled in table	*6*	51	of the 5d electron	*4*
16	the excitation spectra of	*5*	52	allowed fd transition in	*4*
17	are shown in fig	*5*	53	level positions of the	*4*
18	the position of the	*5*	54	it was found that	*4*
19	of the lowest 5d	*5*	55	the emission spectrum of	*4*
20	on the type of	*5*	56	transitions from the lowest	*4*
21	can be found in	*5*	57	be found in the	*4*
22	the same holds for	*5*	58	information is available on	*4*
23	lowest 5d level of	*5*	59	in the host crystal	*4*
24	the values for the	*5*	60	the presence of a	*4*
25	it is assumed that	*5*	61	can be seen as	*4*
26	be seen in fig	*5*	62	the type of cations	*4*
27	and the type of	*5*	63	with the crystal field	*4*
28	the location of the	*5*	64	of the host crystal	*4*
29	by means of a	*5*	65	levels relative to the	*4*
30	as a function of	*5*	66	emission spectra were	
31	on the energy of	*5*		recorded	*4*
32	of the 5d configuration	*5*	67	excitation spectra were	
33	crystal field splitting of	*5*		corrected	*4*
34	field splitting of the	*5*	68	spectra were corrected for	*4*
35	the crystal field splitting	*5*	69	nm is attributed to	*4*
36	difference between the		70	the bottom of the	*4*
	lowest	*5*			

Таблиця 4.10

Послідовності вживання з 4 слів, характерні для A. Meijerink

п\н	**Послідовність**	к-ть текстів	**п\н**	**Послідовність**	к-ть текстів
1	the intensity of the	*10*	36	to the ground state	*5*
2	rare earth ions in	*9*	37	the crystal growth melt	*5*
3	the energy difference		38	excitation spectra were	
	between	*9*		corrected	*5*
4	energy difference between the	*9*	39	spectra were corrected for	*5*
5	the excitation spectrum of	*9*	40	emission spectra in the	*5*
6	the crystal field splitting	*8*	41	measurements were	
7	of rare earth ions	*8*		performed at	*5*

8	excitation spectrum of the	*8*	42	figure 3 shows the	*5*
9	in the excitation spectrum	*8*	43	in agreement with the	*5*
10	the fact that the	*7*	44	the transition to the	*5*
11	were corrected for the	*7*	45	zero phonon line and	*5*
12	the positions of the	*7*	46	to the 4f levels	*5*
13	is due to the	*7*	47	the crystal field parameters	*5*
14	at the desy synchrotron	*7*	48	the assignment of the	*5*
15	levels of rare earth	*7*	49	energy transfer from the	*5*
16	vacuum ultraviolet		50	is shown in fig	*5*
	spectral region	*6*	51	relative intensities of the	*5*
17	the zero phonon line	*6*	52	an overview of the	*5*
18	the position of the	*6*	53	the opportunity to use	*5*
19	it is not possible	*6*	54	opportunity to use the	*5*
20	is not possible to	*6*	55	and found to be	*5*
21	by the crystal field	*6*	56	upon excitation in the	*5*
22	the authors are grateful	*6*	57	the spectrum in fig	*5*
23	authors are grateful to	*6*	58	are assigned to transitions	*5*
24	from hasylab for the	*6*	59	the emission spectra of	*5*
25	hasylab for the opportunity	*6*	60	be explained by the	*5*
26	for the opportunity to	*6*	61	was supported by the	*5*
27	spectra and energy levels	*6*	62	the observation of the	*5*
28	and energy levels of	*6*	63	on the basis of	*5*
29	energy levels of rare	*6*	64	a single crystal of	*5*
30	earth ions in crystals	*6*	65	the splitting of the	*4*
31	can be explained by	*6*	66	to transitions from the	*4*
32	good agreement with the	*6*	67	of the 4f levels	*4*
33	in the uv and	*5*	68	fine structure in the	*4*
34	of the high energy	*5*	69	of the zero phonon	*4*
35	the energies of the	*5*	70	of the energy levels	*4*

Таблиця 4.11

Послідовності вживання з 4 слів, характерні для G. Stryganyuk

п\н	**Послідовність**	к-ть текстів	**п\н**	**Послідовність**	к-ть текстів
1	in the range of	*10*	34	princeton instruments ccd	
2	in the case of	*9*		detector	*5*
3	of the decay time	*8*	35	at the secondary arc 5	*5*
4	the decay time constant	*8*	36	the secondary arc	*5*
5	the excitation spectrum of	*7*		monochromator	*5*
6	the decay kinetics of	*7*	37	200 ns time gate	*5*

7	for the case of	*7*	38	the nonradiative decay of	*5*
8	luminescence excitation and emission	*7*	39	upon excitation in the	*5*
9	excitation and emission spectra	*7*	40	the energy range of	*5*
10	spectra as well as	*7*	41	has been revealed for	*5*
11	luminescence excitation spectra were	*6*	42	upon the excitation within	*5*
12	excitation spectra were scanned	*6*	43	using the facility of	*5*
13	spectra were scanned with	*6*	44	after the excitation pulse	*5*
14	at t 8 k	*6*	45	measurements of luminescence excitation	*5*
15	in the excitation spectra	*6*	46	of luminescence excitation and	*4*
16	be caused by the	*6*	47	for the excitation of	*4*
17	the excitation spectra of	*6*	48	30 cm monochromator spectrograph	*4*
18	and emission spectra as	*6*	49	photomultiplier at the secondary	*4*
19	emission spectra as well	*6*	50	ns time gate defined	*4*
20	decay kinetics were performed	*6*	51	time gate defined by	*4*
21	the emission spectrum of	*6*	52	defined by the excitation	*4*
22	in the emission spectrum	*6*	53	by the excitation pulse	*4*
23	excitation in the range	*6*	54	excitation pulse repetition upon	*4*
24	by the primary 2	*6*	55	storage ring operation in	*4*
25	the primary 2 m	*6*	56	upon the excitation of	*4*
26	primary 2 m monochromator	*6*	57	the emission spectra of	*4*
27	2 m monochromator in	*6*	58	spectral kinetic characteristics of	*4*
28	the measurements were carried	*5*	59	be explained by the	*4*
29	measurements were carried out	*5*	60	decay time constant for	*4*
30	emission spectra were measured	*5*	61	the near activator exciton	*4*
31	spectra were measured with	*5*	62	the energy transfer from	*4*
32	czerny turner mounting equipped	*5*	63	excitation quanta in the	*4*
33	turner mounting equipped with	*5*	64	in the formation of	*4*
			65	decay time constant of	*4*
			66	been revealed for the	*4*
			67	the temperature dependence of	*4*
			68	temperature dependence of the	*4*
			69	at t 300 k	*4*
			70	the decay kinetics for	*4*

Таблиця 4.12

Послідовності вживання з 4 слів, характерні для G. Zimmerer

п\н	Послідовність	к-ть текстів	п\н	Послідовність	к-ть текстів
1	at the superlumi station	*7*	36	spectroscopy of localized	
2	in the case of	*7*		atomic	*3*
3	on the other hand	*6*	37	the luminescence spectra of	*3*
4	in the present paper	*5*	38	the fact that the	*3*
5	the energy of the	*5*	39	the shape of the	*3*
6	in the range of	*5*	40	vacuum ultraviolet	
7	as an excitation source	*5*		radiation physics	*3*
8	excitation spectrum of the	*5*	41	photon excitation in the	*3*
9	in the vuv spectral	*5*	42	a key role in	*3*
10	the vuv spectral range	*5*	43	has to be mentioned	*3*
11	the excitation spectrum of	*5*	44	due to the fact	*3*
12	energy transfer from the	*4*	45	to the fact that	*3*
13	were performed at the	*4*	46	point of view of	*3*
14	performed at the superlumi	*4*	47	the spectral resolution of	*3*
15	the superlumi station of	*4*	48	the excitation spectra of	*3*
16	superlumi station of hasylab	*4*	49	excitation spectra of different	*3*
17	of hasylab at desy	*4*	50	excitation spectra of the	*3*
18	emission and excitation		51	for the first time	*3*
	spectra	*4*	52	energy levels of the	*3*
19	is of the orde	*4*	53	are shown in fig	*3*
20	of the order of	*4*	54	given in the figure	*3*
21	as a function of	*4*	55	would like to point	*3*
22	is due to the	*4*	56	like to point out	*3*
23	of the exciting radiation	*4*	57	with respect to the	*3*
24	is shown in fig	*4*	58	it is not the	*3*
25	in the excitation spectrum	*4*	59	is not the purpose	*3*
26	excitation spectra in the	*4*	60	not the purpose of	*3*
27	in the vacuum ultraviolet	*4*	61	to the excitation pulse	*3*
28	of the energy transfer	*3*	62	50 mrad of sr	*3*
29	the energy transfer from	*3*	63	mrad of sr from	*3*
30	the measurements were		64	of sr from a	*3*
	performed	*3*	65	sr from a bending	*3*
31	measurements were		66	from a bending magnet	*3*
	performed at	*3*	67	luminescence of solid xe	*3*
32	station of hasylab at	*3*	68	luminescence excitation spectra in	*3*
33	the energy range of	*3*	69	in the vicinity of	*3*
34	in the region of	*3*	70	in the near future	*3*
35	relaxation from higher lying	*3*			

Найчастіше вживаною послідовністю чотирьох слів у текстах P. Dorenbos є *the energy of the*, яка побудована за моделлю **article + noun + preposition + article**. Наприклад, *The energy of the 5d excited states of the trivalent lanthanides and their location relative to those of the 4 fn configuration is important for the luminescence properties of lanthanide activated phosphors. The energy of the highest 5d level, the centroid position, the energy of the lowest 5d level, and the energy of emission from the relaxed lowest energy 5d level to the 2F5/2 ground state are shown in Fig. 3.* З-поміж перерахованих 70 послідовностей чотирьох слів (табл. 4.9), які вживаються у текстах P. Dorenbos, можна виділити такі часто вживані моделі:

article + noun + preposition + article (*the energy of the, the size of the, the shape of the, the location of the, the position of the, the values for the, the location of the, the end of the, the presence of a the bottom of the, the top of the*);

preposition + article + noun + preposition (*in the case of, with the type of, on the type of, on the energy of*);

article + noun + noun + preposition (*the energy difference between, the excitation spectrum of, the excitation spectra of, the emmision spectrum of*);

pronoun + auxiliary verb + verb + conjunction (*it is assumed that, it was shown that, it was found that, it is known that*);

auxiliary verb + verb + preposition + article (*is attributed to the, is related to the, be found in the*);

auxiliary verb + verb + preposition + noun (*are shown in fig, are compiled in table, be seen in fig*);

preposition + article + adjective + noun (*of the 5d level, of the 5d configuration, of the 5d electron*).

Також у досліджуваного автора зустрічаються моделі типу: adjective + noun + preposition + noun (*5d levels of cerium*); article + noun + noun + noun (*the crystal field splitting*); modal verb + auxiliary verb + verb + preposition (*can be found in*); noun + noun + preposition + article (*energy difference between the*);

preopsition + article + noun + noun (*of the lanthanide ion*); preposition + noun + preposition + article *(by means of a*).

Послідовність чотирьох слів *the intensity of the* є найчастіше вживаною у текстах A. Meijerink (табл. 4.10). Вона утворена за моделлю **article + noun + preposition + article**, що переважає серед інших у словнику найчастіше вживаних послідовностей чотирьох слів A. Meijerink (*the position of the, the transition of the, the assignment of the*). Для текстів A. Meijerink припадає найбільша кількість різних комбінацій моделей щодо перших 30 найчастіше вживаних послідовностей. Так, автор часто використовує послідовності чотирьох слів за моделями:

article + noun + noun + preposition (*the excitation spectrum of, the energy difference between*);

noun + noun + preposition + article (*excitation spectrum of the, energy transfer from the*);

preposition + article + noun + preposition (*on the basis of, for the opportunity to*).

У текстах A. Meijerink серед перших 70 найчастіше вживаних послідовностей чотирьох слів наявні також моделі: adjective + noun + preposition + article (*good agreement with the*); article + adjective + noun + noun (*the zero phonon line*); article + noun + conjunction + article (*the fact that the*); article + noun + noun + noun (*the crystal field splitting*); article + noun + preposition + noun (*the spectrum in fig*); auxiliary verb + verb + preposition + article (*were corrected for the*); modal verb + auxiliary verb + verb + preposition (*can be explained by*); noun + preposition + adjective + noun (*levels of rare earth*); noun + noun + noun + noun (*vacuum ultraviolet spectral region*); noun + noun + preposition adjective (*energy levels of rare*); noun + noun + preposition noun (*earth ions in crystal*); preposition + article + adjective + noun (*of the rare earth*); preposition + article + noun + noun (*at the desy synchrotron*); preposition + noun + preposition + article (*upon excitation in the*).

Серед розглянутих послідовностей чотирьох слів можна виділити такі послідовності, об'єднавши які можна відтворити частину речення, адже існують певні набори (групи) послідовностей чотирьох слів, що вживаються в текстах з однаковою частотою. Здебільшого, це відбувається, коли автор копіює речення з одного тексту в інший та частково його змінює. Наприклад, такі послідовності, як *the authors are grateful, authors are grateful to, from hasylab for the, hasylab for the opportunity, for the opportunity to, the opportunity to use, opportunity to use the* зустрілися в 5 текстах A. Meijerink з однаковою частотою та в тому ж контексті. Вони утворили речення, наприклад, *The authors are grateful to Dr. P. Gürtler and Dr. S. Petersen from HASYLAB for the opportunity to use the excellent facilities for VUV spectroscopy at the DESY synchrotron* (стаття M1, Додаток Е). А от послідовності *spectra and energy levels, and energy levels of, energy levels of rare, earth ions in crastals* є яскравим прикладом розмежування речення на послідовності чотирьох слів, що повністю відтворюють назву праці, на яку автор посилався у п'яти із десяти своїх статей G. H. Dieke, Spectra and Energy Levels of Rare Earth Ions in Crystals ~Interscience, New York, 1968.

G. Stryganyuk найчастіше використовує послідовність *in the range of* (табл. 4.11), яка є у всіх десяти текстах автора. Серед перших 70 найчастіше вживаних послідовностей чотирьох слів можна виділити такі моделі:

preposition + article + noun + preposition (*in the range of, in the case of, for the case of, upon the excitation within, for the excitation of, upon the excitation of, in the formation of, for the formation of*);

article + noun + noun + preposition (*the excitation spectrum of, the excitation spectra of, the energy range of, the emission spectra of, the enarga transfer from, the temperature dependence of*);

noun + noun + auxiliary verb + verb (*excitation spectra were scaned, emission spectra were measured*);

noun + auxiliary verb + verb + preposition (*spectra were scaned with, spectra were measured with*);

preposition + article + noun + noun (*in the excitation spectra, in the emission spectrum, after the excitation pulse, by the excitation pulse*);

auxiliary verb + verb + preposition + article (*be caused by the, be explained by the, been revealed for the*).

Наведемо приклади й інших моделей, які використовуються у текстах досліджуваного автора: noun + conjunction + noun + noun (*excitation and emission spectra*), conjunction + noun + noun + conjunction (*and emission spectra as*), noun + preposition + article + noun (*excitation in the range*), article + adjective + noun + preposition (*the nonradiative decay of*), preposition +noun + noun +conjunction (*of luminescence excitation and*), adjective + noun + noun + preposition (*spectral kinetic characteristics of*), verb + preposition + article + noun (*defined by the excitation*), article + noun + preposition + noun (*the facility of superlumi*).

У текстах G. Stryganyuk, так само як і у текстах A. Meijerink, вдалося виділити набір послідовностей чотирьох слів (*luminescence excitation and emission, excitation and emission spectra, and emission spectra as, emission spectra as well, spectra as well as*), який наявний у 5 текстах в однаковому контексті та з однаковою частотою (табл. 4.11). Наприклад, *Measurements of luminescence excitation and emission spectra as well as luminescence decay kinetics were performed at the Deutsches Elektronen Synchotron (DESY, Hamburg)* (S2, Додаток Е). *Measurements of luminescence excitation and emission spectra as well as decay kinetics were performed using the facility of the SUPERLUMI station [10] at HASYLAB* (S3, Додаток Е). *Measurements of luminescence excitation and emission spectra as well as luminescence decay kinetics were performed at Deutsches Elektronen Synchotron (DESY, Hamburg) using the synchrotron radiation from DORIS III storage ring and facility of SUPERLUMI experiment at HASYLAB [15]* (S5, Додаток Е). *Measurements of the luminescence excitation and emission spectra upon the excitation of KCl-LaCl3 (0.5 mol.%)-PrCl3 (0.05 mol.%) with the synchrotron radiation from DORIS III storage ring were performed at HASYLAB (DESY, Hamburg) using the facility of*

SUPERLUMI experiment [13] (S6, Додаток Е). *Time-resolved measurements of luminescence excitation and emission spectra as well as the luminescence decay kinetics were performed upon the excitation by the synchrotron radiation from DORIS III storage ring (DESY, Hamburg) using the facility of SUPERLUMI station at HASYLAB [6]* (S7, Додаток Е). В усіх наведених прикладах перша частина речення "*measurements of luminescence excitation and emission spectra as well as decay kinetics were performed"* повністю збігається.

Послідовність *at the superlumi station* є найчастіше вживаною у текстах G. Zimmerer (табл. 4.12). Схема **preposition + article + noun + noun** є домінантною серед перших 70 найчастіше вживаних послідовностей чотирьох слів. У текстах G. Zimmerer часто зустрічаються послідовності, що утворені за моделями:

preposition + article + noun + preposition (*in the case of, in the range of, of the order of, in the region of, in the vicinity of*);

article + noun + noun + preposition (*the excitation spectrum of, the superlumi station of, the energy transfer from, the excitation spectra of*);

preposition + article + noun + noun (*in the excitation spectrum, in the vacuum ultraviolet, of the energy transfer, to the excitation pulse*);

verb + preposition + article + noun (*performed at the superlumi station, is of the order, given in the fig*).

Автор використовує також моделі: preposition+ article + adjective + noun (*in the present paper*), article + noun + preposition + article (*the energy of the*), conjunction + article + noun + noun (*as an excitation source*), noun + noun + preposition + article (*excitation spectrum of the*), auxiliary verb + verb + preposition article (*were performed at the*), noun + noun + preposition + noun (*superlumi station of hasylab*), preposition + noun + preposition + article (*of hasylab at desy*), article + noun + auxiliary verb + verb (*the measurements were performed*), noun + auxiliary verb + verb + preposition (*measurements were performed at*), preposition + noun + preposition + article (*with respect to the*), preposition + noun +preposition + noun (*of hasylab at desy*).

За допомогою групи послідовностей *were performed at the, performed at the superlumi, the superlumi station of, superlumi station of hasylab, of hasylab at desy*, які автор уживає у чотирьох текстах з подібною частотою, можна простежити, як формується речення: *The measurements were performed at the SUPERLUMI station of HASYLAB at DESY under the excitation by synchrotron radiation from the DORIS storage ring*. (Z1, Додаток З).

У G. Zimmerer перша найчастіше вживана послідовність чотирьох слів *at the superlumi station* зустрічається лише у семи текстах з десяти, друга та третя найчастіше вживані послідовності *in the case of, on the other hand* – у шести текстах з десяти, наступні вісім послідовностей *in the present paper, the energy of the, in the range of, as an excitation source, excitation spectrum of the, in the vuv spectral, the vuv spectral range, the excitation spectrum of* – у п'ятьох текстах з десяти (табл. 4.12). У автора відсутня послідовність чотирьох слів, яка б зустрічалась у всіх десяти текстах.

З-поміж 70 найчастіше вживаних послідовностей чотирьох слів P. Dorenbos, A. Meijerink, G. Strganyuk та G. Zimmerer можна виділити такі, що виражають припущення (*it is assumed, can be explained by, can be found in*); візуально представляють результати дослідження (*be seen in fig, are shown in fig, are compiled in table, the spectrum in fig, figure 3 shows the*), згадують місце проведення експерименту, назви приладів (*at the desy synchrotron, measurements were carried out, the secondary arc monochromator, princeton instruments ccd detectror, at the superlumi station, of hasylab at desy*). Ці послідовності чотирьох слів відображають такі складові частини наукової статті, як "Опис експерименту" та "Результати", де автор найбільше може розкрити свій авторський стиль. Варто зазначити, що у жодній із досліджених найчастіше вживаних послідовностей чотирьох слів не зустрічаються назви хімічних сполук, що відкидає можливість групування текстів за спільною тематикою.

Висновки до розділу 4

Перспективною виявилася спроба атрибуції англомовних наукових текстів із залученням таких сучасних методів аналізу, як метод одночасного групування текстів та відповідних їм слів (у математиці відомий як метод головних компонент) і метод ентропії. Для зіставлення ефективності атрибуції із залученням усталених та сучасних підходів до аналізу текстів використано метод хі-квадрат. Ці методи мають спільний лінгвістичний параметр аналізу – статистику послідовності вживання слів (n-грам).

Особливість застосування методу одночасного групування текстів та відповідних їм слів полягає в тому, що він не потребує попереднього опрацювання тексту, тобто вибору характерних параметрів атрибуції. Метод одночасного моніторингу групування текстів та відповідних їм слів, крім здійснення процедури атрибуції, дозволяє виділити групи слів, характерні для текстів певного автора або відповідної тематики, а також забезпечує графічну візуалізацію групування текстів у просторі головних компонент. Кожна головна компонента – певна характеристика текстів (тематика, автор), яка може об’єднувати сотні словоформ. Метод ентропії має інші переваги, простіший алгоритм процедури атрибуції та менші витрати машинного часу.

Тематична атрибуція наукових текстів здійснена на прикладі аналізу наукових праць VI-ої Міжнародної конференції LUMDETR 2006. Знаючи теми наукових праць та проаналізувавши просторовий розподіл текстів у просторі головних компонент, вдалося виділити області, які об’єднують групи текстів певної тематики конференційної секції. Для тематичної атрибуції наукових текстів ефективним виявився аналіз послідовності з одного слова, збільшення розміру послідовності слів не давало покращення. Виділено 5 тематичних секцій конференції із 9, визначених організаторами конференції. У просторі головних компонент чітко згруповано три групи текстів (“Дозиметричні матеріали”, “Запасаючі та інші фосфори”, “Домішки, дефекти, пастки”). Такий розподіл можна пояснити новизною напрямків

досліджень. Праці п’яти тематичних секцій конференції згруповано в одну тематичну секцію в області початку координат. Таке групування може бути зумовлене тим, що для наукових праць цих секцій характерні усталені та широко розповсюджені множини словоформ даної галузі знань. Важливо, що поряд із розподілом текстів метод одночасного моніторингу групування текстів задає розподіл слів у просторі головних компонент. Це дозволяє накласти область текстів на область словоформ і виділити слова, що найбільше характеризують виділені області. Користуючись таким накладанням, виявлено область праць Z, яка не відповідає тематикам секцій, однак тексти у ній об’єднані спільними об’єктами дослідження – кристалами, плівками зі структурою гранату. Виділення області Z демонструє можливості даного методу виявляти приховані ознаки текстів.

Якщо відсутній доступ до повних текстів наукових статей, то інформацію про тексти можна отримати, аналізуючи менші за розміром від статей їх заголовки, анотації або тези. Підставою для цього є близьке просторове групування елементів пар “теза-стаття” та тріад “заголовок-анотація-стаття”. Заголовки та анотації статей розглянуто як елементи вибірки з високим умістом ключових слів, які є визначальними для тематичної атрибуції. Просторове розташування елементів пар “теза-стаття” та тріад “заголовок-анотація-стаття” можна використати для оцінки відмінностей між опублікованою статтею та заявлених початково тез.

Авторська атрибуція наукових текстів апробована для двох масивів текстів, де слід було ідентифікувати: а) одного автора (Г. Стриганюка) серед багатьох авторів, б) чотирьох авторів (P. Dorenbos, A. Meijerink, G. Strganyuk, G. Zimmerer) серед багатьох авторів. Здійснивши аналіз впливу зміни розміру параметра послідовності вживання слова (від одного до десяти слів), виявлено, що лінгвістичний параметр послідовності вживання із чотирьох слів є характерним для авторської атрибуції англомовних наукових текстів, якщо використовувати метод одночасного моніторингу групування текстів та відповідних їм слів. Оптимальне групування текстів відповідає максимальній

кількості знайдених послідовностей вживання слів у тексті. Атрибуція наукових текстів Г. Стриганюка на основі послідовності вживання з одного слова відбувається швидше тематична, ніж авторська. Покращити авторську атрибуцію можна за умови вилучення словоформ, пов'язаних з об'єктами та методами дослідження. Найкраще прояв стилю Г. Стриганюка виявлено у використанні послідовності вживання з чотирьох слів, характерних для: опису спостережуваних об'єктів, вираження припущень, представлення результатів. Роботи автора розмежовано на такі, що виконані у співавторстві, або такі, де остаточне редагування належить автору. Також і для одночасної авторської атрибуції текстів чотирьох авторів серед багатьох авторів отримано високу ефективність розпізнавання текстів для методу одночасного моніторингу групування текстів та відповідних їм слів у поєднанні саме із послідовністю вживання із чотирьох слів. Для навчальної вибірки статті компактно групуються у 4 області відповідно до автора статей. Окреслені області в просторі головних компонент для навчальної вибірки можна використовувати для розпізнавання невідомих текстів тестової вибірки. Для цього до масивів відомих текстів додають невідомі тексти і до нової сукупності текстів застосовують аналіз головних компонент. Тоді потрапляння невідомих текстів у відому область дозволить провести їх ідентифікацію. Так, до масиву відомих текстів було додано тестову статтю S10, яка правильно потрапила в область групування статей Г. Стриганюка.

Укладено словник найчастіше вживаних послідовностей чотирьох слів для текстів P. Dorenbos, A. Meijerink, G. Strganyuk, G. Zimmerer. Для наукових англомовних текстів спостерігається загальна тенденція щодо повторюваності вживання послідовності чотирьох слів у різних текстах одного автора: ймовірність появи однієї й тієї ж послідовності чотирьох слів у всіх текстах одного автора є дуже малою. Проте і тут між авторами можна простежити різну тенденцію до повторюваності одних і тих же послідовностей у своїх текстах. Найменша кількість послідовностей чотирьох слів, що повторюються у декількох текстах, є у G. Zimmerer, а

найбільша – A. Meijerink. Так, у текстах G. Zimmerer жодна послідовність чотирьох слів не зустрілась у всіх десяти текстах автора, перша найчастіше вживана послідовність наявна лише у семи текстах з десяти. A. Meijerink, натомість, має декілька послідовностей, які він використовує у кожному з текстів. У текстах G. Strganyuk спостерігається найменша кількість спільних послідовностей чотирьох слів з іншими авторами, у текстах P. Dorenbos – найбільше.

Запропоновано поєднати метод ентропії, зокрема її різновид – дивергенцію Кульбака-Лайблера та параметр послідовності вживання слів для проведення авторської атрибуції англомовних наукових текстів. Порівняння текстів із залученням методу дивергенції Кульбака-Лайблера проведено зі словником “функціональних” слів або словником опорних текстів. Групування текстів за словником “функціональних слів” виявилось придатним для ідентифікації одноосібних праць P. Dorenbos з ефективністю 70%. Ефективність атрибуції наукових статей у співавторстві виявилась низькою для G. Strganyuk та G. Zimmerer. Ефективність атрибуції статей P. Dorenbos із використанням дивергенції Кульбака-Лайблера та послідовності вживання слів на навчальній вибірці для 4 авторів у випадку вибору різних опорних текстів значимо не відрізняються у межах інтервалу $q_{сер}\pm 2\sigma$. Найбільшу ефективність $q_{сер}$=80 % досягнуто для 2-х слів за умови вибору тексту D1 як опорного. Текст D1 використано як опорний для визначення ефективності атрибуції у випадку апробації алгоритму на тестовій вибірці. Середня ефективність атрибуції із використанням дивергенції Кульбака-Лайблера на навчальній вибірці для 4 авторів складає: P. Dorenbos – 80% (2 слова), A. Meijerink – 85% (2 слова), G. Strganyuk – 93% (3 слова), G. Zimmerer – 60% (3 слова). Оптимальні значення атрибуції досягнуто у випадку n=2 та n=3. Зазначимо, що ефективність розпізнавання текстів для кожного з авторів суттєво не відрізняється. Тестування алгоритму групування текстів методом Кульбака-Лайблера проведено на тестовій вибірці. Ефективність атрибуції на тестовій вибірці статей P. Dorenbos склала

q=100% для n=3. Розмір послідовності з трьох слів для тестової вибірки збігається із розміром для навчальної вибірки (2 або 3 слова).

З метою порівняння ефективності визначення автора тексту за різними мірами розходження проаналізовано використання статистики χ^2 для встановлення відстані між досліджуваними науковими текстами. Розрахунок проводили на тій же вибірці текстів, що і в ентропії. Міру χ^2 розраховували для текстів залежно від вибору опорного тексту та розміру послідовності вживання слів. Максимальна ефективність атрибуції статей P. Dorenbos у випадку вибору різних опорних текстів досягнута для послідовності вживання із одного слова ($q_{сер}$=80 %). Для всіх авторів ефективність атрибуції складає $q_{сер}$=66 %. Збільшення розміру послідовності вживання слова у випадку міри χ^2 погіршує ефективність атрибуції текстів. Найкращі результати розпізнавання для дивергенції Кульбака-Лайблера (3 слова, $q_{сер}$=75±24) та міри χ^2 (1 слово, $q_{сер}$=66±18) у межах 2σ інтерквантильного інтервалу не відрізняються. Зіставлення ефективності авторської атрибуції наукових текстів показало, що міра χ^2 ефективна при аналізі послідовності вживання з одного слова, а дивергенція Кульбака-Лайблера – трьох слів. З урахуванням 95% інтерквантильного інтервалу різниці між ефективностями атрибуції для цих методів немає.

Отже, застосування методів одночасного моніторингу групування текстів та відповідних їм слів та ентропії до параметра послідовності вживання слів наукових текстів дозволяє проводити їх тематичну та авторську атрибуцію з можливістю перерозподілу ваги тематичних та авторських ознак шляхом зміни розміру послідовності вживання слів. У випадку послідовності з одного слова відбувається тематична атрибуція текстів. Авторський стиль більше виявляється, коли збільшувати розмір послідовності вживання слів.

Основні положення цього розділу висвітлено у працях автора [22; 23; 24; 25; 250; 251].

РОЗДІЛ 5
АВТОРСЬКА АТРИБУЦІЯ
АНГЛО-, НІМЕЦЬКО- ТА УКРАЇНОМОВНИХ ХУДОЖНІХ ТЕКСТІВ

Як показано у попередньому розділі дисертаційної праці, характер атрибуції текстів залежить від розміру прийнятих до розгляду послідовностей вживання слів, і найбільша здатність методів атрибуції до виявлення авторського стилю виявляється при збільшенні розміру послідовності до трьох, чотирьох слів. Метод одночасного моніторингу групування текстів та відповідних їм слів застосовано до художніх англо-, німецько- та україномовних текстів з метою апробації ефективності даного підходу для авторської атрибуції художніх текстів [26; 27].

5.1 Авторська атрибуція англомовних художніх текстів

Для перевірки ефективності запропонованої у дисертаційній роботі методики визначення автора тексту, авторська атрибуція була проведена для англомовних художніх текстів. До розгляду обрано наступні художні твори XXI століття: *M. Albom* “The five people you meet in heaven” (A1), “Have a little faith” (A2), “Tuesdays with Morrie” (A3), “Time keeper” (A4); *N. Gaiman* “Murder mysteries” (G1), “Neverwhere” (G2), “The Sandman: the dream hunters” (G3), “The Graveyard Book” (G4); *J. Harries* “Chocolat” (H1), “Five quarters of the orange” (H2), “Sleep pale sister” (H3), “Blue eyed boy (H4); *J. Rowling* “Harry Poter and the Sorcerers stone” (R1). У дужках після назви кожного з проаналізованих текстів наведено скорочене позначення тексту, яке використано для візуального представлення результатів аналізу.

У табл. 5.1 наведені перші 35 сполучень із трьох слів, що найчастіше зустрічаються у вибраних англомовних творах кожного з досліджуваних авторів, а також у сумарному масиві проаналізованих текстів. Послідовності вживання з трьох слів були упорядковані у табл. 5.1 за кількістю їх вживання

у текстах кожного з авторів та загальному масиві текстів. Якщо проаналізувати список найчастіше вживаних слів у всіх текстах разом та список найчастіше вживаних слів кожного з авторів, то спільними найчастіше вживаними словами можна назвати: *there was a, out of the, one of the, the back of.* За умов відмінності авторських стилів написання твору, слід очікувати різний набір послідовностей вживання слів, що входять до числа перших, найчастіше вживаних. Дані, наведені в табл. 5.1, чітко відображають вищезгадану закономірність. Так, не всі найчастіше вживані словосполучення автором N. Gaiman притаманні стилю інших трьох проаналізованих авторів. Також не варто очікувати появу найчастіше вживаних слів одного автора у списку найчастіше вживаних слів загального масиву текстів при зіставленні перших десятків найчастіше вживаних сполучень із трьох слів. Слід зауважити, що найбільше навантаження для ідентифікації авторського стилю мають слова, які є притаманними одному автору та відсутні у творах інших авторів, або ж слова, частота вживання яких суттєво відрізняється при розгляді творів різних авторів. Тому не варто очікувати ефективної авторської атрибуції винятково за розгляду найчастіше вживаних слів, оскільки ймовірність перекриття вживання тих самих слів різними авторами є великою. Звичайно, при розгляді послідовності вживання з двох та більше слів така ймовірність перекриття словника різних авторів зменшується. Для англомовних художніх текстів серед найчастіше вживаних 35 послідовностей трьох слів найбільш уживаними виявилися послідовності, побудовані за моделями:

1) **article+noun+preposition** (артикль + іменник + прийменник):
the end of, the back of, the rest of, the sound of, the side of, a piece of;

2) **preposition + article + noun** (прийменник + артикль + іменник):
for a moment, on the floor, at the end, for a while;

3) **conjunction + pronoun + verb** (сполучник + займенник + дієслово):
and I was, and he was, but it was.

Таблиця 5.1

Найчастіше вживані сполучення трьох слів в англомовних текстах

	Загальні	N. Gaimann	J. Harris	M. Albom	J. Rowling
1.	there was a	*there was a*	for a moment	the end of	*out of the*
2.	for a moment	*out of the*	*there was a*	*there was a*	*there was a*
3.	out of the	and then he	a kind of	in front of	*it was a*
4.	it was a	*it was a*	I could see	*it was a*	in front of
5.	in front of	side of the	in spite of	the first time	you know wh
6.	one of the	that he was	looked at me	for a moment	seemed to be
7.	there was no	for a moment	*out of the*	shook his head	*one of the*
8.	the first time	*one of the*	the first time	he had been	the end of
9.	the end of	he had been	there was no	it was the	out of his
10.	it was the	he did not	that I was	I want to	was going to
11.	the back of	*the back of*	I could not	*one of the*	be able to
12.	the rest of	as if it	I tried to	and he was	as though he
13.	side of the	and then she	I told her	one of those	the rest of
14.	for the first	the bottom of	a moment I	there was no	*the back of*
15.	that he was	it had been	her voice was	to be a	we have got
16.	shook his head	and he was	I told him	end of the	there was no
17.	the sound of	it was not	I began to	over the years	back to the
18.	was going to	the end of	the sound of	he tried to	end of the
19.	the side of	in the dark	for the first	*out of the*	going to be
20.	he had been	if he had	*it was a*	at the end	three of them
21.	a couple of	in the darkness	the rest of	a lot of	he was going
22.	to be a	at the bottom	in front of	for the first	in the air
23.	as if it	on the floor	gave me a	there is a	up in the
24.	on the floor	a couple of	I could hear	seemed to be	rest of the
25.	going to be	as if she	I did not	on the floor	was trying to
26.	a piece of	and it was	I had to	went to the	a lot of
27.	back to the	the edge of	*one of the*	he used to	to his feet
28.	and i was	the sound of	and I was	the rest of	he had a
29.	and he was	was going to	for a second	as if it	front of the
30.	that it was	he was not	*the back of*	do you know	he tried to
31.	it had been	and then it	for a while	*the back of*	the other two
32.	seemed to be	but it was	might have been	it was not	what are you
33.	but it was	and he had	it was the	in and out	on top of
34.	at the end	on the other	a couple of	he had to	to get past
35.	for a while	to be a	all the time	he had a	of the way

M. Albom, наприклад, будує моделі у “Time Keeper” таким чином: *Only God can write the end of your story. Sitting now in the back of the limo, Victor realized they had never told him what the watch cost. For the rest of the afternoon, Dor explained his ideas. He crouched on the floor, staring at the incandescent water, desperate, as man grows alone, for the sound of another soul. Upon reaching land, Dor pulled himself up the side of a shipping dock. He removed a piece of clay plugged inside a hole in the upper bowl—the one Nim had mocked—and the water began to drip through, one silent splash after another. She thought for a moment, then shook her head. He turned his back and slid down, sitting on the floor but feeling no floor beneath him. Had he been spared the smaller mistakes in life only to make the biggest one at the end? And he was not about to argue with her.*

Зіставлення найчастіше вживаних послідовностей трьох слів у текстах N. Gaimann, J. Harris, M. Albom та J. Rowling виявило притаманні для кожного з досліджуваних авторів сполучуваності слів (табл. 5.1). Так, N. Gaimann найчастіше вживає послідовність “*there was a*”, яка побудована за моделлю **adverb + verb + article**. Для текстів J. Harris найчастотнішою є послідовність “*for a moment*”, яка побудована за моделлю **preposition + verb + article**. Послідовність “*the end of*” (**article + noun + preposition**) є найчастотнішою для словника M. Albom. Аналіз словника J. Rowling показав, що авторка найчастіше використовує послідовність “*out of the*”, яка побудована за моделлю **adverb + preposition + article**.

Крім виділеної найчастіше вживаної послідовності трьох слів та моделі її побудови для кожного з авторів, серед найчастіше вживаних послідовностей можна виділити й інші, які притаманні для стилю кожного із досліджуваних авторів. З-поміж 35 найчастіше вживаних послідовностей трьох слів для текстів:

1) N. Gaimann можна виділити послідовності трьох слів, утворені за моделлю: **article + noun + preposition** (*the back of, the bottom of, the end of, a couple of, the edge of, the sound of*); **preposition + article + noun** (*for a moment,*

in the dark, at the bottom, on the floor); **conjunction + pronoun + verb** (*and he was, and it was, but it was, and he had*);

2) J. Harris можна виділити сполучення, побудовані за моделлю: **article + noun + preposition** (*a kind of, the sound of, the rest of, the back of, a coupl of*); **pronoun + verb + preposition** (*I tried to, I began to, I had to*); **pronoun + verb + article** (*it was a, it was the*); **pronoun + verb + pronoun** (*I tried to, I told him*);

3) M. Albom можна виділити сполучення, побудовані за моделлю: **article + noun + preposition** (*the end of, a lot of, the rest of, the back of*); **preposition + article + noun** (*for a moment, over the years, at the end, on the floor*); **pronoun + verb + article** (it was a, it was the, he had a); **pronoun + verb + preposition** (*he tried to, he used to, he had to*);

4) J. Rowling можна виділити сполучення, побудовані за моделлю: **noun + preposition + article** (end of the, rest of the, front of the); **pronoun + verb + article** (it was a, he had a).

Як видно із зіставного аналізу найчастіше вживаних послідовностей трьох слів, для кожного з авторів притаманною є як наявність схем послідовностей виділених зі списку найчастіше вживаних послідовностей трьох слів для всіх художніх англомовних текстів, так і послідовностей, які притаманні йому і не належать до характерних найчастіше вживаних послідовностей у інших зіставлюваних авторів.

Послідовність із трьох слів виявилась найбільш ефективним параметром для розмежування особливостей авторських стилів у розглянутих англомовних художніх текстах. Всього було проаналізовано послідовності з двох, трьох та чотирьох слів. Словник послідовності з трьох слів для аналізованих англомовних текстів становить 413605 елементів.

Із залученням методу одночасного моніторингу групування текстів та відповідних їм слів розглянуто можливість одночасної атрибуції текстів чотирьох авторів (M. Albom, N. Gaiman, J. Harries та J. Rowling), де для трьох авторів було прийнято до аналізу по чотири тексти, а для J. Rowling – лише один текст. Групи текстів, що складалися більше, ніж з одного тексту, не

дозволили розмежувати текст R1 (J. Rowling) від загальної сукупності проаналізованих текстів. Для покращення розділення груп текстів різних авторів з розгляду виключено групу творів N. Gaiman (G1, ... , G4). Результати подальшого аналізу, проведеного для двох груп з чотирьох творів (M. Albom та J. Harries) та однієї групи з одного твору (J. Rowling), представлено на рис. 5.1.

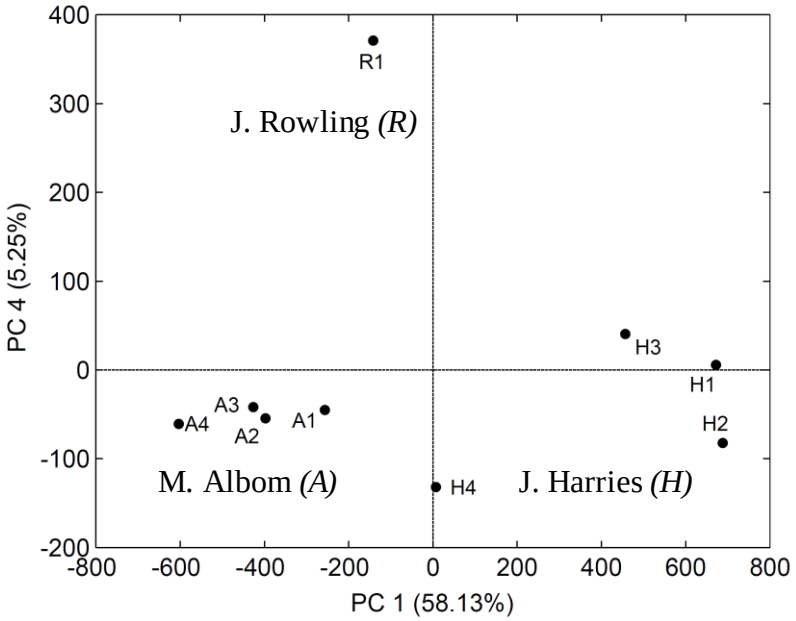


Рис. 5.1. Розподіл текстів за автором із залученням послідовності із трьох слів для англомовних текстів (три автори).

Як видно з розподілу текстів (рис. 5.1), усі проаналізовані тексти були успішно розділені у просторі першої та четвертої головних компонент на три групи відповідно до їх автора. Незважаючи на те, що J. Rowling була представлена одним єдиним твором (R1), стиль цього автора вдалося чітко розділити з-поміж інших авторів.

Аналіз результатів проведеної вище атрибуції текстів дозволяє зробити висновок про доцільність представлення результатів аналізу творів n-авторів у n-1 вимірному просторі. У випадку тривимірного представлення результатів атрибуції (n-1=3), проведеного методом одночасного моніторингу групування текстів та відповідних їм слів, цей метод забезпечує оптимальні результати для чотирьох груп текстів (n=4), що належать різним авторам.

5.2 Авторська атрибуція німецькомовних художніх текстів

Для визначення автора німецькомовних художніх текстів XXI століття проаналізовано: *A. Friedrich* "Süden" (F1), "Totsein verjährt nicht" (F2), "Verzeihen" (F3), "Wie Licht schmeckt" (F4); *K. Gier* "Rubinrot. Liebe geht durch alle Zeiten" (G1); *F. Schätzing* "Keine Angst" (S1), "Lautlos" (S2), "Der Schwarm" (S3), "Tod und Teufel" (S4) та *P. Süskind* "Das Parfum. Die Geschichte eines Mörders" (Su1), "Die Taube" (Su2), "Ein Kampf" (Su3), "Der Kontrabaß" (Su4). У художніх творах німецькомовних авторів виявляються такі ж особливості атрибуції текстів, як і у випадку англомовних художніх текстів. Оптимальних результатів атрибуції німецькомовних текстів за авторським стилем досягнуто при розгляді послідовності трьох слів.

Визначальною особливістю німецькомовних текстів є великий обсяг словника послідовності трьох слів, що сягає 646050 елементів. Спільними найчастіше вживаними словами для всіх аналізованих німецькомовних текстів разом та для кожного автора зокрема є: *Ich weiß nicht, es war ein, an der Wand, schüttelte den Kopf* (табл. 5.2). У німецькомовних художніх текстах найчастотнішою послідовністю трьох слів є *Ich weiß nicht*, яка відповідає моделі **pronoun + verb + particle** (займенник + дієслово + частка). Серед 35 найчастіше вживаних послідовностей трьох слів характерною виявилась також модель типу: **preposition + article + noun** (прийменник + артикль + іменник). Для цієї схеми найчастіше вживаними є послідовності: *in der Hand, auf den Tisch, in der Lage, in der Nacht, in die Augen, in die Höhe, auf der Straße, auf dem Weg, an der Wand, in den Nacke*.

Так, до прикладу, Frank Schätzing вживає виділені найчастіше вживані послідовності трьох слів у "Der Schwarm": *Johanson stand reglos da, den Ärmel in der Hand. Dann kamen zur Weihnachtszeit stattliche Portionen auf den Tisch. Wir sind außerdem in der Lage, die Bewegungen der Gruppen via Echoortung zu*

lokalisieren. In der Nacht ihrer verpatzten Romanze hatte er vermutet, sie werde tags drauf zurück nach Trondheim fahren wollen, aber so war es nicht. Dann sah er Johanson in die Augen und nickte. Der spitze Bug stand steil in die Höhe. Er stand draußen auf der Straße. Jetzt war er auf dem Weg zurück in die Stadt. Er legte den Kopf in den Nacken und schaute in den Himmel.

Для авторів німецькомовних художніх текстів характена така ж тенденція щодо вживання конструкцій найчастіше вживаних послідовностей трьох слів, як і для авторів англомовних художніх текстів. У словнику найчастіше вживаних полідовностей трьох слів наявні як моделі послідовностей, характерні для всіх німецькомовних творів, так і моделі, які притаманні певному автору. Так, у кожного з авторів наявна модель типу **preposition + article + noun**, проте вона є не настільки поширеню у текстах F. Schätzing, як у інших авторів. K. Gier, наприклад, використовує меншу кількість різних типів моделей послідовностей трьох слів, що входять до числа перших 35 послідовностей, ніж інші автори. Наведемо приклади найчастіше вживаних послідовностей трьох слів та їх моделі для кожного з авторів.

Таблиця 5.2

Найчастіше вживані сполучення трьох слів у німецькомовних текстах

	Total	A. Friedrich	K. Gier	F. Schätzing	P. Süskind
1.	*ich weiß nicht*	in der Nähe	in der Zeit	*schüttelte den Kopf*	und wenn er
2.	hin und her	in der Hand	auf jeden Fall	machte eine Pause	in seinem Leben
3.	*es war ein*	*ich weiß nicht*	in die Vergangenheit	einen Moment lang	in der Tat
4.	nach einer weile	in der Nacht	*schüttelte den Kopf*	nach einer weile	für einen Moment
5.	in der Hand	die ganze Zeit	um die ecke	*ich weiß nicht*	gar nicht mehr
6.	den Kopf und	und ich hab	vor meinen Augen	so gut wie	dass er sich
7.	zum ersten mal	*es war ein*	sagte sie und	den Kopf und	er hatte sich
8.	die ganze Zeit	auf den Boden	die Tür hinter	im selben Moment	in die Augen
9.	vor sich hin	zum ersten	*es war ein*	*es war ein*	aus der Rue

		mal			
10.	auf den Tisch	auf der Straße	die ganze Zeit	hin und her	in der Lage
11.	in der Lage	in der Küche	zum ersten mal	runzelte die Stirn	*ich weiß nicht*
12.	sich auf die	in der Stadt	auf der Stelle	ich weiß nicht	er in der
13.	in der Nacht	das hab ich	zuckte mit den	aber es war	und das war
14.	in der nähe	in der Wohnung	und es war	sich auf die	hatte er sich
15.	im selben moment	*schüttelte den Kopf*	nur ein paar	und so weiter	auf der Straße
16.	ich weiß nicht	den ganzen Tag	dass ich mich	blick auf die	am Ende des
17.	in die Augen	er sich nicht	das erste mal	in den letzten	und so weiter
18.	sich auf den	an der Tür	auf der Treppe	im nächsten Moment	es war nicht
19.	hatte er sich	vor sich hin	und sah mich	einen Blick auf	in der Stadt
20.	in die Höhe	sie sich nicht	*an der Wand*	sich auf den	auf der Welt
21.	aber es war	und er hatte	in der zwischenzeit	sich in den	wenn er sich
22.	auf diese weise	das war ein	ich gar nicht	der anderen seite	nach und nach
23.	dass er sich	hin und her	in die Höhe	sich über die	*es war ein*
24.	der anderen Seite	und er hat	noch gar nicht	auf die Idee	das war der
25.	er sich nicht	auf dem Tisch	drehte sich zu	beugte sich vor	in der Hand
26.	auf der anderen	*an der Wand*	gar nicht mehr	den Kopf in	*an der Wand*
27.	den Kopf in	auf dem Weg	*ich weiß nicht*	hatte er sich	nicht einmal mehr
28.	er hatte sich	und wenn ich	sah aus wie	vor sich hin	und in der
29.	drehte sich um	in die höhe	doch gar nicht	in die luft	auf jeden fall
30.	auf der Straße	und ich dachte	dass ich das	ein weiteres mal	sich nicht mehr
31.	und so weiter	er in der	mit der Hand	drehte sich um	wenn man sie
32.	auf dem Weg	auf dem Boden	noch nicht mal	hin und wieder	sich in der
33.	*an der Wand*	in seinem Kopf	nur ein bisschen	auf der anderen	er sich nicht
34.	in den Nacken	und ich hatte	aus dem Gesicht	*an der wand*	und als er
35.	*schüttelte den Kopf*	sich auf den	sagte ich zu	in den Nacken	*schüttelte den Kopf*

Для словника послідовностей трьох слів A. Friedrich притаманними є моделі типу: **preposition + article + noun** (*in der Nähe, in der Hand, in der Nacht, auf dem Boden*); **conjunction + pronoun + verb** (*und Ich hab, und er hatte, und er hat, und Ich dachte, und Ich hatte*); **article + adjective + noun** (*die ganze Zeit, den ganzen Tag*).

У текстах K. Gier найчастіше зустрічаються моделі: **preposition + article + noun** (*in der Zeit, in die Vergangenheit, um die Ecke, auf der Stelle, auf der Treppe, an der Wand, in der Zwischenzeit, in die Höhe, mit der Hand, aus dem Gesicht*); **verb + pronoun + preposition** (*drehte sich zu, sagte Ich zu*).

Серед найчастотніших моделей уживання послідовності трьох слів у текстах F. Schätzing можна виділити: **preposition + article + noun** (*nach einer Weile, auf die Idee, in die Luft, an der Wand, in den Nacken*); **verb + article + noun** (*schüttelte den Kopf, machte eine Pause, runzelte die Stirn*); **preposition + adjective + noun** (*im selben Moment, im nächsten Moment*); **pronoun + preposition + article** (*sich auf den, sich in den*); **article + adjective + noun** (ein weiteres Mal, der anderen Seite); **article + noun + preposition** (*den Kopf in, einen Blick auf*).

У словнику послідовностей трьох слів P. Süskind характерними моделями є: **preposition + article + noun** (*in der Tat, für einen Moment, in die Augen, aus der Rue, in der Lage, auf der Straße, in der Stadt, auf der Welt, in der Hand, an der Wand*); **conjunction + conjunction + pronoun** (*und wenn er, und als er*); **pronoun + verb + particle** (*Ich weiß nicht, es war nicht*); **pronoun + preposition + article** (*er in der, sich in der*).

Групування німецькомовних художніх текстів за автором здійснено із залученням методу одночасного моніторингу групування текстів та відповідних їм послідовностей трьох слів. Одночасний аналіз текстів чотирьох авторів, де представлено як групи, куди входили по декілька творів одного автора (група ***S*** - F. Schätzing, ***F***- A. Friedrich та ***Sü*** - P. Süskind), так і група, де взято до аналізу лише один текст автора (група ***G*** - *K. Gier*), показав, що важко було досягти позитивних результатів групування у

випадку, коли автор представлений лише одним текстом. Тому відповідно до закономірностей, виявлених для авторської атрибуції англомовних текстів, кількість німецькомовних авторів була зменшена до трьох (S, F, G) з метою досягнення оптимальної авторської атрибуції для її представлення у двовимірному просторі.

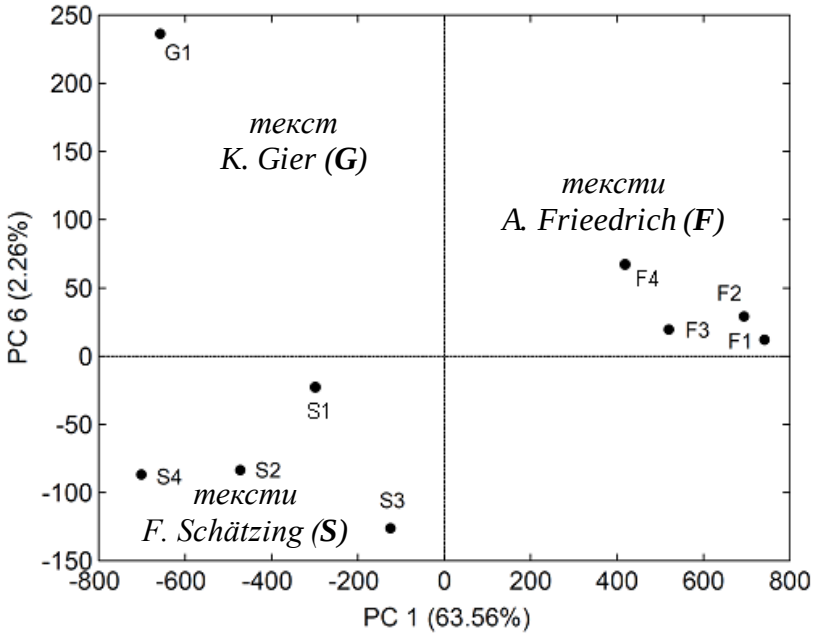


Рис. 5.2. Розподіл текстів за автором із залученням послідовності трьох слів для німецькомовних текстів (три автори).

Результати атрибуції творів трьох авторів F. Schätzing (група S), A. Friedrich (група F) та K. Gier (група G) у двовимірному просторі представлено на рис. 5.2. Досягнуто чіткого розмежування груп творів A. Friedrich та F. Schätzing за першою компонентою, а розмежування твору K. Gier від творів інших авторів – за шостою головною компонентою (рис. 5.2).

5.3 Авторська атрибуція україномовних художніх текстів

Авторська атрибуція україномовних художніх текстів методом одночасного моніторингу групування текстів та відповідних їм слів апробована для таких творів XXI століття: *Ю. Андрухович* “Дванадцять обручів” (A1); *О. Забужко* “Дівчатка” (Z1), “Музей покинутих секретів” (Z2),

“Польові дослідження з українського сексу” (Z3), “Сестро, сестро” (Z4) (група Z); *Л. Костенко* “Берестечко” (K1), “Маруся Чурай” (K2), “Скіфська одіссея” (K3), “Записки українського самашедшого” (K4); *Ю. Покальчук* “Безмежність” (P1), “Озерний вітер” (P2), “Хлопці від Катеринки” (P3), “Заборонені ігри” (P4). Як і для англо- та німецькомовних художніх текстів, послідовність трьох слів виявилась найбільш ефективним параметром для розмежування особливостей авторських стилів у розглянутих україномовних художніх текстах. Обсяг словника послідовності з трьох слів для розглянутих україномовних текстів становить 318137 елементів.

На відміну від художніх англомовних та німецькомовних творів, в україномовних художніх творах виявлено дещо менше лексичне перекриття у словнику послідовностей трьох слів при аналізі його перших 35 елементів. Так, серед перших 35 найчастіше вживаних послідовностей трьох слів спільними виявлено лише дві послідовності, притаманні як для загального словника, так і для словників кожного із досліджуваних авторів: *так і не, що в нього* (табл. 5.3). Така особливість україномовних художніх текстів може бути пояснена значно чисельнішим набором можливих загальновживаних комбінацій послідовностей з трьох слів, що суттєво зменшує перекриття словників художніх україномовних творів різних авторів. В українській мові відсутні граматичні кліше, притаманні англійській та німецькій мовам.

Таблиця 5.3

Найчастіше вживані сполучення з трьох слів в україномовних текстах

	Загальні	Ю. Андрухович	О. Забужко	Л. Костенко	Ю. Покальчук
1.	*так і не*	до того ж	*так і не*	це ж не	і я не
2.	та ще й	*так і не*	раз у раз	та ще й	раз по раз
3.	раз у раз	аж ніяк не	та ще й	я ж не	ні з ким
4.	що все знають	на всі боки	на відміну від	я вже не	все буде добре
5.	до того ж	час до часу	ще й як	вже й не	і я вже
6.	я вже не	швидше за все	так воно й	у них там	але ж і
7.	це ж не	чи то пак	на той час	у нас є	я не можу
8.	час від часу	одна з них	я не знаю	а в нас	що з ним
9.	я ж не	той же час	можна було б	час від часу	не було в

10.	я не знаю	з усіх сил	що в неї	що ж ти	що я не
11.	вже й не	він так і	*що в нього*	що у нас	та ще й
12.	що ж ти	час від часу	я вже не	не те що	от і все
13.	і я не	так само не	як і в	*так і не*	час від часу
14.	але ж і	але він не	до того ж	ті ж самі	ні про що
15.	а в нас	і тут таки	в ту мить	то що ж	не можна було
16.	аж ніяк не	та що там	ні до чого	уже й не	у нас в
17.	але я не	з іншого боку	тато з мамою	я вже й	і все тут
18.	на той час	як не дивно	його вже не	а що ж	і одного разу
19.	не те що	але все це	не можна було	але ж і	що це не
20.	можна було б	але це не	вже не було	чого ж ти	в цю мить
21.	у нас тут	в одному з	в себе в	то чого ж	в мене вже
22.	в цю мить	в нього з	тут таки й	а це вже	а в мене
23.	на відміну від	там не було	все таки не	це ж треба	як і всі
24.	що я не	один із них	не в змозі	як на мене	ніби й не
25.	у них там	й так само	але то вже	це у нас	в ту мить
26.	все ж таки	і тільки тут	та й не	а ще ж	що з тобою
27.	а у нас	ніхто з них	а як же	а то ж	все ж таки
28.	і я вже	що все це	це ж не	а я ж	що я тобі
29.	я ніколи не	а відтак і	що ж ти	наче й не	він прийшов до
30.	я нічого не	сам на сам	так уже й	чого ж ви	й не було
31.	*що в нього*	*що в нього*	не може бути	що я тут	*що в нього*
32.	що з ним	з огляду на	на мене з	по той бік	він дивився на
33.	ні з ким	і все ж	й без того	а тут ще	ніхто не знає
34.	вже не було	як би це	не інакше як	на весь світ	але я не
35.	а що ж	можна було б	не те щоб	*що в нього*	*так і не*

Послідовність трьох слів “*так і не*” є найчастіше вживаною у проаналізованих україномовних художніх текстах і будується за моделлю **прислівник + сполучник + частка**. Іншими типовими моделями для послідовності трьох слів в україномовних художніх текстах є:

1) **сполучник + займенник + частка** (*і я не, але я не, що я не, і я вже*);

2) **займенник + частка + частка** (*я вже не, це ж не*);

3) **сполучник + прийменник + займенник** (*а в нас, але я не, що я не*).

До прикладу, наведемо такі речення з тексту Ю. Андруховича “Дванадцять обручів”: *Альказар – так, здається, називалася фортеця, хоч минуло багато років <u>і я не</u> можу пам'ятати з певністю. Богдан того разу змовчав, <u>але я не</u> подумав би ніколи, що зі страху.*

Найчастіше вживаною послідовністю трьох слів у творах Ю. Андруховича є "*до того ж*", яка будується за моделлю **прийменник + займенник + частка**. Порівняно з іншими авторами, Ю. Андрухович не так часто використовує займенники у виявлених найчастіше вживаних послідовностях трьох слів. Серед інших 35 найчастіше вживаних послідовностей трьох слів можна виділити такі, що будуються за моделями: **іменник + прийменник + займенник** (*одна з них, один із них*), **сполучник + займенник + частка** (*так само не*).

О. Забужко найчастіше вживає послідовність "*так і не*", яка відповідає моделі **прислівник + сполучник + частка**. Виділяється також у її творах послідовність *тато з мамою,* яка не є притаманнною для списку перших 35 найчастіше вживаних слів як для україномовних, так і для англо- й німецькомовних досліджуваних авторів. Характерними для цієї авторки є моделі: **сполучник + прийменник + займенник** (*що в неї, що в нього*).

Послідовність "*це ж не*" є найчастотнішою для словника послідовностей трьох слів Л. Костенко, яка утворена за моделлю **займенник + частка + частка.** Для цієї авторки, на відміну від інших україномовних авторів, характерна тенденція частого використання у перших 35 найчастіше вживаних послідовностях трьох слів частки "*ж*", наприклад, *я ж не, ті ж самі, а що ж, чого ж ти, а я ж* тощо. Часто вживаними моделями є також: **займенник + частка +частка** (*я ж не, я вже не*); **сполучник + прийменник +займенник** (*а в нас, що у нас*).

У текстах Ю. Покальчука найчастіше зустрічається послідовність трьох слів "*і я не*", що відповідає моделі **сполучник + займенник + частка**. Для цього автора характерні й послідовності трьох слів, які можна розглядати як прості речення, наприклад: *все буде добре, що з ним, що з тобою*. У жодного з інших досліджуваних авторів не було виявлено серед перших 35 найчастіше вживаних послідовностей трьох слів таких, що можуть утворювати повноцінне речення. У Ю. Покальчука наявні послідовності трьох слів, які утворюються за моделлю: **сполучник + прийменник + займенник** (*що з*

ним, а в мене, що в нього); **займенник + дієслово + прийменник** (*він прийшов до, він дивився на*).

Результати аналізу україномовних художніх текстів методом одночасного моніторингу групування текстів та відповідних їм послідовностей з трьох слів, представлено на рис. 5.3. Для випадку україномовних текстів, результати атрибуції творів 4 авторів є вже задовільними у їх двовимірному представленні. Значно менше перекриття словників найчастіше вживаних послідовностей трьох слів для української мови (на противагу англійській та німецькій) дозволяє проводити атрибуцію із залученням аналізу вживання високочастотних слів, що виявляється у результатах авторської атрибуції, представленої у просторі перших головних компонент. Якщо одночасно аналізувати твори чотирьох авторів, то чітко виділяються твори Ю. Покальчука (група P) та твір Ю. Андруховича (A1), проте групи творів О. Забужко (група Z) та Л. Костенко (група K) добре не розмежовуються.

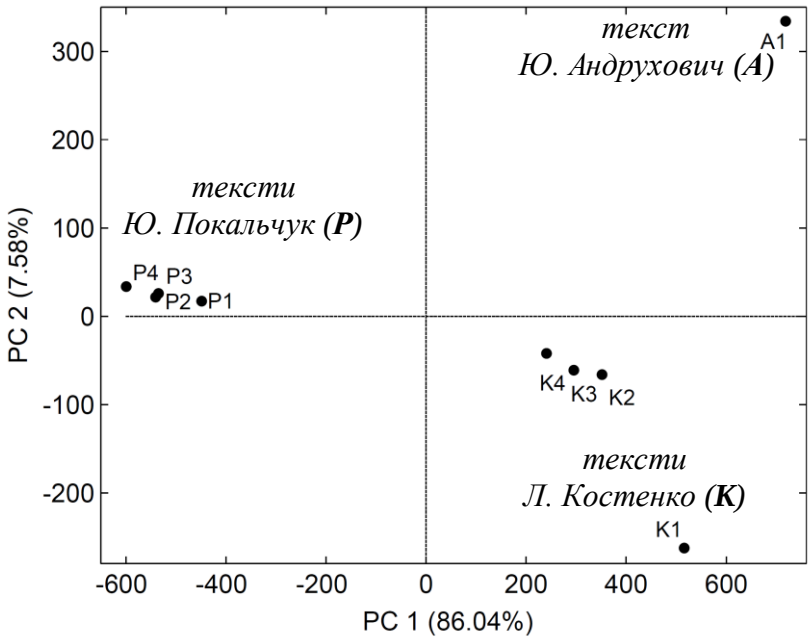


Рис. 5.3. Розподіл текстів за автором із залученням послідовності вживання трьох слів для україномовних текстів (три автори).

Як видно з рисунку 5.3, розділення текстів різного авторства значно покращується для художніх текстів україномовних авторів, якщо

наблизитись до умов, виявлених при аналізі англомовних художніх текстів (представлених творів *n* авторів у *n*-1 вимірному просторі). Досягнуто чіткого розмежування груп творів Ю. Покальчука та Л. Костенко за першою головною компонентою, а твір Ю. Андруховича виділено за першою та другою головною компонентою. Авторська атрибуція текстів виявилась успішною завдяки вживанню авторами послідовностей з трьох слів, які були унікальними і не зустрічались у творах інших авторів.

Висновки до розділу 5

Авторська атрибуція художніх текстів, проведена методом одночасного моніторингу групування текстів та відповідних їм слів, показала, що для англо-, німецько- та україномовних художніх текстів оптимальне розділення за авторським стилем досягається при аналізі послідовностей трьох слів. Зі збільшенням розміру послідовності слів до чотирьох, п'яти або ж його зменшення до двох ефективність атрибуції методом одночасного моніторингу групування текстів та відповідних їм слів суттєво зменшується.

Аналіз найчастіше вживаних послідовностей трьох слів у проаналізованих англо-, німецько- та україномовних художніх текстах дав змогу встановити міру їх перекриття, яка виявилася співмірною для англо- та німецькомовних художніх текстів, однак таке перекриття словників найчастіше вживаних слів виявилось значно меншим для україномовних художніх текстів. Більша кількість однакових елементів у словниках найчастіше вживаних послідовностей трьох слів англо- та німецькомовних художніх текстів (міра їх перекриття) пояснюється наявністю усталених граматичних зворотів, притаманних для англійської та німецької мов. На відміну від англійської та німецької мов, в українській мові відсутнє чітке регулювання порядку слів у реченні, що є причиною виявленої малої ймовірності перекриття елементів словників найчастіше вживаних послідовностей трьох слів.

Авторська атрибуція методом одночасного моніторингу групування текстів та відповідних їм слів реалізується на основі виявлення притаманних авторському стилю сполучень слів. Внаслідок часткового перекриття словників найчастіше вживаних авторами послідовностей слів англійської та німецької мов, авторська атрибуція художніх текстів даними мовами не може бути ефективно реалізованою із залученням часто вживаних послідовностей слів. У зв'язку з цим, для авторської атрибуції англо- та німецькомовних текстів характерною є необхідність аналізу низькочастотних сполучень слів для яких авторські стилі є добре розділеними. Виявлене для української мови мале перекриття словників найчастіше вживаних авторами послідовностей трьох слів дозволяє ефективно використовувати часто вживані сполучення трьох слів для авторської атрибуції україномовних художніх текстів.

Результати авторської атрибуції художніх текстів методом одночасного моніторингу групування текстів та відповідних їм слів можна візуально представити у двох- або тривимірній Декартовій системі координат. Показано, що оптимальних результатів групування художніх англо-, німецько- та україномовних текстів за автором цим методом атрибуції досягається при представленні результатів розподілу текстів трьох авторів у двовимірній системі координат, або текстів чотирьох авторів у тривимірній системі координат.

Проаналізований фактичний матеріал свідчить про те, що найбільш уживаним послідовностями трьох слів в а) англомовних текстах є *there was a*; б) німецькомовних текстах – *Ich weiß nicht*; в) україномовних текстах – *так і не*. Виявлено спільну тенденцію для трьох досліджуваних мов: ймовірність появи у текстах однакових послідовностей трьох слів, навіть якщо аналізувати перші найчастіше вживані послідовності, є значно меншою порівняно з випадком аналізу частоти появи одного слова. Серед перших 35 найчастіше вживаних послідовностей трьох слів однаковими є лише п'ять послідовностей із трьох слів для англійської мови, чотири – для німецької мови та дві – для української мови.

Зіставлення найчастіше вживаних послідовностей трьох слів та моделей їх утворення в англо-, німецько- та україномовних художніх творах показало, що ймовірність появи однакових схем послідовностей з трьох слів серед перших найчастіше вживаних 35 послідовностей в україномовних авторів є меншою, ніж в англо- та німецькомовних.

Спільною найчастотнішою моделлю для послідовності трьох слів з-поміж проаналізованих англомовних та німецькомовних художніх текстів є “прийменник + артикль + іменник”. Важко простежити наявність спільної моделі для англо-, німецько- та україномовних художніх текстів серед перших 35 найчастіше вживаних послідовностей трьох слів.

Для кожного з авторів визначено перші 35 найчастіше вживаних послідовностей трьох слів. Здійснено їх аналіз та визначено типові моделі, за якими будуються ці послідовності. Зі зростанням порядкового номеру послідовності трьох слів суттєво зменшується ймовірність появи у авторів спільних часто вживаних їх послідовностей. Саме наявність характерних послідовностей трьох слів, притаманних лише певному автору, лежить в основі авторської атрибуції англо-, німецько та україномовних художніх текстів.

Основні положення цього розділу висвітлено у працях автора [26; 27].

ВИСНОВКИ

Теоретико-методологічні засади дисертаційного дослідження ґрунтуються на: 1) розкритті дефініції поняття атрибуції у сучасному мовознавстві, зіставленні термінів стилеметрія, атрибуція, класифікація, кластеризація; 2) особливостях атрибуції наукових текстів, які є результатом роботи колективу авторів, що ставить високі вимоги до вибору оптимальних параметрів та ефективності методів атрибуції тексту; 3) класифікації вихідних параметрів атрибуції текстів на мовні (синтаксичні, лексичні, морфологічні, помилки) та позамовні (структурні дані) параметри; 4) виборі оптимальних методів для опрацювання масивів багатовимірних даних; 5) застосуванні лінгвостатистичних методів, які забезпечують об'єктивність результатів атрибуції текстів. Практична основа виконання завдань дисертації – розробка методів та програм для: а) апроксимації рангового розподілу слів у текстах; б) підрахунку абсолютних та відносних частот появи слів та словосполучень у тексті в) формування частотної матриці появи слів та словосполучень у тексті; г) обчислення міри хі-квадрат текстів та ентропії (дивергенція Кульбака-Лайблера), порівняно зі словником опорного тексту, або словником "функціональних" слів.

Розроблена комплексна методика аналізу стильової, тематичної та авторської атрибуції англо-, німецько- та україномовних наукових і художніх текстів виявилась ефективною для стильової, тематичної та авторської атрибуції досліджуваних текстів. Методика передбачає формування репрезентативної вибірки текстів, вибір оптимальних параметрів тексту, визначення їх абсолютних частот, обробку статистичних даних різними лінгвостатистичними методами.

Основним критерієм розмежування англо-, німецько- та україномовних текстів наукового і художнього стилів є залежність рангово-частотного розподілу слів у тексті. Нові можливості у проведенні стильової атрибуції текстів надають модифіковані формули, що забезпечують апроксимацію

рангово-частотного розподілу слів у тексті. Жодна із наявних модифікацій рангово-частотного закона Ципфа не забезпечує точного відтворення розподілу слів у тексті одночасно для високочастотних та низькочастотних слів. Дослідження широкого вибору модифікацій закону Ципфа дозволило виділити: а) закон Мандельброта, який описує розподіл високочастотних слів у тексті та б) закон Юла-Саймона, який описує розподіл низькочастотних слів у тексті. Запропоновано модифікацію показникової функції Ципфа у записі Лавалетті *f(k;q;s;n)*, що передбачає два параметри (*q* та *s*) для апроксимації рангово-ймовірнісного розподілу слів. Перевага запропонованої формули над іншими в тому, що вона точно і з малою кількістю параметрів описує розподіл слів у тексті. Вперше для математичного параметра *s* запропоновано лінгвістичне тлумачення – відповідає за функціональний стиль. Параметр *s* набуває характерних значень залежно від стилю тексту – наукового або художнього. Стрімкість спаду ймовірності появи слова виявилась меншою у наукової літературі, а отже, і значення параметра *s* для наукових текстів є меншим, порівняно з художньою літературою. Детальне дослідження зміни величини параметра *s* залежно від приналежності тексту до наукового або художнього стилів дозволило встановити діапазон коливань кількісних показників параметра *s* для англо-, німецько- та україномовних наукових і художніх текстів. Інтерквантильні інтервали параметра *s* для вибірок англомовних текстів наукової літератури [0.85 – 1.01] та художньої літератури [1.04 – 1.12] не перекриваються. Розбіжності між текстами наукового та художнього стилів англійської мови за параметром *s* є статистично значимими, а самі параметри апроксимаційної кривої для наукових *s*=0.93±0.08 і художніх текстів *s*=1.08±0.04 можуть бути використані для визначення приналежності тексту до художнього або наукового стилю англійської мови. Для німецькомовних текстів наукового та художнього стилю розбіжності між значеннями параметра *s* є також статистично значимими. Параметр *s* для наукових текстів в межах *2σ* інтерквантильного інтервалу становить [0.86 – 0.94], а художніх – [0.97 –

1.13]. Діапазон коливань параметра *s* в україномовних наукових текстах є [0.82 – 0.86] і не перекривається з межами *2σ* інтерквантильного інтервалу художніх текстів [0.89 – 0.97]. Якщо кількісний показник параметра s певного тексту входить у вище встановлені межі інтерквантильного інтервалу, то цей текст є науковий або художній.

Зіставлено рангово-частотні розподіли слів англо-, німецько- та україномовних наукових і художніх текстів. Спільною у науковому тексті для трьох мов є наявність у першій тридцятці найчастіше вживаних слів лише службових частин мови та дієслова *бути*, відсутність займенників, іменників. Спільним у художніх текстах є висока частота вживання займенників (займенник "*I*", "*Ich*", "*Я*" має подібну частоту вживання у зіставлюваних мовах), найчастотнішим сполучником є "*and*", "*und*","*і*". Для англо-, німецько- та україномовних текстів спільними найчастіше вживаними є слова: 1) *in – in – в, also – auch – також, and – und – і (та), from – von – від, with – mit – з, as – als – як, is – ist – є, on – an (auf) – на, not – nicht – не, be – werden – бути, by (at) –bei – при* (науковий текст); 2) *and – und – і (та), I – Ich – я, he – er – він, was – war – було, in – in – в (у), it – das – це, you – du – ти, on – auf – на, she – sie – вона, said – sagte – казав, with – mit – з, but – aber – але, as – als – як* (художній текст). Серед визначених перших 300 найчастіше вживаних слів для наукових текстів характерним є вживання загальнонаукових термінів, а для художніх текстів – слів на позначення частин тіла та періодів дня.

Тематична атрибуція вузькоспеціалізованих наукових робіт з фізики була успішно здійснена за використання методу одночасного моніторингу групування текстів і відповідних їм слів (метод аналізу головних компонент). Застосування цього методу дозволяє виділити основні характеристики аналізованих текстів та здійснити розподіл текстів на групи відповідно до прояву в них виділених основних характеристик. Метод не вимагає попереднього опрацювання тексту, тобто вибору характерних параметрів атрибуції текстів, таких, як довжина слова, речення, частота вживання різних

частин мови тощо. У роботі апробовано його ефективність для тематичної атрибуції наукових праць VI-ої Міжнародної конференції LUMDETR – 2006 у галузі люмінесцентного матеріалознавства і виділено п’ять тематичних секцій конференції замість дев’яти, заявлених у програмі конференції. На результат розподілу вплинули наявність тематики зі спільними, усталеними та поширеними множинами словоформ, поява нових напрямів досліджень, присутність робіт зі спільними об’єктами досліджень. У дисертації показано, що одночасний аналіз статей та відповідних їм тез, анотацій і заголовків як окремих елементів вибірки сприяє кращому розподілу основних характеристик текстового масиву та може бути використаний для оцінки відмінності текстів опублікованої статті від початково заявлених тез.

Запропоноване у дисертації поєднання методу одночасного моніторингу групування текстів і відповідних їм слів із аналізом послідовності вживання чотирьох слів є ефективним для авторської атрибуції наукових текстів. У роботі показано, що стиль автора виявляється у вживанні характерних послідовностей чотирьох слів для: 1) опису спостережуваних об’єктів; 2) вираження припущень; 3) представлення та порівняння результатів. Зіставлення словників послідовностей чотирьох слів різних авторів у наукових текстах показало, що у авторів наукових текстів, порівняно з авторами художніх текстів, дуже мала кількість спільних послідовностей чотирьох слів. Так, серед перших 70 найчастіше вживаних послідовностей виявлено лише одну спільну послідовність чотирьох слів “*the excitation spectrum of*” у текстах P. Dorenbos, A. Meijerink, G. Strganyuk та G. Zimmerer. Послідовність “*in the case of*” є спільною тільки для статей P. Dorenbos, G. Stryhanyuka і G. Zimmerer. У кожного автора наявна лише одна послідовність чотирьох слів, яка вживалася б у всіх його текстах (“*the energy of the*” – P. Dorenbos, “*the intensity of the*” – A. Meijerink, “*in the range of*” – G. Strganyuk, “*at the superlumi station*” – G. Zimmerer). У словнику довільного автора можна зустріти і такі послідовності, які відтворюють навіть частину речення. Це відбувається, коли автор копіює речення з однієї статті

та вставляє його в іншу. Найчастіше вживані послідовності чотирьох слів у наукових текстах можна згрупувати на такі, що: 1) виражають припущення; 2) візуально представляють результати дослідження; 3) згадують місце проведення експерименту, назви приладів. Спільними найчастіше вживаними моделлями послідовностей чотирьох слів у наукових текстах є: article + noun + preposition + article (наприклад, *the size of the, the transition of the*) та preposition + article + noun + preposition (наприклад, *on the basis of, for the formation of*).

Для авторської атрибуції англо-, німецько- та україномовних художніх текстів, визначення автора методом одночасного моніторингу групування текстів і відповідних їм слів сягає максимальної ефективності при аналізі послідовності трьох слів. Зменшення оптимального розміру послідовності слів для авторської атрибуції художніх текстів пояснюється відсутністю назв об'єктів дослідження у художній літературі. Аналіз найчастіше вживаних послідовностей трьох слів у проаналізованих художніх текстах дав змогу встановити послідовності, що зустрічаються у всіх текстах: для англо- та німецькомовних художніх текстів (*there was a, out of the, one of the, the back of* та *Ich weiß nicht, es war ein, an der Wand, schüttelte den Kopf*), однак для україномовних художніх текстів кількість таких послідовностей є меншою (*так і не, що в нього*). *There was a, Ich weiß nicht, так і не* – найчастіше вживані послідовності трьох слів у художніх текстах зіставлюваних мов. Для художніх текстів характерними є послідовності, побудовані за моделями: 1) article+noun+preposition (*the end of*), preposition + article + noun (*for a moment*), conjunction + pronoun + verb (*and I was*) – англомовні тексти; 2) pronoun + verb + particle (*Ich weiß nicht*), preposition + article + noun (*in der Hand*) – німецькомовні тексти; 3) сполучник + займенник + частка (*і я не*), займенник + частка + частка (*я вже не*), сполучник + прийменник + займенник (*а в нас*) – україномовні тексти. Кожен із досліджуваних авторів має різну найчастіше вживану послідовність трьох слів: *there was a* (N. Gaimann), *for a moment* (J. Harris), *the end of* (M. Albom), *out of the* (J. Rowling),

in der Nähe (A. Friedrich), *in der Zeit* (K. Gier), *schüttelte den Kopf* (F. Schätzing), *und wenn er* (P. Süskind), *до того ж* (Ю. Андрухович), *так і не* (О. Забужко), *це ж не* (Л. Костенко), *і я не* (Ю. Покальчук). З-поміж досліджуваних авторів можна виділити Ю. Покальчука. Для текстів Ю. Покальчука притаманні послідовності трьох слів, які можна розглядати як ціле речення (*все буде добре, що з ним, що з тобою*). Таке використання найчастіше вживаних послідовностей трьох слів не простежується в інших досліджуваних авторів.

Перспективами подальших досліджень є 1) здійснення тематичної (авторської) атрибуції текстів різних функціональних стилів методом одночасного моніторингу групування текстів та відповідних їм слів та методом ентропії для різних груп мов; 2) створення нових тематичних словників на базі семантично зв’язаних слів, що формують основні характеристики тексту; 3) визначення автора перекладу статей у наукових журналах, які перекладаються з української на англійську мову; 4) зіставлення закономірностей зміни рангово-частотного розподілу слів для одного наукового (художнього) тексту, перекладеного різними мовами.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Айвазян С. А. Классификация многомерных наблюдений / С. А. Айвазян, З. И. Бежаева, О. В. Староверов. – М. : Статистика, 1974. – 239 с.

2. Айвазян С. А. Прикладная статистика. Основы моделирования и первичная обработка данных / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. – М. : Финансы и статистика, 1983. – 471 с.

3. Алексеев П. М. О нелинейных формулировках закона Ципфа / П. М. Алексеев // Вопросы кибернетики. – М. ; Л. – 1978. – Вып. 41.– С. 53–65.

4. Андреев С. Н. Многомерный анализ разноуровневых признаков английских аффиксальных глаголов / С. Н. Андреев, Ю. А. Тулдава // Уч. зап. Тартус. ун-та. – Вып. 872: Квантитативная лингвистика и автоматический анализ текстов. – Тарту, 1989. – С. 3–11.

5. Балли Ш. Французская стилистика / Ш. Балли // Под ред. Е. Г Эткинда. – М. : Изд-во иностр. лит., 1961. – 341 с.

6. Баранов А. Н. Введение в прикладную лингвистику / Анатолий Николаевич Баранов. – М. : Эдиториал УРСР, 2001. – 360 с.

7. Батов В. И. Опыт построения методики для установления авторства текстов / В. И. Батов, Ю. А. Сорокин // Язв. АН СССР. Сер. Литература и язык. – М. – 1977. – Т. 36. – № 4. – С. 345–347.

8. Берков П. Н. Об установлении авторства анонимных и псевдонимных произведений XVIII века / П. Н. Берков // Русская литература. – 1958. – № 2. – С. 180–189.

9. Бобкова Н. С. Типология заглавий в научно-технической периодике / Н. С. Бобкова, В. К. Курчаева // Отраслевая терминология и ее структурно-типологическое описание : межвуз. сб. науч. тр. – Воронеж : Изд. Воронежского университета, 1988. – С. 150–158.

10. Бойко Ю. Диференційні параметри речення як детермінанта авторського стилю / Ю. Бойко // Проблеми квантитативної лінгвістики : зб. наук. пр. – Чернівці: Рута, 2005. – С. 292–305.

11. Бондаренко Т. Ф. Выделение морфологических типов словоизменения в различных функциональных стилях / Т. Ф. Бондаренко // Вопросы статистической стилистики / Отв. ред. Б. Н. Головин. – К. : Наукова думка, 1974. – С. 86–96.

12. Браверман Э. М. Структурные методы обработки эмпирических данных / Э. М. Браверман, И. Б. Мучник. – М. : Наука, 1983. – 203 с.

13. Бук С. Н. Основи статистичної лінгвістики / С. Н. Бук / Відп. ред. проф. Ф. С. Бацевич. – Львів : Видавничий центр ЛНУ імені Івана Франка, 2008. – 124 с.

14. Васильєв Ю. А. О влиянии композиционно-смысловой организации научного текста на его языково-стилистические характеристики / Ю. А. Васильєва // Стиль научной речи / Отв. ред. Е. С. Троянская. – М. : Наука, 1978. – С. 75–94.

15. Вашак П. Длина слова и длина предложения в текстах одного автора / П. Вашак // Вопросы статистической стилистики / Отв. ред. Б. Н. Головин. – К. : Наукова думка, 1974. – С. 314–329.

16. Визе Е. Концепция авторской “точки зрения” и вопросы ее выражения в текстах научного стиля / Е. Визе // Словообразование. Стилистика. Текст. Номинативные средства в текстах разных функциональных стилей. – Казань : Изд-во Казан. ун-та, 1990. – С. 33–40.

17. Виноградов В. В. О принципах определения авторства в связи с общими проблемами теории и истории литературы / В. В. Виноградов // Научная сессия, тезисы докладов и сообщений. – Л. – 1960. – 18 с.

18. Виноградов В. В. О теории поэтической речи / В. В. Виноградов // Вопросы языкознания. – М. – 1962. – № 2. – С. 3–17.

19. Виноградов В. В. Стилистика. Теория поэтической речи. Поэтика / Виктор Владимирович Виноградов. – М. : Изд-во АН СССР, 1963. – 255с.

20. Волошиновська І. Лексикографічна обробка текстових даних як засіб визначення спрямованості текстів / І. Волошиновська // Вісник Національного університету “Львівська політехніка”. Проблеми української термінології. – Львів : Вид-во нац. ун-ту “Львівська політехніка”, 2003. – № 409. – С. 147–151.

21. Волошиновська І. А. Модифікація функції розподілу Лавалетті як адаптація рангово-частотного закону Зіпфа для текстового корпусу природної мови / І. А. Волошиновська // Лінгвістичні студії : зб. наук. пр. – Донецьк : ДонНУ, 2008. – Вип. 16. – С. 334–339.

22. Волошиновська І. А. Аналіз просторової моделі текстового корпусу як метод формування тематичних підрозділів та розпізнавання авторської ідеї у колективних роботах / І. А. Волошиновська // Науковий вісник Волинського національного університету імені Лесі Українки. Серія : Філологічні науки. – Луцьк : Редакційно-видавничий відділ “Вежа” ВНУ імені Лесі Українки, 2008. – № 5. – С. 375–379.

23. Волошиновська І. А. Розділення тематичних напрямків та виявлення спорідненості спеціалізованих наукових праць / І. А. Волошиновська // Лінгвістичні студії : зб. наук. пр. – Донецьк : ДонНУ, 2008. – Вип. 17. – С. 282–287.

24. Волошиновська І. А. Особливості авторизації вузько-спеціалізованих наукових праць / І. А. Волошиновська // Нова філологія : зб. наук. пр. – Запоріжжя : ЗНУ, 2009. – № 35. – С. 36–43.

25. Волошиновська І. А. Ефективність авторської та тематичної атрибуції текстів науково-технічного спрямування / І. А. Волошиновська // Лінгвістичні студії : зб. наук. пр. – Донецьк : ДонНУ, 2011. – № 23. – С. 242–247.

26. Волошиновська І. Авторська атрибуція англо-, німецько- та україномовних художніх текстів / І. Волошиновська // XI Міжнародна міждисциплінарна конференція студентів, аспірантів та молодих вчених “Шевченківська весна: 2013”. – Київ, 2013. – С. 23–27.

27. Волошиновская И. Сравнительный анализ авторской атрибуции художественных текстов (по материалам английского, немецкого и украинского языков) / И. Волошиновская // Научная дискуссия: вопросы филологии, искусствоведения и культурологии: материалы IX международной заочной научно-практической конференции. (05 марта 2013 г.) – Москва: Изд. "Международный центр науки и образования", 2013. – С. 104 – 109.

28. Герд А. С. Специальный текст как предмет прикладного языкознания / А. С. Герд // Прикладное языкознание / Отв. ред. А. С. Герд. – СПб. : Изд-во С.-Петербург. ун-та, 1996. – С. 68–90.

29. Глинський В. В. Статистический анализ / В. В. Глинський, В. Г. Ионин. – М. : Филин, 1998. – 257 с.

30. Гніздечко О. М. Авторизація наукового дискурсу: комунікативно-прагматичний аспект (на матеріалі англомовних статей сучасних європейських та американських лінгвістів) : автореф. дис. на здобуття наук. ступеня канд. філол. наук : спец. 10.02.04 "Германські мови" / О. М. Гніздечко. – К., 2005. – 20 с.

31. Головач Ю. Лис Микита і мережі мови / Ю. Головач, В. Пальчиков // Журнал фізичних досліджень. – 2007. – Т. 11. – №1. – С. 22–33.

32. Головин Б. Н. Язык и статистика / Борис Николаевич Головин – М. : Просвещение, 1970. – 190 с.

33. Гринбаум О. Н. Компютерные аспекты стилеметрии. Стилеметрия / О. Н. Гринбаум // Прикладное языкознание / Отв. ред. А. С. Герд. – СПб. : Изд-во С.-Петербург. ун-та, 1996. – С. 451–465.

34. Гудков В. N-граммы в лингвистике / Гудков В., Гудкова Е. // Вестник Челябинского государственного университета. Филология. Искусствоведение. – Челябинск, 2011. – Вып. 57. – С. 69–71.

35. Дарчук Н. П. Индивидуальное и общее в лексической системе авторского стиля (на материале современной украинской художественной

прозы) : автореф. дис. на соискание степени канд. филол. наук : спец. 10.02.21 “Прикладная лингвистика”/ Н. П. Дарчук. – Киев, 1975. – 20 с.

36. Дорош А. К. Теорія ймовірностей та математична статистика / А. К. Дорош, О. П. Коханівський. – К. : НТУУ “КПІ”, 2006. – 268 с.

37. Електронний ресурс: Режим доступу http://fantlab.ru/article374

38. Електронний ресурс: Режим доступу www.jgaap.com

39. Електронний ресурс: Режим доступу http://www.philocomp.net/humanities/signature

40. Електронний ресурс: Режим доступу http://www.rusf.ru/books/analysis/index.htm

41. Електронний ресурс: Режим доступу http://www.sciencedirect.com/

42. Електронний ресурс: Режим доступу http://www.smalt.karelia.ru/

43. Електронний ресурс: Режим доступу http://www.textology.ru/web.htm

44. Електронний ресурс: Режим доступу http://www.vaal.ru

45. Енквист Н. Е. Параметры контекста / Н. Е. Енквист // Новое в зарубежной лингвистике. Лингвостилистика. – М. : Прогресс, 1980. – Вып. IX. – С. 254–270.

46. Ермоленко Г. В. Лингвистическая статистика (краткий очерк и библиографический указатель) / Георгий Владимирович Ермоленко. – Алма-Ата : Казахский государствений университет имени С. М. Кирова, 1970. – 156 с.

47. Ермоленко Г. В. Анонимные произведения и их авторы. На материале русских текстов второй половины XIX – начала XX в. / Георгий Владимирович Ермоленко. – Минск : Издательство “Университетское”, 1988. – 119 с.

48. Ефимов А. И. Стилистика художественной речи / Александр Иванович Ефимов. – М. : Изд-во МГУ, 1961. – 298 с.

49. Єріна А. М. Статистичне моделювання та прогнозування: Навч. посібник / Антоніна Михайлівна Єріна. – К. : КНЕУ, 2001. – 170 с.

50. Загнітко А. П. Лінгвістика тексту : Теорія і практикум / Анатолій Панасович Загнітко. – Донецьк : ТОВ “Юго-Восток, Лтд”, 2007. – 313 с.

51. Звегинцев В. А. Предложение и его отношение к языку и речи / Владимир Андреевич Звегинцев. – М. : Эдиториал УРСС, 2001. – 306 с.

52. Зубов А. В. Информационные технологии в лингвистике / А. В. Зубов, И. И. Зубова. – М. : Изд. центр “Академия”, 2004. – 208 с.

53. Ільченко О. М. Англійська мова як lingua franca наукової дискурсивної спільноти / О. М. Ільченко // Проблеми семантики, слова, речення та тексту. – К. : КЛДУ, 2000. – Вип. 3. – С. 88–91.

54. Иванова-Маркова Л. П. Сопоставительный анализ употребления системы односоставных и двусоставных предложений в русской и украинской поетической речи / Л. П. Иванова-Маркова // Вопросы статистической стилистики / Отв. ред. Б. Н. Головин. – К. : Наукова думка, 1974. – С. 243–251.

55. Караулов Ю. Н. Лингвистическое конструирование и тезаурус литературного языка / Юрий Николаевич Караулов. – М. : Наука, 1981. – 367 с.

56. Карпіловська Є. А. Машинні версії традиційних словників як основа для укладання компютерних словників та тезаурусів / Є. А. Карпіловська // Мовознавство. – 1996. – № 4 – 5. – С. 19–22.

57. Карпіловська Є. А. Вступ до комп’ютерної лінгвістики / Євгенія Анатоліївна Карпіловська. – Донецьк : ТОВ “Юго-Восток, Лтд”, 2006. – 188 с.

58. Клейн Л. С. Анатомия “Илиады” / Лев Самуилович Клейн. – СПб. : Изд-во С.-Петерб. ун-та, 1998. – 560 с.

59. Коваленко А. М. Заголовок англомовного журнального мікротексту-повідомлення : структура, семантика, прагматика (на матеріалі тижневика Newsweek) : автореф. дис. на здобуття наук. ступеня канд. філол. наук : спец. 10.02.04 “Германські мови” / А. М. Коваленко. – Київ, 2002. – 19 с.

60. Коваль А. П. Структура научного текста (научный стиль современного украинского языка) : автореф. дис. на соискание степени доктора филол. наук : спец. № 661 / А. П. Коваль – К., 1970. – 40 с.

61. Коваль А. П. Науковий стиль сучасної української літературної мови. Структура наукового тексту. / Алла Петрівна Коваль. – К. : Видавництво київського університету, 1970. – 307 с.

62. Колегаєва И. М. Текст как единица научной и художественной коммуникации : [монография] / Ирина Михайловна Колегаєва. – Одесса : Редакционно-издательський отдел областного управления по печати ОГУ имени И. И. Мечникова, 1991. – 121 с.

63. Коновалова Т. И. Функционирование сложносочиненных предложений в художественной речи : автореф. дис. на соискание степени канд. филол. наук : спец. 10.02.02 “Языки народов Российской Федерации” / Т. И. Коновалова – К., 1983. –20 с.

64. Крамер Г. Математические методы статистики / Гаральд Крамер. – М.: Мир, 1976. – 648 с.

65. Краус І. Квантитативні характеристики функціональних та індивідуальних стилів / І. Краус // Мовознавство. – К. : Наукова думка, 1967. – № 6. – С. 32–36.

66. Критская В. И. Знаки препинания как текстообразующие единицы (автоматизация анализа и редактирования в научном тексте) / В. И. Критская. – автореф. дис. на соискание степени канд.филол. наук : спец. 10.02.21 “Прикладная лингвистика” / В. И. Критская. – Киев, 1991. – 20с.

67. Кукушкина О. В. Определение авторства текста с использованием буквенной и граматической информации / О. В. Кукушкина, А. А. Поликарпов, Д. В. Хмелёв // Проблемы передачи информации. – 2001. – Т. 37. – Вып. 3. – С. 96–109.

68. Кэррол Дж. Факторный анализ стилевых характеристик прозы / Дж. Кєррол // Семиотика и искусствометрия. – М. : Мир, 1972. – С. 183–197.

69. Левицкий В. В. Квантитативные методы в лингвистике / Виктор Васильевич Левицький. – Вінниця : Нова Книга, 2007. – 264 с.

70. Лесохин М. М. Введение в математическую лингвистику Лингвистическое приложение основ математики / М. М. Лесохин, К. Ф. Лукьяненков, Р. Г. Пиотровский. – Минск : Наука и техника, 1982. – 263 с.

71. Мальцева Г. Ф. Некоторые количественные приемы описания индивидуального авторского стиля / Г. Ф. Мальцева // Статистика текста / Под ред. Р. Г. Пиотровского. – Минск. – 1969. – Т. 1. – 206 с.

72. Марков А. А. Об одном применении статистического метода [Електронний ресурс] / А. А. Марков // Известия Имп. Акад. наук. – 1915. – Т. 10. – Сер. VI. – № 4. –С. 239–242. – Режим доступу до статті http:/www.textology.ru/library/book.aspx?bookId=8&textId=2

73. Мартыненко Г. Я. Классификационные задачи стилеметрии / Г. Я. Мартиненко, С. В. Чебанов // Уч. зап. Тартус. ун-та. – Вып. 827: Квантитативная лингвистика и автоматический анализ текстов. – Тарту, 1988. – С. 119–136.

74. Мартыненко Г. Я. Основы стилеметрии / Григорий Яковлевич Мартыненко. – Л. : Изд. Ленинградского ун-та, 1988. – 173 с.

75. Мартыненко Г. Я. Стилеметрия / Г. Я. Мартиненко, С. В. Чебанов // Прикладное языкознание / Отв. ред. А. С. Герд. – Спб. : Изд-во С.-Петербург. ун-та, 1996. – С. 420–434.

76. Марусенко М. А. Атрибуция анонимных и псевдонимных литературных прозведений методами распознавания образов / Михаил Александрович Марусенко. – Л. : Изд. Ленинградского университета, 1990. – 168 с.

77. Марусенко М. А. Атрибуция анонимных и псевдоанонимных текстов методами прикладной лингвистики / М. А. Марусенко // Прикладное языкознание / Отв. ред. А. С. Герд. – СПб. : Изд-во С.-Петербург. ун-та, 1996. – С. 466–479.

78. Маслов В. П. О законе Ципфа и ранговых распределениях в лингвистике и семиотике / В. П. Маслов, Т. В. Маслова // Математические заметки. – 2006. – Т. 80. – № 5. – С. 718–732.

79. Медведев А. Р. К вопросу о микрокомпозиции научно-технических текстов / А. Р. Медведев // Стиль научной речи / Отв. ред. Е. С. Троянская. – М. : Наука, 1978. – С. 95–106.

80. Методи структурного дослідження мови / Відп. ред. В. С. Перебийніс, Л. О. Кадомцева. – К. : Наукова думка, 1968. – 187 с.

81. Митрофанова О. Д. Научный стиль речи : проблеми обучения / Ольга Даниловна Митрофанов. – М. : Рус. язык, 1985. – 231 с.

82. Муравицька М. П. Статистичні лінгвістичні дослідження та їх розвиток в українському мовознавстві / М. П. Муравицька // Мовознавство. – К. : Наукова думка, 1967. – № 5 – С.47–58.

83. Орлов Ю. К. Модель частотной структуры лексики / Ю. К. Орлов // Исследования в области вычислительной лингвистики и лингвостатистики. – М. : МГУ, 1978. – С. 59–118.

84. От Нестора до Фонвизина. Новые методы определения авторства / Под ред. Л. В. Милова. – Магадан, 2009. – 448 с.

85. Перебейнос В. И. Методы и уровни моделирования нулевого стиля / В. И. Перебейнос // Вопросы статистической стилистики / Отв. ред. Б. Н. Головин. – К. : Наукова думка, 1974. – С. 16–35.

86. Перебийніс В. І. Статистичні методи для лінгвістів / Валентина Ісидорівна Перебийніс. – Вінниця : Нова Книга, 2002. – 168 с.

87. Перебейнос В. С. Широкомасштабные лингвостатистические исследования в Украине / В. С. Перебейнос // Проблеми квантитативної лінгвістики. – Чернівці : Рута, 2005. – С. 89–99.

88. Пещак М. М. Наличие-отсутствие составных элементов целого – одна из стилистических черт произведения / М. М. Пещак // Вопросы статистической стилистики / Отв. ред. Б. Н. Головин. – К. : Наукова думка, 1974. – С. 217–228.

89. Пещак М. М. Стиль деловых документов XIV века : автореф. дис. на соискание степени доктора филол. наук : спец. 10.02.04 “Германские языки”/ М. М. Пещак. – К., 1980. – 51 с.

90. Пещак М. М. Нариси з комп’ютерної лінгвістики / Марія Михайлівна Пещак. – Ужгород : Закарпаття, 1999. – 199 с.

91. Пещак М. М. Атрибуція / М. М. Пещак // Українська мова : Енциклопедія, 2000. – С. 36.

92. Пиотровский Р. Г. Математическая лингвистика / Р. Г. Пиотровский, К. Б. Бектаев, А. А. Пиотровская. – М. : Высшая школа, 1977. – 383 с.

93. Радзієвська Т. В. Текст як засіб комунікації / Тетяна Вадимівна Радзієвська. – К. : Ін-т укр. мови, 1993. – 194 с.

94. Разинкина Н. М. Стилистика английской научной речи / Нина Марковна Разинкина. – М. : Наука, 1972. – 168 с.

95. Разинкина Н. М. Развитие языка английской научной литературы (лингвостилистическое исследование) / Нина Марковна Разинкина. – М. : Наука, 1978. – 212 с.

96. Разинкина Н. М. Функциональная стилистика английского языка / Нина Марковна Разинкина. – М. : Высш. шк., 1989. – 182 с.

97. Рогов А. А. Автоматизированная система обработки и анализа литературных текстов “СМАЛІТ” / А. А. Рогов, Ю. В. Сидоров, А. В. Король // Труды и материалы II-го Международного конгресса исследователей русского языка “Русский язык: исторические судьбы и современность” (18-21 марта). – М. : МГУ, 2004. – С. 485–486.

98. Рогов А. А. Программный комплекс “СМАЛІТ” / А. А. Рогов, Г. Б. Гурин, А. А. Котов, Ю. В. Сидоров, Т. Г. Суровцова // Труды 10-й Всероссийской научной конференции “Электронные библиотеки : перспективные методы и технологии, электронные коллекции” (Дубна, Россия, 7–11 октября). – 2008. – С.155–160.

99. Романов А. С. Структура программного комплекса для исследования подходов к идентификации авторства текстов / А. С. Романов //

Доклады Томского государственного университета систем управления и радиоэлектроникию. – 2008. – № 2 (18). – Ч. 1. – С. 106–109.

100. Романов А. С. Методика идентификации автора текста на основе аппарата опорных векторов / С. А. Романов // Доклады Том. гос. ун-та систем управления и радиоэлектроники. – 2009. – № 1 (19). – Ч. 2. – С. 36–42.

101. Романов А. С. Методика и программный комплекс для идентификации автора неизвестного текста : автореф. дис. … канд. техн. наук : спец. 05.13.18. “Математическое моделирование, численные методы и комплексы программ” / А. С. Романов – Томск, 2010. – 23 с.

102. Севбо И. П. Графическое представление синтаксических структур и стилистическая диагностика / Ирина Платоновна Севбо. – К. : Наукова думка, 1981. – 372 с.

103. Сенкевич М. П. Стилистика научной речи и литературное редактирование научных произведений / Майя Петровна Сенкевич. – М. : Высш. шк., 1976. – 263 с.

104. Слепак Б. Я. Некоторые теоретико-методологические предпосылки качественно-количественной концепции стиля / Б. Я. Слепак // Вопросы сопостовительной и прикладной лингвистики / Уч. зап. Тартус. гос. ун-та. – Тарту. – 1982. – Вып. 619. – С. 107–117.

105. Соловьёв В. И. Заглавие реферата, его роль и функциональные свойства / В. И. Соловьёв // Научно-техническая информация. –1971. – Сер 1. – № 7. – С. 20–23.

106. Статистичні параметри стилів / Відп. ред. В. С. Перебийніс. – К. : Наукова думка, 1967. – 260 с.

107. Струве П. Б. Кто первый указал на применение статистики к филологическим исследованиям / П. Б. Струве // Известия Российской АН. – 1918. –Сер. 6 – № 12. – С. 1317–1318.

108. Суровцова Т. Г. Многомерный количественный анализ и классификация текстов на основе лингвостатистических характеристик : диссертация ... канд.. техн. наук : 05.13.18 “Математическое моделирование,

численные методы и комплексы программ” / Суровцова Т. Г. – Петрозаводск, 2008. –134 с.

109. Сушко С. О. Частоти повторюваності букв і біграм у відкритих текстах українською мовою [Електронний ресурс] / С. О. Сушко, Л. Я. Фомичова, Є. С. Барсуков // Захист інформації. – 2010. – Том 3 (48). – С. 94–102. – Режим доступу до статті http://www.nbuv.gov.ua/portal/natural/Zi/2010_3/15.pdf

110. Тетерина Т. С. Становление научного стиля английского языка (опит статистического описания) : автореф. дис. … канд. филол. наук : спец. 10.02.04 “Германские языки” / Т. С. Тетерина. – М., 1973. – 23 с.

111. Тулдава Ю. А. Частотная структура текста и закон Ципфа / Ю. А. Тулдава // Уч. зап. Тартус. ун-та. – Вып. 711: Квантитативная лингвистика и автоматический анализ текстов. – Тарту, 1985. – С. 93–116.

112. Тулдава Ю. А. О частотном спектре лексики текста / Ю. А. Тулдава // Уч. зап. Тартус. ун-та. – Вып. 745: Квантитативная лингвистика и автоматический анализ текстов. – Тарту, 1986. – С. 139–162.

113. Тулдава Ю. Проблемы и методы квантитативно-системного исследования лексики / Юхан Тулдава. – Таллин : Валгус, Тартуский государственный Университет, 1987. – 204 с.

114. Тураєва З. Я. Лингвистика текста : (Текст: структура и семантика) / Зинаида Яковлевна Тураєва. – М. : Просвещение, 1986. – 127 с.

115. Федосюк М. Ю. Синтаксические особенности научно-технических рефератов и формул изобретений : автореф. дис. … канд. филол. наук : спец. / М. Ю. Федосюк. – М., 1977. – 20 с.

116. Хетсо Г. Стиль и норма / Г. Хетсо // Лингвистика текста и стилистика. Учен. зап. Тартуского гос. ун-та. – 1981. – Вып. 585. – С. 48-61.

117. Хмелёв Д. В. Распознавание автора текста с использованием цепей А. А. Маркова [Електронний ресурс] / Д. В. Хмелёв // Вестник МГУ. Сер. 9 Филология. – 2000. – № 2. – С. 115–126. – Режим доступу до статті: www.rusf.ru/books/analysis/vestnik2000win.htm

118. Хмелёв Д. В. Классификация и разметка текстов с использованием методов сжатия данных. Краткое введение [Електронний ресурс] / Д. В. Хмелёв. – 2003. – Режим доступу до статті: http://compression.ru/download/articles/classif/intro.html

119. Шайкевич А. Я. О статистическом словаре языка Достоевского / А. Я. Шайкевич // Русский язык в научном освещении. – 2001. – № 2. – С. 122–149.

120. Шайкевич А. Я. Статистический словарь языка Достоевского / А. Я. Шайкевич, В. М. Андрющенко, Н. А. Ребецкая. – М. : Языки славянской культуры, 2003. – 832 с.

121. Шевелев О. Г. Разработка и исследование алгоритмов сравнения стилей текстовых произведений: автореф. дис. … канд. техн. наук : спец. 05.13.18. “Математическое моделирование, численные методы и комплексы программ” / О. Г. Шевелев. – Томск, 2006. – 20 с.

122. Штокмар М. П. Анализ языка и стиля как средство атрибуции / М. П. Штокмар // Вопросы текстологии. – 1960. – Вып. 2 – С. 100–145.

123. Эсбенсен К. Анализ многомерных данных. Избранные главы / К. Эсбенсен / Пер. с англ. С.В. Кучерявского. – Черноголовка : Изд-во ИПХФ РАН, 2005. – 158с.

124. Яхонтова Т. В. Заголовок у науковому тексті : структурні, семантичні та стилістичні особливості / Т. В. Яхонтова // Вісн. Львів. Ун-ту. Сер. : Іноземні мови. – Львів, 2005. – Вип. 12. – С. 24–31.

125. Яхонтова Т. В. Лінгвістичні характеристики англомовного жанру “тези доповідей наукової міжнародної конференції” / Т. В. Яхонтова // Філологічні студії. – Луцьк : Волин. акад. дім, 2007. – № 1–2. – С. 293–298.

126. Яхонтова Т. В. Функціонально-смислові та мовні особливості анотації сучасних англомовних наукових статей / Т. В. Яхонтова // Гуманітарний вісник. Сер. : Іноземна філологія : зб. наук. пр. – Черкаси : ЧДТУ, 2007. – № 11. – С.553–557.

127. Яхонтова Т. В. Лінгвістична генологія наукової комунікації : [монографія] / Тетяна Вадимівна Яхонтова. – Львів : Видавничий центр ЛНУ імені Івана Франка, 2009. – 420 с.

128. Яхонтова Т. В. Структурно-композиційні особливості сучасної англомовної наукової статті / Т. В. Яхонтова // Мовознавство. – 2009. – № 6. – С. 51–58.

129. Abbasi A. Applying Authorship Analysis to Extremist-group Web Forum Messages / A. Abbasi, H. Chen // IEEE Intelligent Systems. – 2005. – Vol. 20(5). – P. 67–75.

130. Abbasi A. Writeprints: a Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace / A. Abbasi, H. Chen // ACM Transactions on information systems. – 2008. – Vol. 26(2). – P. 7:1–7:29.

131. Al-Harbi S. Automatic Arabic Text Classification / S. Al-Harbi, A. Almuhareb, A. Al-Thubaity, M. S. Khorsheed, A. Al-Rajeh // 9es Journées internationales d'Analyse statistique des Données Textuelles, 2008. – P. 77-83.

132. Argamon S. Style Mining of Electronic Messages for Multiple Authorship Discrimination: First Results / S. Argamon, M. Saric, S. Stein // Proceedings of the 9th International Conference on Knowledge Discovery and Data Mining (Washington DC, USA, August 24 – 27). – 2003. – P. 475–480.

133. Argamon S. Measuring the Usefulness of Function Words for Authorship Attribution [Електронний ресурс] / S. Argamon, S. Levitan. – 2005. – Режим доступу до статті : http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.71.6935&rep=rep1&type=pdf.

134. Argamon S. Interpriting Burrows's Delta: Geometric and Probabilistic Foundations / S. Argamon // Literary and Linguistic Computing. – 2008. – Vol. 23(2). – P. 131–147.

135. Baayen H. Outside the cave of shadows: using syntactic annotation to enhance authorship attribution / H. Baayen, H van Halteren, F. Tweedie // Literary and Linguistic Computing. – 1996. – Vol. 11 (3). – P. 121–132.

136. Baayen H. An Experiment in Authorship Attribution / H. Baayen, H. van Haltern, A. Neijt, F. Tweedie // Proceedings of 6es Journ´ees internationales d'Analyse statistique des Donn´ees Textuelles (Saint-Malo, France, March 13 – 15). – 2002. – P. 29–37.

137. Basile C. An Example of Mathematical Authorship Attribution / Chiara Basile, Dario Benedetto, Emanuele Caglioti, Mirko Degli Esposti // Journal of Mathematical Physics. – 2008. – Vol. 49 (12). – P. 125211–125230.

138. Bellegarda J. R. Exploiting both Local and Global Constraints for Multispan Statistical Language Modeling [Електронний ресурс] / J. R. Bellegarda // Proceedings of International Conference on Acoustics, Speech, and Signal Processing (Seattle, Washington, USA, May 12-15). – 1998. – Vol. 2. – P. 677–680. – Режим доступу до статті: http://www.ece.umassd.edu/Faculty/acosta/ICASSP/Icassp_1998/pdf/scan/ic981164.pdf

139. Binongo J.N.G. Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution. / J. N. G. Binongo // Chance. – New York : Springer, 2003. – Vol. 16 (2). – P. 9–17.

140. Biskub I. Applied and Computational Linguistics: [підручник англ. мовою] / Iryna Biskub. – Луцьк : РВВ "Вежа" Волин. держ. ун-ту імені Лесі Українки, 2007. – 304с.

141. Book of Abstracts (LUMDETR 2006 – VI European Conference on Luminescent Detectors and Transformers of Ionizing Radiation, June 19-23, 2006 Lviv, Ukraine). – Lviv : Liga Press, 2006. – 268 p.

142. Breiman L. Classification and Regression Trees / L. Breiman, J. Friedman, R. Olshen, C. Stone. – New York : Wadsworth Statistics, 1984. – 358 p.

143. Buk S. N. Rank-Frequency Analysis for Functional Style Corpora of Ukrainian / S. N. Buk, A. A. Rovenchak // Journal of Quantitative Linguistics.— 2004.— V. 11 (3). – P. 161–171.

144. Buk S. Word-Length-Related Parametrs of Text Genrs in Ukrainian Language. A Pilot Study / S. Buk, O. Humenchyk., L. Mal'tseva, A. Rovenchak // Text and Language: Structures - Functions - Interrelations. Quantitative perspectives. – 2010. – P. 13–19.

145. Burges C. J. C. Tutorial on Support vector machines for Patern Recognition / C. J. C. Burges // Data Mining and Knowledge Discovery. – Vol. 2 (2). – 1998. – P. 121–167.

146. Burrows J. F. Word Patterns and Story Shapes : The Statistical Analysis of Narrative Style / J. F. Burrows // Literary and Linguistic Computing. – 1987. – Vol. 2. – P. 61–70.

147. Burrows J. F. Not Unless You Ask Nicely: the Interpretative Nexus between Analysis and Information / J. F. Burrows // Literary and Linguistic Computing. – 1992. – Vol. 7. – P. 91–109.

148. Burrows J. “Delta”: a Measure of Stylistic Difference and a Gide to Likely Authorship / J. Burrows // Literary and Linguistic Computing. – 2002. – Vol. 17 (3). – P. 267–287.

149. Burrows J. Questions of Authorships: Attribution and Beyond / J. Burrows // Computers and the Humanities. – 2003. – Vol. 37 (1). – P. 5–32.

150. Calix K. Stylometry for E-mail Author Identification and Authentication [Електронний ресурс] / K. Calix, M. Connors, D. Levy, H. Manzar, G. McCabe, and S. Westcott // Proceedings of CSIS Research Day (Pace University, USA, May 2). – 2008.– Режим доступу до статті : http://csis.pace.edu/~ctappert/srd2008/c2.pdf:

151. Cavnar W. B. N-Gram-Based Text Categorization / W .B. Cavnar, J. M. Trenkle // Proceedings of 3-rd Annual Symposium on Document Analysis and Information Retrieval. – Las Vegas, 1994. – P. 161–175.

152. Clement R. Ngram and Bayesian Classification of Documents for Topic and Authorship / R. Clement, D. Sharp // Literary and Linguistic Computing. – 2003. – Vol 18 (4). – P. 423–447.

153. Diederich J. Authorship attribution with support vector machines / J. Diederich, J. Kindermann, E. Leopold, G. Paass // Applied Intelligence. – 2003. – Vol. 19 (1-2). – P. 109–123.

154. Dumais S. T. Inductive Learning Algorithms and Reprezentainions for Text Categorization / S. T. Dumais , J. Platt, D. Heckerman, M. Sahami // Procedings of International Conference on Information and Knowledge Management (Bethesda, Maryland, USA, November 3-7). – 1998 – P. 148–155.

155. Estoup J. B. Gammes Sténographique. Recueil de Textes Choisis pour L'acquisition Mythodique de la Vitesse [Електронний ресурс] / J. B. Estoup. – Paris : Institut stenographique, 1912. – 142 p. – Режим доступу до статті : http://torvald.aksis.uib.no/ corpora/2002-2/0070.html.

156. Farringdon J. M. Analysing for Authorship: A Guide to the Cusum Technique / J. M. Farringdon, A. Q. Morton, M. G. Farringdon. – Cardiff : University of Wales Press, 1996. – 324 p.

157. Ferrer i Cancho R. Two Regimes in the Frequency of Words and the Origins of Complex Lexicons : Zipf's Law Revisited / R. Ferrer i Cancho, R. V. Sole // Journal of Quantitative Linguistics. – 2001. –Vol. 8 (3). – P. 165–173.

158. Ferrer i Cancho R. Can simple models explain Zipf's law for all exponents? / R. Ferrer i Cancho, V. D. P. Servedio // Glotometrics. – 2005. – Vol. 11. – P. 1–8.

159. Frank E. Text categorization using compression models / E. Frank, C. Chui, I. H. Witten // Proceedings of Data Compresion Conference (Snowbird, Utah, March) – Los Alamitos : IEEE Press, 2000. – P. 200–209.

160. Fung G. The Disputed Federalist Papers: SVM Feature Selection Via Concave Minimization [Електронний ресурс] / G. Fung // In TAPIA'03 : Proceedings of the conference on Diversity in computing. – New York : ACM Press, 2003. – P. 42–46. − Режим доступу до статті : http://pages.cs.wisc.edu/~ gfung/federalist.pdf.

161. Glasman-Deal H. Science Research Writing for Non-native Speakers of English / Hilary Glasman-Deal. – Imperial College Press, 2010. – 257 p.

162. Glover A. Detecting Stylistic Inconsistencies in Collaborative Writing / A. Glover, G. Hirst // Writers at Work : Professional Writing in the Computerized Environment. – L. : Springer-Verlag, 1995. – P. 147–168.

163. Grieve J. Quantitative authorship attribution : An evaluation of techniques / J.Grieve // Literary and Linguistic Computing. – 2007. – Vol 22 (3). – P. 251–270.

164. Grzybek P. Quantitative Text Typology: The Impact of Word Length / P. Grzybek, E. Stadlober, E. Kelih, G. Antić // Classification. The Ubiquitous Challenge. –Heidelberg : Springer, 2005. – P. 53–64.

165. Hastie T. The Elements of Statistical Learning. Data Mining, Inference, and Prediction / T. Hastie, R. Tibshirani, J. Friedman // Springer Series in statistics. – New York : Springer, 2008. – 745 p.

166. Hayes J. Authorship Attribution: a Principal Component and Linear Discriminant Analysis of the Consistent Programmer Hypothesis / J. Hayes// International journal of computers and their applications. – 2008. – Vol. 15 (2) – P. 79–99.

167. Holmes D. A Stylometric Analysis of Mormon Scripture and Related Texts / D. Holmes // Journal of the Roal Statistical Society. Series A (Statistics in Society). – 1992. – Vol. 155 (1). –P. 91–120.

168. Holmes D. Authorship Attribution / D. Holmes // Computers and the Humanities. – 1994. – Vol. 28 (2). – P. 87–106.

169. Holmes D. I. The ‘Federalist’ Revisited: New Directions in Authorship Attribution / D. I. Holmes, R. S. Forsyth // Literary and Linguistic Computing. – 1995. – Vol. 10. – P. 111–127.

170. Holmes D. The Evolution of Stylometry in Humanities Scholarship / D. Holmes // Literary and Linguistic Computing. – 1998. – Vol. 13 (3). – P. 111–117.

171. Holmes D. I. Stephen Crane and the New-York Tribune: A Case Study in Traditional and Non-Traditional Authorship Attribution / D. I. Holmes, M. Robertson, R. Paez // Computers and the Humanities. – 2001. – Vol. 35. – P. 315–331.

172. Holmes D. I. Stylometry and the Civil War: The case of the Pickett Letters / D. I. Holmes // Chance. – New York : Springer, 2003. – Vol. 16 (2). – P. 18–26.

173. Hoorn J. Neural Network Indentifacation of Poets using Letter Sequences / J. Hoorn, S. Frank, W. Kovalczyk, F. Van der Ham // Literary and Linguistic Computing. – 1999. – Vol. 14 (3). – P. 311-338.

174. Hoover D. L. Statistical Stylistics and Authorship Attribution: an Empirical Investigation / D. L. Hoover // Literary and Linguistic Computing. – 2001. – Vol. 16 (4). – P. 421–444.

175. Hoover D. L. Frequent Word Sequence and Statistical Stylistics / D. L. Hoover // Literary and Linguistic Computing. – 2002. –Vol. 17 (2). – P. 157–179.

176. Hoover D. L. Multivariate Analysis and the Study of Style Variation / D. L. Hoover // Literary and Linguistic Computing. – 2003. – Vol. 18 (4). – P. 341–360.

177. Hoover D.L. Testing Burrow's Delta / D.L. Hoover // Literary and Linguistic Computing. – 2004. – Vol. 19 (4). – P. 453–475.

178. Hotelling H. Analysis of a Complex of Statistical Variables into Principal Components / H. Hotelling // Journal of Educational Psychology. – 1933. – Vol. 24. – P. 417–441, 498–520.

179. Jackson J. E. Principal Components and Factor Analysis: Part 1-Principal Components / J. E. Jackson // Journal of Quality Technology. – 1981. – Vol. 13. – P. 201–213.

180. Jackson P. Text Retrieval, Extraction & Categorization, Natural Language Processing for Online Applications / P. Jackson, I. Moulinier. – Amsterdam : John Benjamins Publishing Co., 2002. – 226 p.

181. Jockers M. A Comparative Study of Machine Learning Methods for Authorship Attribution / Matthew L. Jockers, Daniela M. Witten // Literary and Linguist Computing. – 2010. – Vol. 25 (2). – P. 215–223.

182. Jolliffe I. Principal Component Analysis / Ian Jolliffe // Springer Series in Statistics. – New York : Springer, 2002. – 487 p.

183. Juola P. The Time Course of Language Change / P. Juola // Computers an the Humanities. – 2003. – Vol. 37 (1). – P. 77–96.

184. Juola P. A Prototype for Authorship Attribution Studies / P. Juola, J. Sofko, P. Brennan // Literary and Linguistic Computing. – 2006. – Vol. 21 (2). – P. 169–178.

185. Juola P. Authorship Attribution / P. Juola //Foundations and Trends in Information Retrieval. – 2006. – Vol. 1 (3). – P. 233–334.

186. Kelih E. Classification of Author and / or Genre? The Impact of Word Length / E. Kelih, G. Antic, P. Grzybek, E. Stadlober // Classification. The Ubiquitous Challenge. – Heidelberg: Springer, 2005. – P. 498–505.

187. Kelih E. Graphemhäufigkeiten in Slawischen Sprachen: Stetige Modelle / E. Kelih // Glottometrics. – 2009. – Vol. 18. – P. 52–68.

188. Kešelj V. N-gram-based Author Profiles for Authorship Attribution / V. Kešelj, F. C. Peng, N. Cercone, C. Thomas [Електронний ресурс] // Proceedings of Pacific Association for Computational Linguistics. − Halifax, Nova Scotia, Canada, 2003. – P. 256–264. − Режим доступу до статті : http://web.cs.dal.ca/~vlado/papers/ pacling03.pdf.

189. Khmelev D. V. Disputed Authorship Resolution through Using Relative Empirical Entropy for Markov Chains of Letters in Human Language Text / D. V. Khmelev // Journal of Quantitative Linguistics. – 2000. – Vol. 7 (3). – P. 201–207.

190. Khmelev D. V. Using Markov Chains for Identification of Writers / D. V. Khmelev, F. J. Tweedie // Literary and Linguistic Computing. – 2001. – Vol. 16 (3). – P. 299–307.

191. Kjell B. Authorship Determination using Letter Pair Frequency Features with Neural Network Classifiers / B. Kjell // Literary and Linguistic Computing. – 1994. – Vol. 9 (2). – P. 119–124.

192. Kjell B. Authorship Determination of Text Samples using Neural Networks and Bayesian Classfiers / B. Kjell // Proceedings of the IEEE

International Conference on Systems, Man and Cybernetics (San Antonio, TX, 2-5 October . – 1994. – Vol. 2. — P. 299–307.

193. Koppel M. Exploiting Stylistic Idiosyncrasies for Authorship Attribution [Електронний ресурс] / M. Koppel, J. Schler // Proceedings of the IJCAI: Workshop on Computational Approaches to Style Analysis and Synthesis (Acapulco, Mexico, August 9-15). – 2003. – P. 69–72. − Режим доступу до статті : http://www.cs.biu.ac.il/ ~koppel/papers/ijcai-idios.

194. Koppel M. Measuring Differentiability: Unmasking Pseudonymous Authors / M. Koppel, J. Schler, E. Bonchek-Dokow // Journal of machine learning research. – 2007. – Vol. 8 (1). – P. 1261–1276.

195. Krsul I. H. Authorship Analysis and Identifying the Author of a Program / I. Krsul, H. Spafford // Computers and Security. – 1997. – Vol. 16 (3). – P. 233–257.

196. Kullback S. On Information and Sufficiency / S. Kullback, R. A. Leibler // Annals of Math. Stat. – 1951. – Vol. 22. – P. 79–86.

197. Kullback S. The Kullback-Leibler distance / S. Kullback // The American Statistician. – 1987. – Vol. 41. – P. 340–341.

198. Kuperman V. Productivity in the Internet Mailing List: A Bibliometric analysis / V. Kuperman // Journal of the American Society for Information Science and Technology. – 2006. – Vol. 57 (1). – P. 51–59.

199. Lafferty J.D. Grammatical Trigrams: A Probabilistic Model of Link Grammar [Електронний ресурс] / J. D. Lafferty, D. Sleator, D. Temperly // Proceedings of AAAI Fall Symp. Probabilistic Approaches to Natural Language, Cambridge, 1992. − P. 74-81. − Режим доступу до статті : http://www.link.cs.cmu.edu/

200. Lavalette D. Facteur d’impact: Impartialite ou Impuissance? [Електронний ресурс] / D. Lavalette // Internal Report, INSERM U350, Institut Curie – Recherche, 1996. – Режим доступу: http:/www.curie.u-psud.fr/U350.

201. Le H. Q. Extension of Zipf's Law to Word and Character N-grams for English and Chinese / H. Q. Le, E. I. Sicilia-Garcia, J.I. Ming, F.J. Smith //

Computational Linguistics and Chinese Language Processing. – 2003. – Vol. 8 (1). – P. 77–102.

202. Le H. Q. Zipf and Type-Token Rules for the English and Irish Languages [Електронний ресурс] / H. Q. Le , Francis J Smith // MIDL, Paris, 29-30 novembre 2004. – P.65–70 − Режим доступу до статті : http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.134.4095

203. Li Y.H. Classification of Text Documents / Y. H. Li, A. K. Jain // The Computer Journal – 1998. – Vol. 41. – № 8. – P. 537–546.

204. Mahajan M. Improved Topic-dependent Language Modeling using Information Retrieval Techniques / M. Mahajan, D. Befferman, X. D. Huang // Proceedings of International Conference on Acoustics, Speech and Signal Processing (Arizona, USA, 15-19 March). – 1999. – Vol. 1 – P. 541–544.

205. Malyutov M. Authorship Attribution of Texts: a Review / M. Malyutov // Electronic Notes in Discrete Mathematics. – 2005. – Vol. 21. – P. 353–357.

206. Mandelbrot B. An Informational Theory of the Statistical Structure of Language / B. Mandelbrot // Communication theory / Ed. Willis Jackson. – L. : Butterworths, 1953. − P. 486−502.

207. Manning C. D. Foundations of Statistical Language Modelling / C. D. Manning, H. Schütze. – Camridge, Massachusetts : The MIT Press, 1999. – 680 p.

208. Martin S. C. Adaptive Topic-Dependent Language Modeling Using Word-Based Varigrams / S. C. Martin // Proceedings of the 5-th European conference on speech communication and technology (Rhodes, Greece, September 22-25). – 1997. – Vol. 3.−P. 1447−1450.

209. Matthews R. Neural Computation in Stylometry I: An Application to the Works of Shakespeare and Fletcher / Robert A. J. Matthews, Thomas V. N. Merriam // Literary and Linguistic Computing. – 1993. – Vol. 8 (4). – P. 203–209.

210. Mendenhall T. C. Characteristic Curves of Composition / T. C. Mendenhall // The Popular Science Monthly. – 1904. – Vol. LXV (19). – P. 373–377.

211. Merriam T. Neural Computation in Stylometry II: An Application to the Works of Shakespeare and Marlowe / Thomas V. N. Merriam, Robert A. J. Matthews // Literary and Linguistic Computing. – 1994. – Vol. 9 (1). – P. 1–6.

212. Mikros G. Investigating Topic Influence in Authorship Attribution [Електронний ресурс] / G. Mikros, E. Argiri // Proceedings of the SIGIR Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (Amsterdam, Netherlands, July 27). – 2007. − P. 29–35. Режим доступу до статті : http://sunsite.informatik.rwth-aachen.de/Publicati.

213. Montemurro Marcelo A. Beyond the Zipf-Mandelbrot Law in Quantitative Linguistics / Marcelo A. Montemurro // Physica A. – 2001. – Vol. 300 (3–4). – P. 567–578.

214. Mosteller F. Inference in An Authorship Problem / F. Mosteller, D. L. Wallace // Journal of the American Statistical Association. – 1964. – Vol. 58. – P. 275–309.

215. Nagao Makoto A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese / Makoto Nagao, Shinsuke Mori // Proceedings of the 15th International Conference on Computational Linguistics (Kyoto, Japan, August 5-9). –1994. – Vol. 1. – P. 611–615.

216. Nemeth G. Multilingual Statistical Text Analysis, Zipf's Law and Hungarian Speech Generation / Nemeth Geza, Zainko Csaba // Acta Linguistica Hungarica. – 2002. –Vol. 49 (3-4). – P. 385–405.

217. Nwogu K. The Medical Research Paper: Structure and Functions / K. Nwogu // English for Specific Purposes. – 1997. – Vol. 16 (2). – P. 119–138.

218. Olsson J. Forensic Linguistics / John Olsson – L. : Continuum International Publishing Group, 2008. – 256 p.

219. Patton J. M. Stylometric Analysis of Yashar Kemal's "Ince Memed Tetralogy" / J. M. Patton, F. A. Can // Computers and the Humanities. – 2004. – Vol. 38. – P.457–467.

220. Pearson K. On Lines and Planes of Closest Fit to Systems of Points in Space / K. Pearson // Philosophical Magazine. – 1901. – Vol. 2. – P. 559–572.

221. Peng R. Quantitative Analysis of Literary Styles / R. Peng, N. Hengartner // The American Statistician. – 2002. – Vol. 56 (3). – P. 175–185.

222. Peng F. Language Independent Authorship Attribution Using Character Level Language Models / F. Peng, D. Schuurmans, V. Keselj, S. Wang // Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (Budapest, Hungary, April 12–17). – 2003. – Режим доступу до статті : http://portal.acm.org/citation.cfm?id=1067843.

223. Popescu I. I. On the Lavalette Ranking Law / I. I. Popescu, M. Ganciu, M. Penache, D. Penache // Romanian Reports in Physics. – 1997. – Vol. 49. – P. 3–27.

224. Popescu Ioan-Iovitz. On a Zipf's law Extension to Impact Factors / Ioan-Iovitz Popescu // Glottometrics. – 2003. – Vol. 6. – P. 83–93.

225. Popescu I.-I. Zipf's law – another view / I.-I. Popescu, G. Altmann, R. Köhler // Quality and Quantity. – 2010. – Vol. 44. – P. 713–731.

226. Posner I. R. How People Write Together / I. R. Posner, R. M. Baecker // Proceedings of the twentyfifth annual Hawaii International Conference on System Sciences (Hawaii, January 7-10). – 1992 – Vol. 4. – P. 127–138.

227. Physical Review B [Електронний ресурс] / The American Physical Society. – 2004. – Vol. 69. – Issue (14-16). – P. 140301 – 161101. – Режим доступу до статті : http://prb.aps.org/.

228. Radiation Measurements / Proceedings of the 6th European Conference on Luminescent Detectors and Transformers of Ionizing Radiation (LUMDETR 2006). – 2007. – Vol. 42 (4-5). – P. 509–944.

229. Rijsbergen C. J. van Information Retrieval [Електронний ресурс] / C. J. van Rijsbergen. – L. : Buttersworth. 1979. – 208 p. – Режим доступу до статті : http://www.dcs.gla.ac.uk/Keith/Preface. html.

230. Rudman J. The State of Authorship Attribution Studies: Some Problems and Solutions / J. Rudman // Computers and the Humanities. – 1998. – Vol. 31. – P. 351–365.

231. Rudman J. Non-traditional Authorship Attribution Studies in the *Historia Augusta*: Some Cavets/ J. Rudman // Literary and Linguistic Computing. – 1998. – Vol. 13 (3). – P. 151–157.

232. Rudman J. Authorship Attribution: Statistical and Computational Methods / J. Rudman // Encyclopedia of Language and Linguuistics. – Elsevir Ltd, 2006. – P. 611–617.

233. Salton G. Automatic Information Organization and Retrieval / Gerard Salton. – New York : McGraw-Hill, 1968.

234. Sanderson C. On Authorship Attribution via Markov Chains and Sequence Kernels / C. Sanderson, S. Guenter // Proceedings of the 18th International Conference on Pattern Recognition (Hong Kong, 20-24 August). – 2006. – P. 437–440.

235. Sebastiani F. Machine Learning in Automated Text Categorization / F. Sebastiani // ACM Computing Surveys. – 2002. – Vol. 34 (1). – P. 1–47.

236. Shannon C. E. Prediction and Entropy of Printed English / C. E. Shannon // Bell System Technical Journal. – 1951. – Vol. 30. –P. 50–64.

237. Sichel H. S. Word Frequency Distribution and Type-token Characteristics / H. S. Sichel // Mathematical Scientist – 1986. – Vol. 11. – P. 45–72.

238. Simkin M.V. Re-inventing Willis [Електронний ресурс] / M. V. Simkin, V. P. Roychowdhury. – 2006. – Режим доступу до статті: http://arxiv. org/abs/physics/0601192.

239. Simon Herbert A. On a Class of Skew Distribution Functions / Herbert A. Simon // Biometrika. – 1955. – Vol. 42 (3-4). – P. 425–440.

240. Skern T. Writing Scientific English. A workbook. / Tim Skern. – Wien : Facultas Verlags- und Buchhandels AG, 2009. – 192 p.

241. Stamatatos E. Computer-based Authorship Attribution without Lexical Measures / E. Stamatatos, N. Fakotakis, G. Kokkinakis // Computers and the Humanities. – 2001. – Vol. 35 (2). – P. 193–214.

242. Swales J. M. Research Genres: Exploration and Applications / John M. Swales – Cambridge: Cambridge Univ. Press, 2004. – 316 p.

243. Sylvester J. J. On the Reduction of a Bilinear Quantic of the n^{th} Order to the Form of a Sum of n Products by a Double Orthogonal Substitution / J. J. Sylvester // Messenger of Mathematics. – 1889. – Vol. 19. – P. 42–46. – Режим доступу : http://pca.narod.ru/Sylvester1889.pdf

244. Tearle M. An algorithm for Automated Authorship Attribution using Neural Networks / M. Tearle, K. Taylor, H. Demuth // Literary and Linguistic Computing. – 2008. – Vol. 23 (4). – P. 425–442.

245. The Journal of Applied Linguistics. Oxford University Press. – 1998. – Vol 19 (1). – 155 p.

246. Tweedie F.J. Neural network applications in stylometry: the Federalist papers / F. J. Tweedie, S. Singh, D. I. Holmes // Computers and the Humanities. – 1996. – Vol. 30. – P. 1–10.

247. Vapnik V. N. The Nature of Statistical Learning Theory / Vladimir Naumovich Vapnik. – Berlin : Springer Verlag, 2000. – 314 p.

248. Vel de O. Mining E-mail Content for Author Identification Forensics / O. de Vel., A. Anderson, M. Corney, G. Mohay // SIGMOD Record. – 2001. – Vol. 30 (4). – P. 55–64.

249. Venkatesh J. Handwritten Tamil Character Recognition using SVM / J. Venkatesh, C. Sureshkumar // International Journal of Computer and Network Security. – 2009. – Vol. 1 (3). – P. 29–33.

250. Voloshynovska I. Peculiarity of N-Gram Model Application in the Author Style Recognition / I. Voloshynovska // Proceedings of the III Intrnational Conference on Computer Science and Information Technologies (Lviv, September 25–27). – 2008. – P. 77-79.

251. Voloshynovska I. Employment of N-gram Analysis in Relative Entropy Model for Authorship Identification within Scientific Text Base / I. Voloshynovska // Proceedings of the international conference Intelectual Systems for decision making and problems of computational intelligence. –Kherson : KNTU, 2010. – Vol. 2. – P. 305–306.

252. Voloshynovska I. A. Characteristic Features of Rank-Probability Word Distribution in Scientific and Belletristic Literature / I. A. Voloshynovska // Journal of Quantitative Linguistics. – 2011 – Vol. 18 (3). – P. 274–289.

253. Wyllys Ronald E. The Measurements of Jargon Standardization in Scientific Writing using Rank-frequency Zipf's Curves / Ronald E. Wyllys – PhD. thesis, University of Wisconsin-Madison, 1974. – 107 p.

254. Yang Y. An Evaluation of Statistical Approches to Text Categorization / Y. Yang // Information Retrieval. – 1999. –Vol. 1 (1/2). – P. 69–90.

255. Zanette Damian H. Zipf's Law and the Creation of Musical Context / Damian H. Zanette // Musicae Scientiae. – 2006. – Vol. 10. – P. 3–18.

256. Zhao Y. Effective and Scalable Authorship Attribution using Function Words / Y. Zhao, J. Zobel // Proceedings of the 2nd Asian Information Retrieval Symposium (Jeju Island, South Korea, October 13–15) – 2005. – P. 174-190.

257. Zhao Ying Using Relative Entropy for Authorship Attribution / Ying Zhao, Justin Zobel, Phil Vines // AIRS 2006. – Berlin Heidelberg : Springer-Verlag, 2006. – P. 92–105.

258. Zhao Ying Entropy-Based Authorship Search in Large Document Collections / Ying Zhao, Justin Zobel // Lecture Notes in Computer Science. – 2007. – Vol. 4425. – P. 381–392.

259. Zhao Y. Searching with Style: Authorship Attribution in Classic Literature / Y. Zhao, J. Zobel // In Proceedings of the Thirtieth Australasian Computer Science Conference (Ballarat, Victoria, Australia, January 30 - February 2). – 2007. – P. 59–68.

260. Zheng R. A Framework for Authorship Analysis of Online Messages : Writing-Style Features and Classification Techniques / R. Zheng, Z. Huang, H. Chen // J. Amer. Soc. Inf. Sci. Technol. – 2006. – Vol. 57 (3) – P. 378–393.

261. Zipf G. K. Human Behavior and the Principle of Least Effort / George Kingsley Zipf. – Cambridge, Mass.: Addison-Wesley. – 1949.

262. Zorenko Yu. Single-crystalline Films of Ce-doped YAG and LuAG Phosphors: Advantages Over Bulk Crystals Analogues / Yu. Zorenko, V. Gorbenko, I. Konstankevych, A. Voloshinovskii, G. Stryganyuk, V. Mikhailin, V. Kolobanov, D. Spassky // Journal of Luminescence. – 2005. – Vol.114, Issue 2. – P. 85–94.

СПИСОК ДОВІДНИКОВИХ ДЖЕРЕЛ

263. ВТСУМ 2000 : Великий тлумачний словник сучасної української мови / Уклад. і голов. ред. В. Т. Бусел. – К. ; Ірпінь : ВТФ Перун, 2007. – 1736 с.

264. СЧС 1996 : Словник чужомовних слів / І. Бойків, О. Ізюмов, Г. Калишевський, М. Трохименко. – К. : Родовід, 1996. – 523 с.

265. СІС 1985 : Словник іншомовних слів / За ред. О. С. Мельничука. – К. : Голов. Ред. УРЕ, 1985. – 966 с.

Додаток А

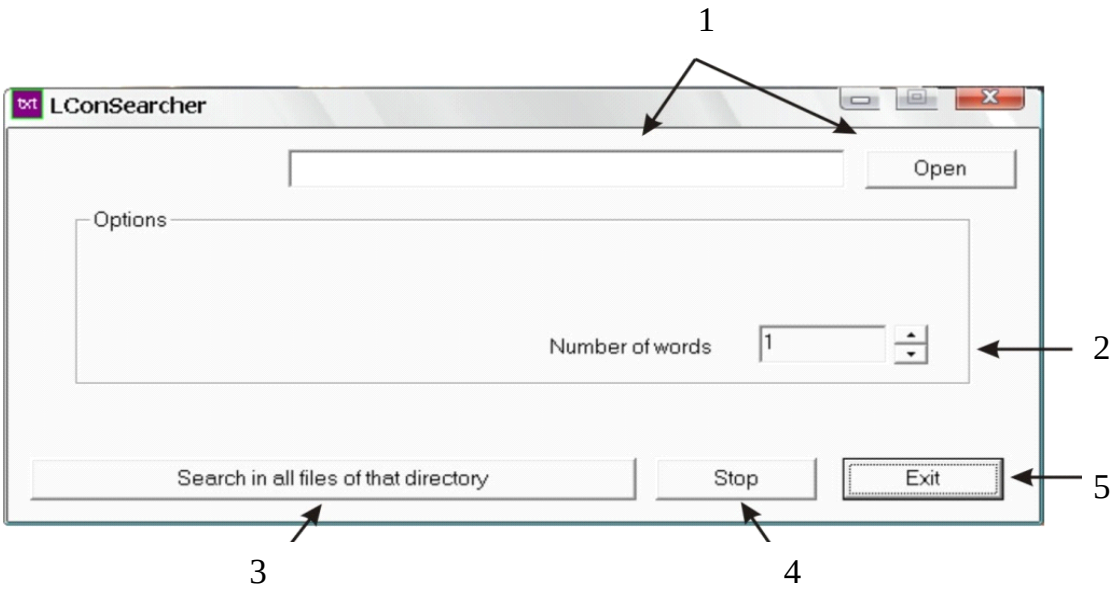


Рис. 1. Інтерфейс програми формування словника послідовності вживання одного і більше слів для вибірки текстів.

Призначення керуючих елементів інтерфейсу: 1 – поле та кнопка для вибору одного із текстів аналізованої групи; 2 – поле, в якому задають кількість слів у *n*-грамі; 3 – кнопка старту роботи програми; 4 – кнопка дострокової зупинки розрахунків; 5 – кнопка завершення роботи з програмою.

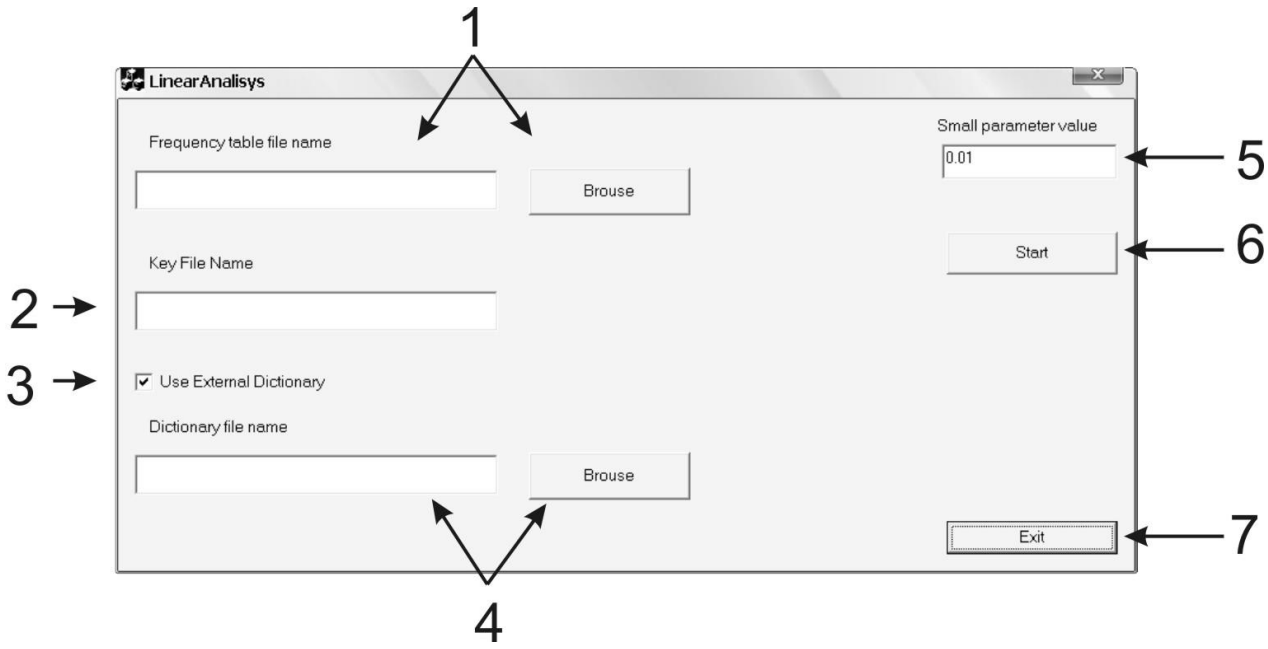


Рис. 2. Інтерфейс користувача програми “LinearAnalisys”.

Призначення керуючих елементів інтерфейсу: 1 – поле та кнопка для задання назви файлу, що містить вхідну частотну таблицю; 2 – поле для задання назви опорного тексту; 3 – елемент управління режимом вибору ключових слів (якщо поле відмічене позначкою “V” – використовують зовнішній словник ключових слів, у протилежному випадку – ключовими вважаються слова, які зустрічаються в опорному тексті); 4 – поле та кнопка для задання назви текстового файлу, який містить зовнішній словник ключових слів (ці елементи активні, коли відмічено поле 3); 5 – поле для задання малого параметра μ; 6 – кнопка старту обчислень; 7 – кнопка завершення роботи з програмою.

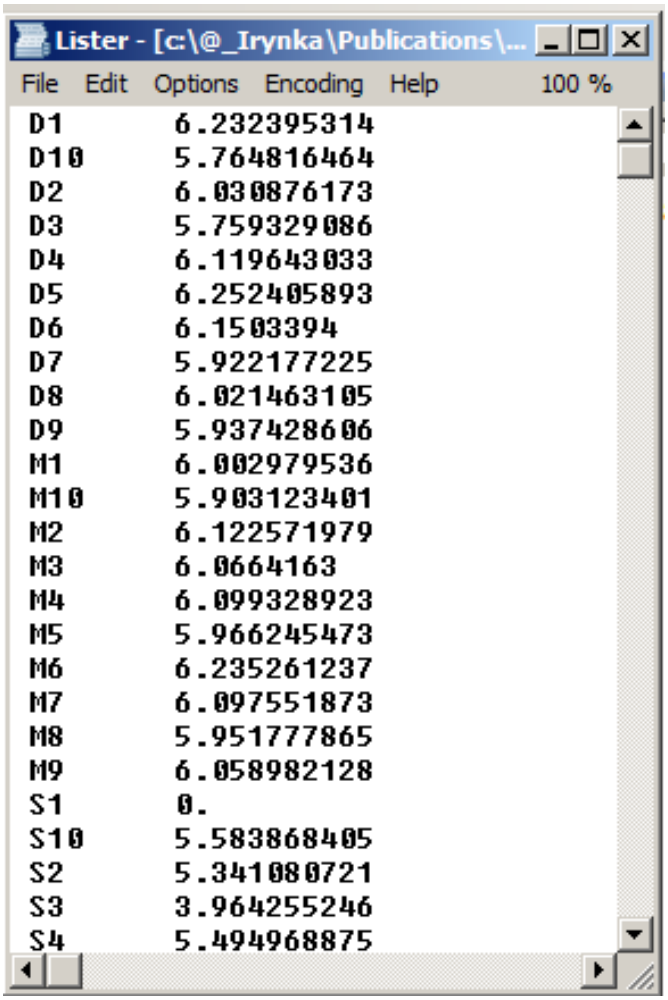


Рис. 3. Фрагмент текстового файлу з результатами розрахунків програми “LinearAnalisys”.

У першій колонці наведено умовні позначення текстів (D1, D10, ... , S1, ...). У другій колонці – дивергенцію для аналізованого тексту та опорного тексту. Дивергенція дорівнює 0 для тексту S1. Це означає, що текст S1 був вибраний як опорний.

Додаток Б

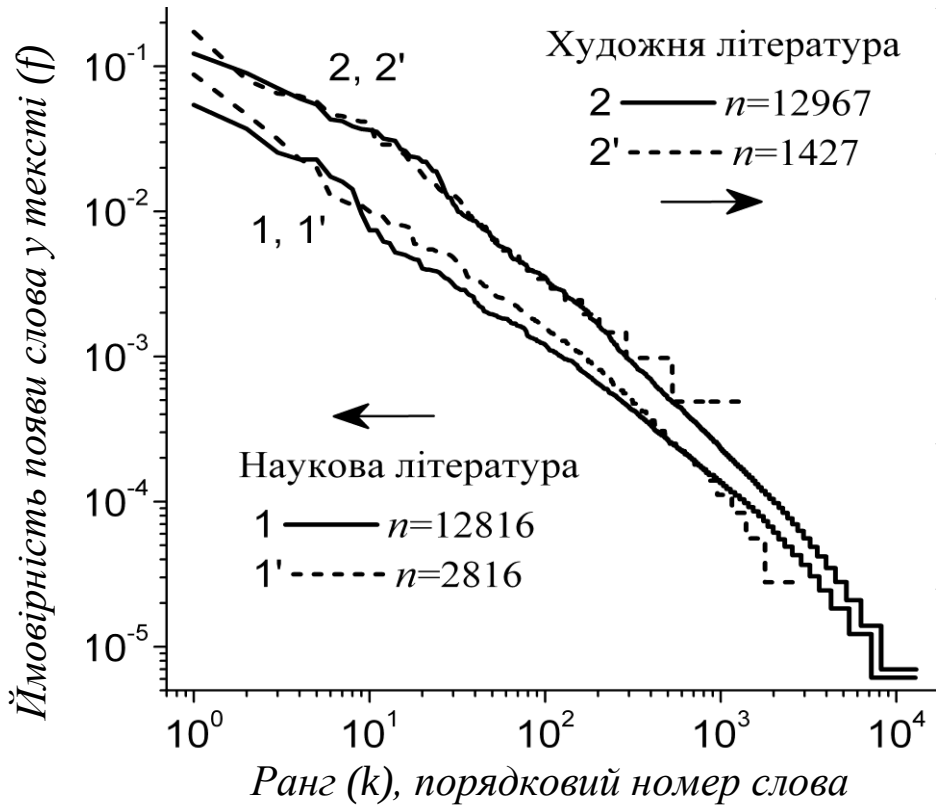


Рис. 1. Залежність рангового розподілу слів від стилю та обсягу текста.

Крива 1: “Crystal Design: Structure and Function” (суцільна крива); крива 1’: Thesis of M. True (штрихова крива); крива 2: “Anna Karenina” (суцільна крива); крива 2’: “The hunting of the Snark” (штрихова крива).

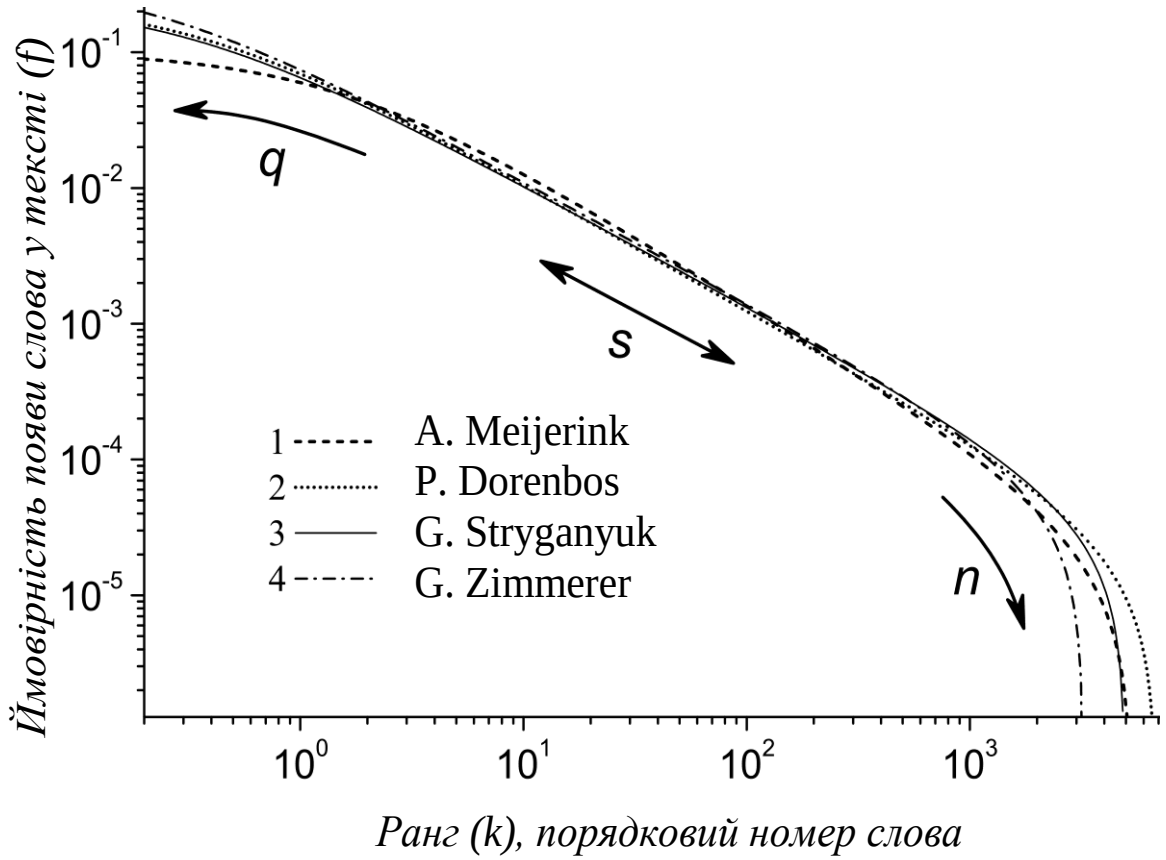


Рис. 2. Апроксимація модифікованою формулою Лавалетті рангово-частотного розподілу послідовності з 1 слова (1-грам) наукових текстів різних авторів.

Додаток В

Тематичні секції VI-ої Міжнародної конференції LUMDETR 2006

Назва тематичної секції		к-ть статей	словоформ загалом	різних словоформ
A	Сцинтиляційні матеріали	14	37469	4379
B	Дозиметричні матеріали	14	33272	4309
C	Запасаючі та інші фосфори	10	28194	3819
D	Наноматеріали та нанотехнології	10	22851	3648
E	Теорія та моделювання	7	16572	2890
F	Люмінесцентні механізми	12	31744	4212
G	Домішки, дефекти, пастки	16	42130	5063
H	SUPERLUMI експеримент	8	22021	3099
I	Технологія люмінесцентних матеріалів	5	12790	2681

Список статей використаних для тематичної атрибуції наукових текстів

PROCEEDINGS OF THE 6-TH EUROPEAN CONFERENCE ON LUMINESCENT DETECTORS AND TRANSFORMERS OF IONIZING RADIATION (LUMDETR 2006)

№ ***A Scintillation materials***

A 1 Laguta V. Luminescence and decay of excitons in lead tungstate crystals / V Laguta, M. Nikl, S Zazubovich [et al.]

A 2 **Single crystalline film scintillators based on Ce- and Pr-doped aluminium garnets** / Y Zorenko, V. Gorbenko, E. Mihokova [et al.]

A 3 **Peculiarities of luminescence and scintillation properties of YAP:Ce and LuAP:Ce single crystals and single crystalline films** / Y. Zorenko, V. Gorbenko, I. Konstankevych [et al.]

A 4 Mares J.A., Beitlerova A., Nikl M. Scintillation and optical properties of

YAG:Ce films grown by liquid phase epitaxy.

A 5 Effect of Eu^{2+} concentration on afterglow suppression in CsI:Tl, Eu / L.A. Kappers, R.H. Bartram, D.S. Hamilton [et al.]

A 6 Multiple doping of Cs:TI crystals and its effect on aftergo / E.E. Ovechkina, V. Gaysinskiy, S.R. Miller [et al.]

A 7 Itoh M. Photo-stimulated luminescence and photo-induced infrared absorption in $ZnWO_4$ / M. Itoh, T. Katagiri, Y. Tani, M. Fujita

A 8 Scintillation properites of ceramics based on zinc oxide / Demidenko V.A., Gorokhova E.I., Khodyuk [et al.]

A 9 Influence of thallium doping on scintillation characteristicsof mixed KDP/ADP crystals / Voronov A.P., Vyday Yu.T., Salo V.I. [et al.]

A 10 Influence of thermal treatment and γ–radiation on absorption, luminescence and scintillation properties of $Lu_3Al_5O_{12}$:Ce single crysalline films / Zorenko Yu., Pavlyk B., Turchak R., Zorenko T. [et al.]

A 11 Defect clusters of variable composition as an origin of coloration of oxide crystals under thermal treatment and irradiation / Burachas S., Ippolitov M., Manko V., Lomonosov V. [et al.]

A 12 Stilbene crystalline powder in polymer base as a new fast neutron detector / Budakovsky S.V., Galunov N.Z., Grinyov B.V., Karavaeva [et al.]

A 13 Bordun O. Luminescence of thin films of bismuth and lead complex oxide compounds / O. Bordun, I. Kukharsky, V. Antonyuk, V. Dmytruk [et al.]

A 14 The inertia properties of Cs_2LiYCl_6(Ce) scintillation crystals / L. Trefilova, V. Cherginets, A. Gektin, B. Grinyov [et al.]

B Dosimetric Materials

B 15 Kortov V. Materials for thermoluminescent dosimetry: actual status and future trends

B16 High-dose characterization of different LiF phosphors / P. Bilski, P. Olko, M. Puchalska, B. Obryk [et al.]

B17 Dotzler C. Dosimetric properties of $RbCdF_3:Mn^{2+}$ / C. Dotzler, G.V.M. Williams, A. Edgar, G.A. Appleby[et al.]

B18 Nonlinear dose dependence in TLD-500 detectors resulting from interactive interaction of traps / V.S. Kortov., I.I. Milman, S.V. Nikiforov, E.V. Moiseykin [et al.]

B19 Magnetic manifestations of thermoluminescence excitation in CaF_2:Mn (TLD-400) / M. Danilkin, A. Kirillov, S. Klimonsky, V. Kuznetsov [et al.]

B20 Dual-step annealing for the stability of glow curve structure and the TL sensitivity of the newly developed LiF:Mg,Cu,Si / J.I. Lee, J.L. Kim, J.S. Yang, A.S. Pradhan [et al.]

B21 Puchalska M. Thermoluminescence glow peak parameters for LiF:Mg, Ti with modified activator concentration / M. Puchalska, P. Bilski, P. Olko

B22 Ptaszkiewicz M. Long term fading of LiF:Mg, Cu, P and LiF:Mg,Ti thermoluminescence detectors with standard and modified activator composition / M. Ptaszkiewicz

B23 Kopec R. A model for distinguishing between static and dynamic exposure of personal thermoluminescence dosemeters / R. Kopec, M. Budzanowski, B. Obryk, P. Olko

B24 Azorin-Vega J.C. Thermoluminescence properties of TiO_2 nanopowder / J.C. Azorin-Vega, J. Azorin-Nieto, M. Garcia-Hipolito, T. Rivera-montalvo

B25 Sommer M. New aspects of a BeO-based optically stimulated luminescence dosimeter / M. Sommer, R. Freudenberg, J. Henniger

B26 Synthesis and physical characteristics of radiophotoluminescent glass dosimeters / Shih-Ming Hsu, David YC Huang, Hsi-Wen Yang [et al.]

B27 Zhydachevskii Ya. Optically stimulated luminescence of $YalO_3:Mn^{2+}$ for radiation dosimetry / Ya. Zhydachevskii, A. Suchocki, M. Berkowski, Ya. Zakharko

B28 Rebisz M The response of thermally stimulated luminescence in CVD diamonds to heavy charged particles / M. Rebisz, B. Voss

C Storage and Other Phosphors

C29 Schweizer S. Fluorozirconate-based glass ceramic x-ray detectors for digital radiography / S. Schweizer, J.A. Johnson

C30 The role of segregations and oxygen doping in the photostimulation mechanism of csBr:Eu^{2+} / S. Hessen, J. Zimmermann, H. Von Seggern [et al.]

C31 Persistent luminescence and synchrotron radiation study of the $Ca_2MgSi_2O_7$:Eu^{2+},R^{3+} materials / T. Aitasalo, J. Hölsä, M. Kirm [et al.]

C32 **Energy transfer to Ce^{3+} ions in $Tb_3Al_5O_{12}$:Ce single crystalline films** / Yu. Zorenko, V. Voznyak, V. Vistovsky [et al.]

C33 **High-pressure luminescence spectroscopy of $EuAl_2O_4$ phosphor** / Yu. Zorenko, V. Gorbenko, M. Grinberg [et al.]

C34 Correlation of the dielectric properties and the PSL-sensitivity in CsBr:Eu image plates / G. Schierning, M. Batenschuk, F. Meister [et al.]

C35 Luminescence spectroscopy of Eu2+ in CsBr:Eu needle image plates (NIPs) / M. Weidner, M.Batenschuk, F. Meister [et al.]

C36 Thermolumunescence properties of copper doped zirconium oxide for UVR dosimetry / T. Rivera, L. Olvera, A. Martinez [et al.]

C37 Photoluminescence, afterglow and thermoluminescence in $SrAl_2O_4$:Eu^{2+},Dy^{3+} irradiated with blue and UV light / V. Chernov, T.M. Piters, R. Melendrez [et al.]

C38 Zorenko Yu. Energy transfer between the Eu2+ dipole and aggregate centers in CsBr:Eu crystals / Yu. Zorenko, R. Turchak, T. Voznjak

D Nanomaterials and Nanotechnologies

D 39 Luminescent properties of nanophosphors / L.G. Jacobson, B.L. Bennett, R.E. Meunchausen [et al.]

D 40 Luminescence of cerium doped YAG nanopowders / V. Pankratov, L. Grigorjeva, D. Millers, T. Chudoba

D 41 Photoluminescence structure of highly excited quantum dots of type II / J. Krasnyj, W. Donderowisz, W. Jacak, M. Tytus

D 42 Revealing the radiation –induced effects in silicon by processing at enhanced temperatures – pressures / A. Misiuk, B. Surma, J. Bak-Misiuk [et al.]

D 43 Smytyna V. The nature of emission centers in CdS nanocrystals / V. Smytyna, V. Skobeeva, N. Malushin

D 44 Luminescence properties of Sn-based microcrystals embedded in csBr matrix / P.V. Savchyn, S.V. Myagkota, A.S. Voloshinovskii [et al.]

D 45 Formation of graded band-gap in CdZnTe by YAG:Nd laser radiation / A. Medvid', L. Fedorenko, B. Korbutjak [et al.]

D 46 Shevchuk V.N. Photo-stimulated transfer and luminescence in rare earth gallium garnet crystals / V.N. Shevchuk

D 47 Luminescence, vibrational and XANES studies of AIN nanomaterials / S. Bellucci, A.I. Popov, C. Balasubramanian [et al.]

D 48 Kavetskyy T. Charged defectes in chalcogenide vitreous semiconductors studied with combined Raman scattering and PALS methods / T. Kavetskyy, M. Vakiv, O. Shpotyuk

E Theory and Modeling

E 49 Galunov N.Z. Ionising radiatio energy exchange in the regions of high activation density of organic scintillators / N.Z. Galunov, E.V. Martynenko

E 50 Electronic structure and optical properties of ABP_2O_7 double phosphates / Yu. Hizhnyi, O. Gomenyuk, S. Nedilko [et al.]

E 51 Electronic energy band parameters of CsCl evaluated on core Bloch states and plane waves / S.V. Syrotyuk, Ya.M. Chornodolskyy, G.B. Stryganyuk [et al.]

E 52 Chruscinska A. Complex OSL signal and the trap independence assumption / A. Chruscinska

E 53 A new methods for the numerical analysis of thermoluminescence glow curve / K.S. Chung, H.S. Choe, J.I. Lee, J.L. Kim

E 54 Weinstein I.A. Evolutionary approach in the simulation of thermoluminescence kinetics / Weinstein I.A., Popko E.A.

E 55 Balabay R. Alteration on the surface of the pore walls of the porous silicon under high temperature ageing: computer simulation / R. Balabay,

E. Chernonog

F Luminescence Mechanisms

F 56 Intrinsic Luminescence in oriented BeO crystals under VUV and inner-shell excitation / V. Ivanov, M. Kirm, V. Pustovarov, A. Kruzhalov [et al.]

F 57 Ogorodnikov I.N. Luminescence of the hydrogen bonded crystals / I.N. Ogorodnikov, M. Kirm, V.A. Pustovarov

F 58 Silica luminescence induced by fast light ions / S.I. Kononenko, O.V. Kalantaryan, V.I. Muratov, V.P. Zhurenko

F 59 $Tb^{3+} \rightarrow Ce^{3+}$ energy transfer in $Y_{3-x-y}Tb_yGd_xAl_5O_{12}$ (x=0.65, y=0.575) doped with Ce^{3+}/ R. Turos-Matysiak, W. Gryk, M Grinberg [et al.]

F 60 Solarz P. Energy transfer processes in $K_5Li_2GdF_{10}$:Eu,Pr / P. Solarz, W. Ryba-Romanowski

F 61 Surdo A.I. Thermoactivated spectroscopy in dosimetric α-Al_2O_3 / A.I. Surdo

F 62 Excited states of molybdenum oxyanion in scheelite and wolframite structures / A. Kotlov, L. Jönsson, H. Kraus [et al.]

F 63 Eu^{2+} luminescence in the $EuAl_2O_4$ concentrated phosphor / F. Meister, M. Batentschuk, S. Dröscher [et al.]

F 64 Sanchez-Munoz L. Radiatio-induced self-structuring of radiative defects complexes in a K-feldspar crystal: a study by thermoluminescence / L. Sanchez-Munoz, J. Garcia-Guinea, V. Correcher, A. Delgado

F 65 Comparison of UV-IR radioluminescence and cathodoluminescence spectra of a potassium feldspar / V. Correcher, L. Sanchez-Munoz, J. Garcia-Guinea [et al.]

F 66 Radio-luminescence efficiency and rare-earth dispersion in Tb doped silica glasses / M. Fasoli, F. Moretti, A. Lauria [et al.]

F 67 Luminescence properties of Cu_6PS_5I nanosized superionic conductors / I.P. Studenyak, R.Ya. Buchuk, V.O. Stephanovich [et al.]

G Impurities, Defects, Traps

G 68 Some aspects of radiation resistance of wide-gap metal osides /

A. Lushchik, E. Feldbach, S. Galaev [et al.]

G 69 Luminescence and excitation energy transfer in new fluride crystals containing rare earth ions / W. Ryba-Romanowski, P. Solarz, M. Gusowski, G. Dominiak-Dzik

G 70 Influence of the crystal structure on the stability of Ln^{2+} in strontium borates / V.P. Dotsenko, I.V. Berezovskaya, N.P. Efryushina [et al.]

G 71 Origin of TSL peaks located at 200-250 K in UV-irradiated $PbWO_4$ crystals / P. Fabeni, A. Krasnikov, V.V. Laguta [et al.]

G 72 Energy transfer features in Eu^{2+} and Ce^{3+} doped $LiCaAlF_6$ crystals / S. Neicheva, A. Gektin, N. Shiran [et al.]

G 73 Zahedifar M. Effect of population of trapping states on kinetic parameters of LiF:Mg,Cu,P (GR-200) using mixed and general order kinetics / M. Zahedifar, M.J. Kavianinia, M. Ahmadi

G 74 The Lu- doping effect on emission and the coloration of pure and Ce- doped BaF_2 crystals / V. Nesterkina, N. Shiran, A. Gektin [et al.]

G 75 Carriers trapping and radiative recombination in Ce, Eu and Pr doped $LiLuF_4$ crystals / V.Voronova, N. Shiran, A. Gektin [et al.]

G 76 Luminescence of molecular O_2 ions in neutron-irradiated Be_2GeO_4 / L. Blagigina, A. Zatsepin, A. Kukharenko[et al.]

G 77 Thermal treatment influence on radiative recombination center stability in ZnSe(X) crystals / L. Gal'chinetskii, B. Grinyov, N. Starzinskiy [et al.]

G 78 Paramagnetic impurity defects in LuAG:Ce thick film scintillators / V.V. Laguta, A.M. Slipenyuk, M.D. Glinchuk [et al.]

G 79 The reason of the scintillation efficiency decrease of CsI(Tl) crystals exposed by the high-dosed radiation / L. Trefilova, B. Grinyov, V. Alekseev [et al.]

G 80 The recombination channels of luminescence excitation in YAG:Yb single crystalline films / Ya. M. Zakharko, A.P. Luchechko, S.B. Ubizskii [et al.]

G 81 Shevchuk V.N. Dipole effects in AWO_4 (A=Pb, Cd) luminescent crystals / V.N. Shevchuk, I.V. Kayun

G 82 Optical, infrared and electron-microscopy studies of metallic $(Cd_i)_n$ clusters in layered CdI_2 crastals / I. Bolesta, S. Veglosh, Yu. Datsiuk [et al.]

G 83 Calculation of point defects ensemble in CdTe films considering transport phenomenon in gas phase / V.V. Kosyak, M.M. Kolesnik, A.S: Opanasyuk, I.Yu. Protsenko

H SUPERLUMI Users Meeting

H 84 Zimmerer G. SUPERLUMI: a unique setup for luminescence spectroscopy with synchrotron radiation / G. Zimmerer

H 85 Makhov V.N. Luminescence excitation spectra of $LiGdF_4$ and $LiLuF_4$ in the region of interconfigurational $4f^n - 4f^{n-1}5d$ transitions in the Gd^{3+} and Lu^{3+} ions / Makhov V.N., Kirm M., Stryganyuk G.

H 86 Fast intrinsic emission in Cs_2CdI_4 single crystal / M.S. Pidzyrailo, V.V. Vistovskyy, A.S. Voloshynovskii [et al.]

H 87 Luminescent properties of Yb-doped $LaSc_3(BO_3)_4$ under VUV excitation / N. Guerassimova, I. Kamenskikh, D. Krasikov [et al.]

H 88 Luminescence and thermoluminescence of alkalinc earth metaborates / I.V. Berezovskaya, N.P. Efryushina, A.S. Voloshynovskii [et al.]

H 89 Luminescence of Bi^{3+} ions in $Y_3Al_5O_{12}$:Bi single crystalline films / Yu. Zorenko, V. Gorbenko, T. Voznyak [et al.]

H 90 Temperature dependence of the PbWO4:F,Eu luminescence / V. Kolobanov, N. Krutyak, M. Pashkovsky, D. Spassky

H 91 Specific features of luminescence of oxygen-deficient centers in nanostructured silicon dioxide / V.S. Kortov, A.F. Zatsepin, V.A. Pustovarov [et al.]

I Technologie of Luminescent materials

I 92 Fabrication of submicron-sized oxide phosphors and their spectroscopic properties / E. Zych, A. Walasek, J. Trojan-Piegza [et al.]

I 93 Glass-ceramics and epoxy-composites for radiation imaging / G.V.M. Williams, A. Bittar, C. Dotzler [et al.]

I 94 Characteristics of defect formation in aluminium oxide reinforced bioactive

glass / S. Szarska, H. Jungner, B. Staniewicz-Brudnik, M. Wiatr

I 95 Zorenko Yu**. Growth peculiarities of the $R_3Al_5O_{12}$ (R=Lu, Yb, Tb, Eu-Y) single crystalline film phosphors by Liquid Phase Epitaxy** / Yu. Zorenko., V. Gorbenko

I 96 Influence of polystyrene scintillator strip methods of production on their main characteristics / V. Senchyshyn, B. Grynyov, S. Melnychuk [et al.]

* Жирним шрифтом виділено статті написані Ю.Зоренком

Додаток Г

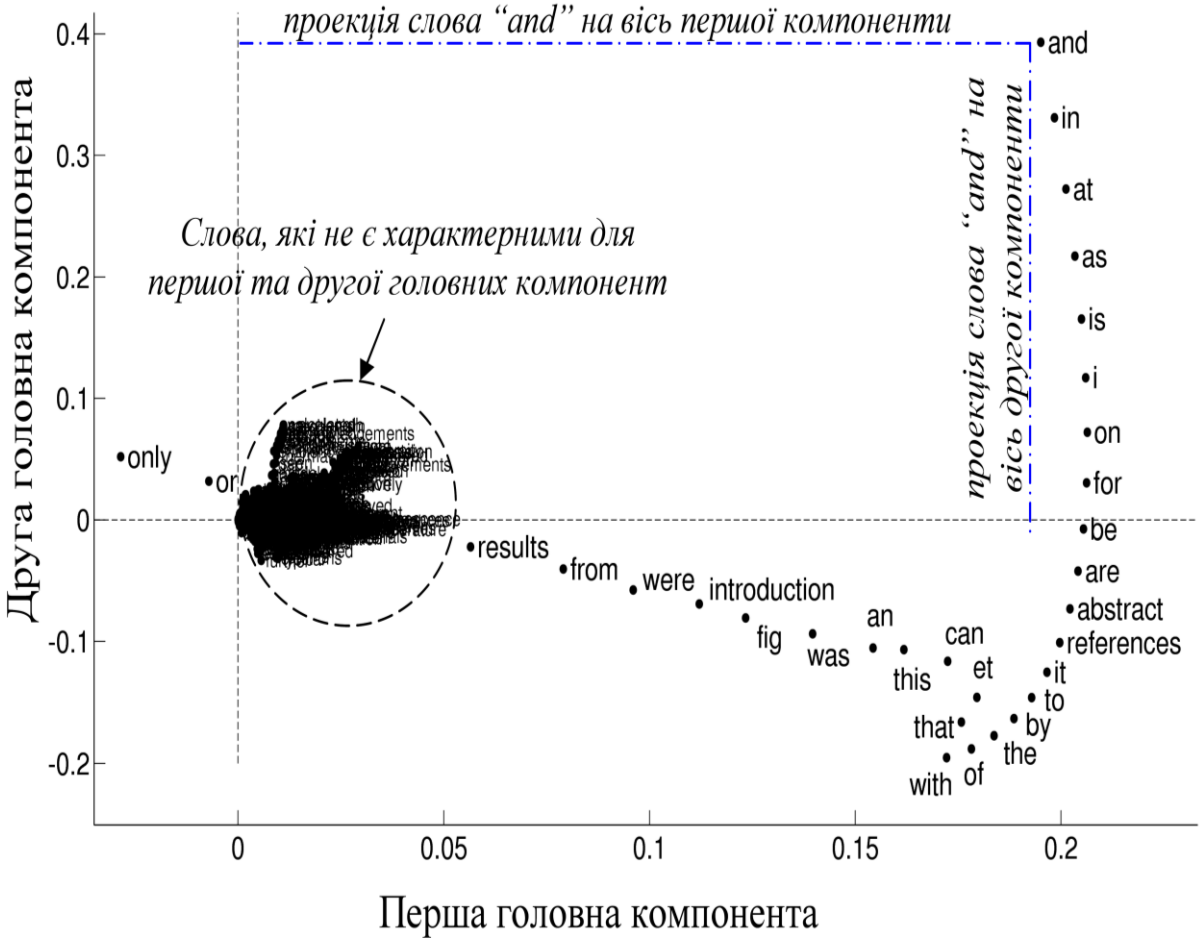


Рис. 1. Розподіл функціональних слів у просторі головних компонент.

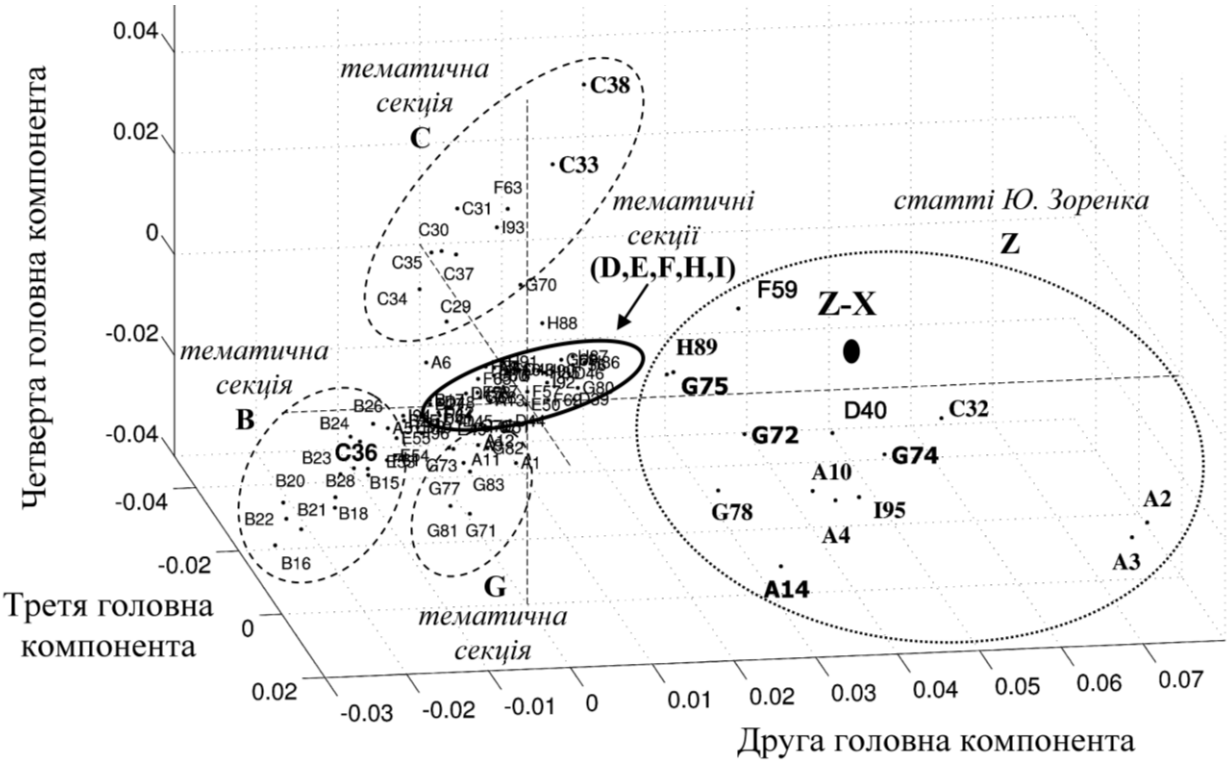


Рис. 2. Розподіл праць конференції у просторі головних компонент.

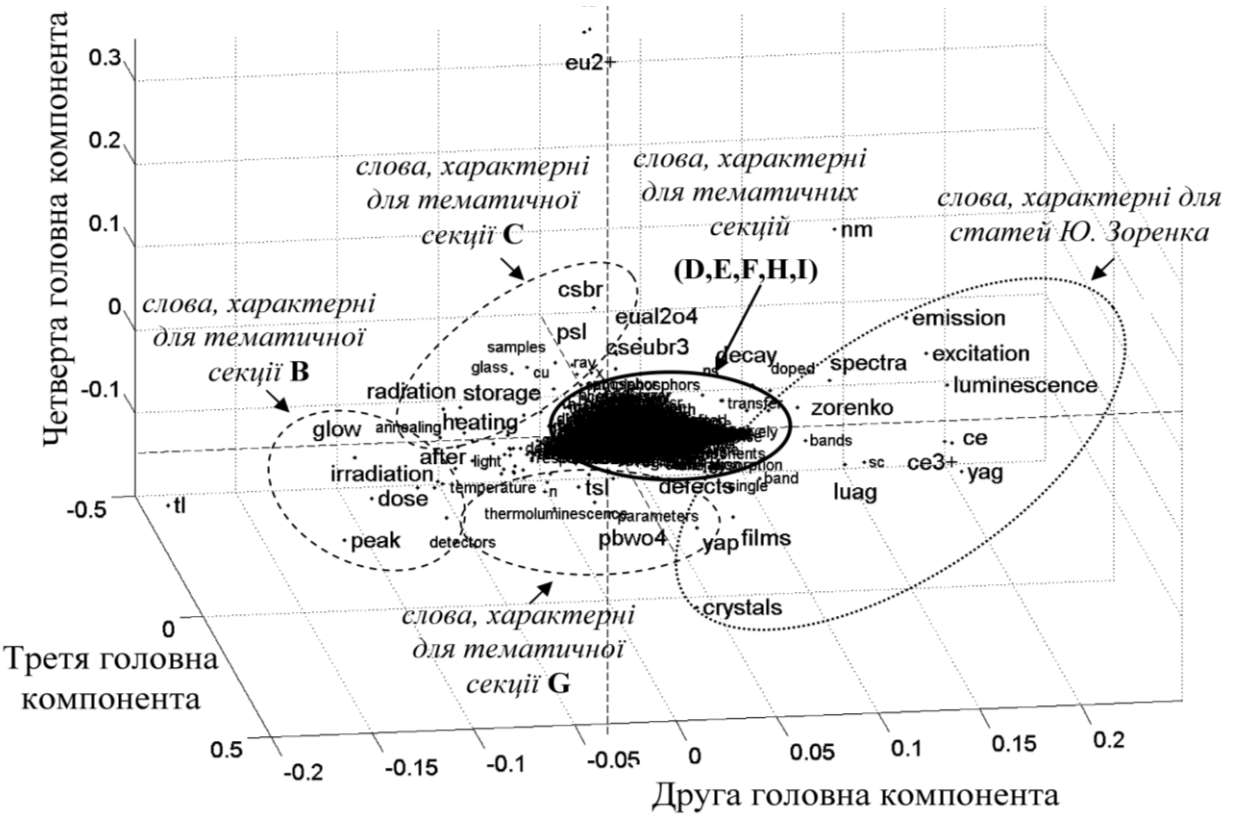


Рис. 3. Розподіл словоформ у просторі головних компонент.

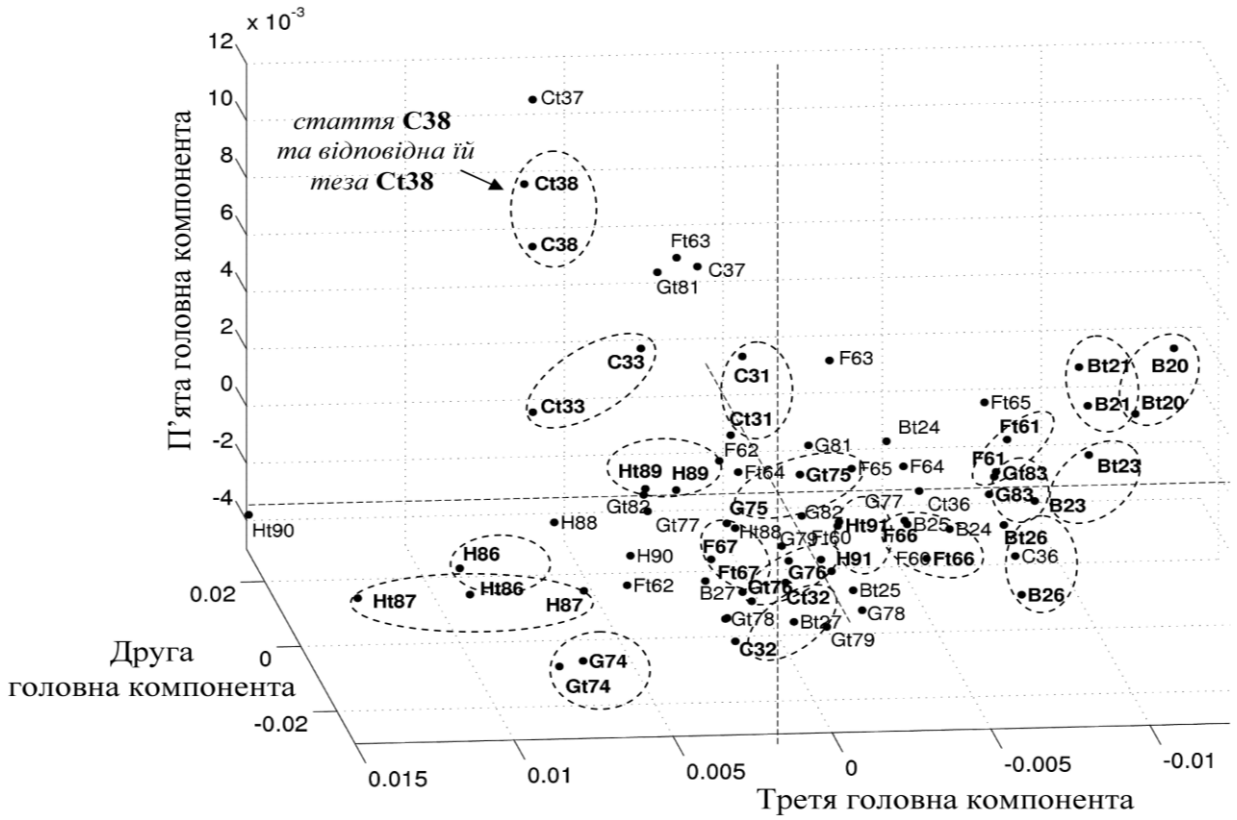


Рис. 4. Тематичне групування статей та відповідних їм тез.

Тексти статей позначені великими латинськими буквами і порядковим номером (наприклад, C33 означає: 33-я стаття збірника праць, що увійшла до групи C), а у позначеннях тез наявна мала літера “t”. Споріднені пари “теза-стаття” обведені пунктирним колом.

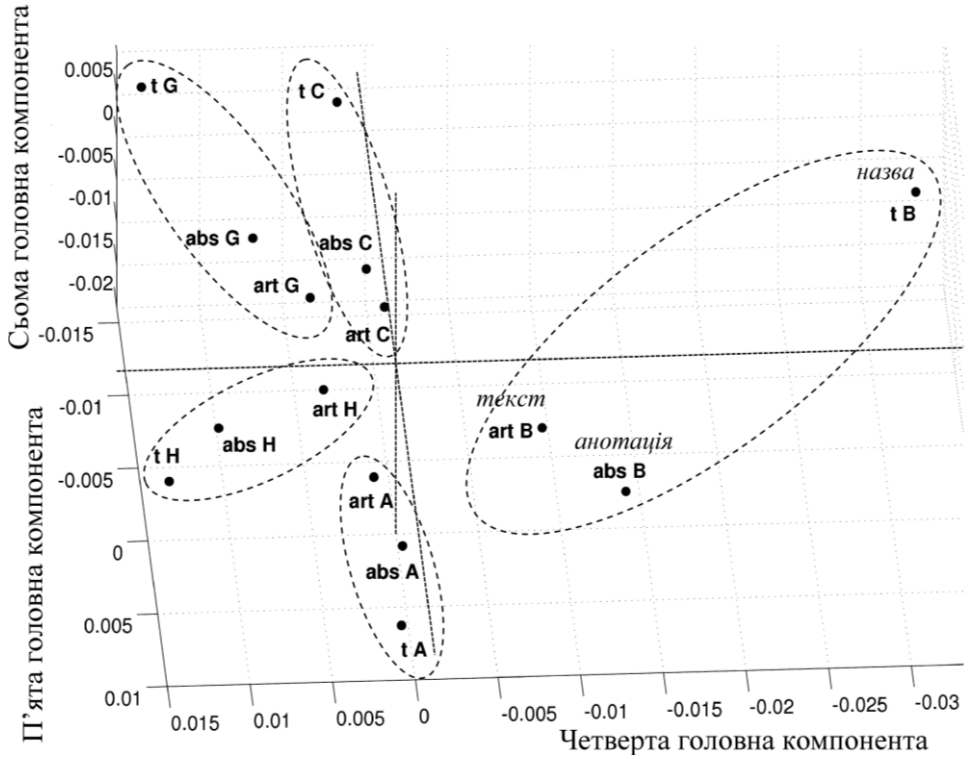


Рис. 5. Розподіл тріад “стаття–анотація–заголовок” у просторі головних компонент.

Статті (*article*) груп A, B, C, G та H позначені art A, art B, art C, art G, art H; анотації (*abstract*) – abs A, abs B, abs C, abs G, abs H; заголовки (*title*) – t A, t B, t C, t G, t H.

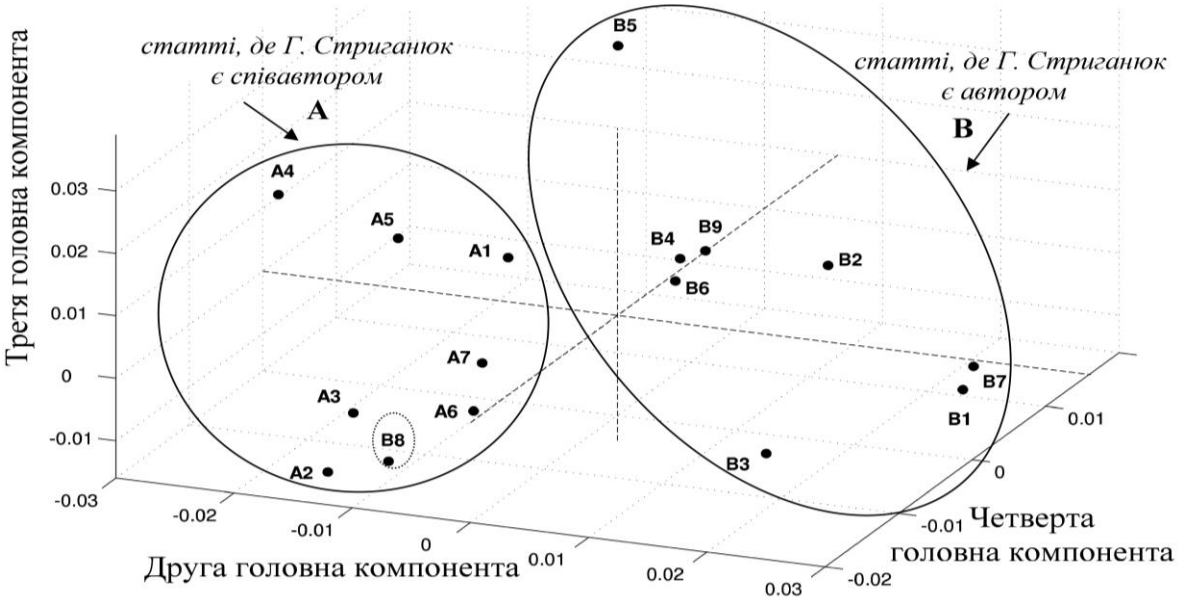


Рис. 6. Групування статей Г. Стриганюка у просторі головних компонент для послідовності з одного слова.

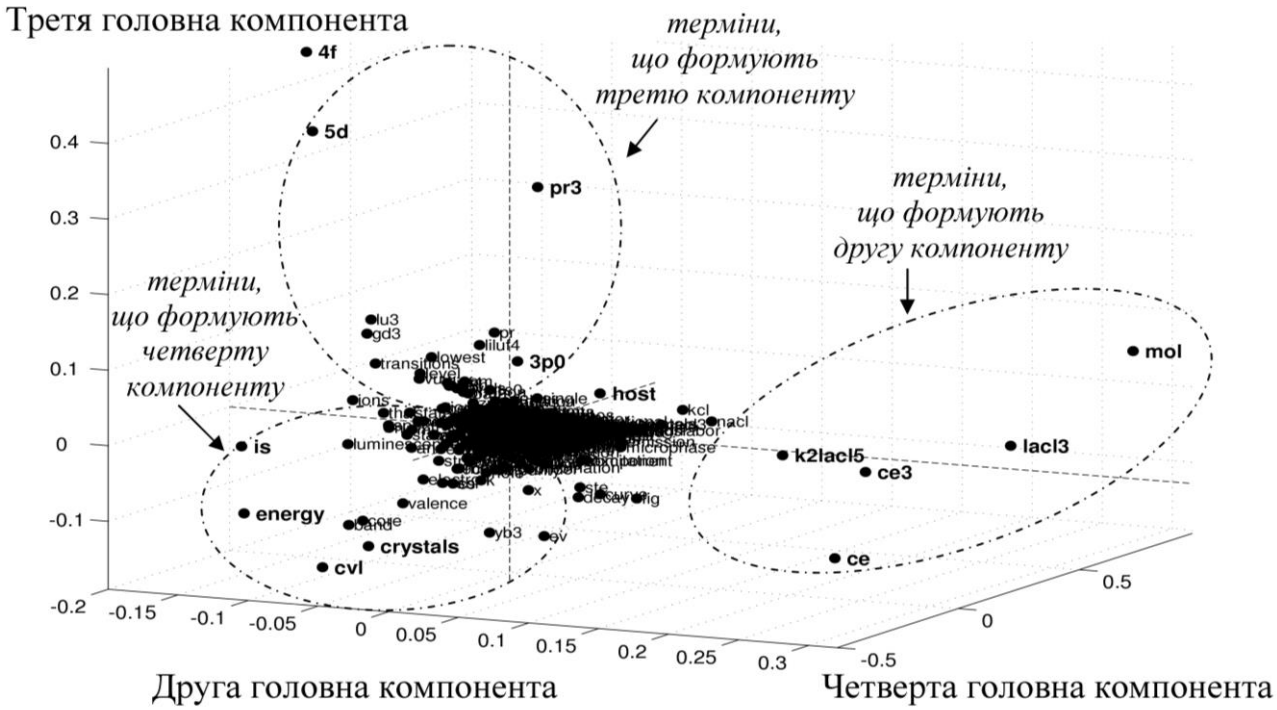


Рис. 7. Групування слів у просторі головних компонент.

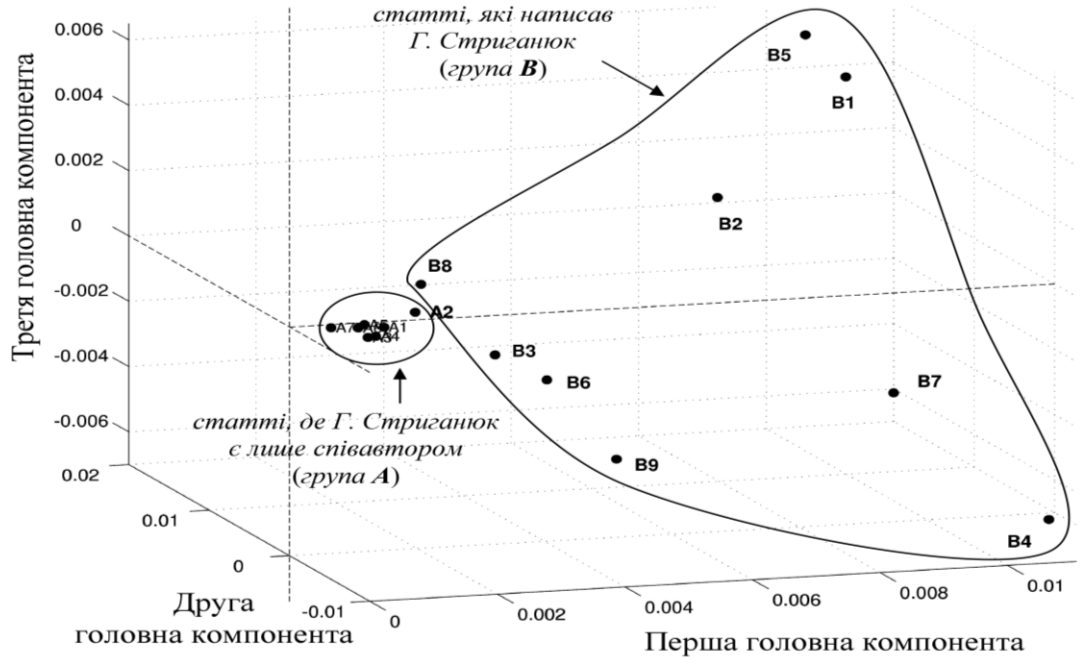


Рис. 8. Групування статей Г. Стриганюка (послідовність 4 слів).

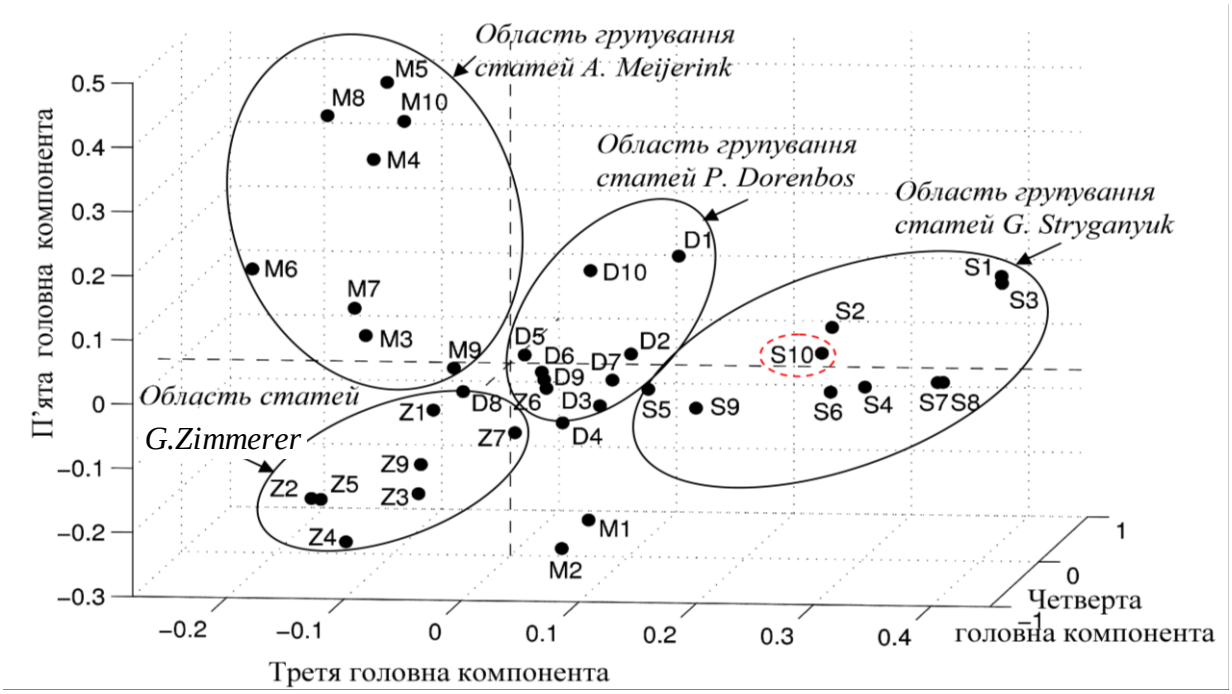


Рис. 9. Групування текстів P. Dorenbos, A. Meijerink, .G. Stryganyuk, G. Zimmerer (послідовність 4 слів).

Додаток Д

Список наукових статей Г. Стриганюка

Група А

A1 Variation of 5*d*-level position and emission properties of BaF_2:Pr crystals / P. A. Rodnyi, G. B. Stryganyuk, C. W. E. van Eijk et al. // Physical Review B – 2005. – Vol. 72 (19) – P. 1951121.1-1951121.6 .

A2 Features of the core-valence luminescence and electron energy band structure of $A_{1-x}Cs_xCaCl_3$ (A=K, Rb) crystal / Ya. Chornodolskyy, G. Stryganyuk, S. Syrotyuk et al. // Journal of Physics: Condensed Matter – 2007. – Vol. 19. – P. 476211.

A3 Europium Luminescence in Fluorite upon High-Energy Excitation / P. Rodny, A. Khadro, A. Voloshinovski, G. Stryganyuk // Optics and Spectroscopy. – 2007. – Vol. 103 (4). – P. 568.

A4 VUV 5d-4f luminescence of Gd^{3+} and Lu^{3+} ions doped into CaF_2 / V. N. Makhov, S. Kh. Batygov, L. N. Dmitruk, M. Kirm, G. Stryganyuk, G. Zimmerer // Phys. Stat. Sol. (C). – 2007. – Vol. 4-3. – P. 881 .

A5 Vacuum-ultraviolet 5d-4f luminescence of Gd^{3+} and Lu^{3+} ions in fluoride matrices / M. Kirm, G. Stryganyuk, S. Vielhauer et al. // Phys. Rev. B – 2007. – Vol. 75. – P. 75111.

A6 Intrinsic luminescence and luminescence of inadverten impurities in LuAP and LuYAP crystals / N. V. Guerasimova, I. A. Kamenskikh, V. V. Mikhailin et al. // Rus. Phys. J. – 2006. – Vol. 49. – P. 44.

A7 Creation of permanent lattice defects via exciton self-trapping into molecular states in Xe matrix / E.Savchenko, A.Ogurtsov, I.Khyzhniy et al. // Phys. Chem. Chem. Phys. – 2005. – Vol. **7**. – P. 785.

Група В

B1 Luminescent characteristics of pure and Ce doped K_2LaCl_5 phase in KCl host / A.Voloshinovskii, G.Stryganyuk, G.Zimmerer et al., // Phys. Stat. Sol. (A). – 2005. – Vol. 202 (9). – R101.

B2 Luminescence of Pr^{3+} doped K_2LaCl_5 microcrystals encapsulated in KCl host / G.B.Stryganyuk, P.V.Savchyn, Z.A.Khapko et al. // Opt. Mater. – 2008. – doi:10.1016/j.optmat.2008.07.001.

B3 Spectral-kinetic study of $LaCl_3$:Ce crystals / A.S.Voloshinovskii, O.T.Antonyak, G.B.Stryganyuk et al. // Ukr. J. Phys. Opt. – 2002. – Vol. 3(3). – P. 194.

B4 Luminescence of Ce^{3+} doped $LaPO_4$ nanophosphors upon Ce^{3+} 4f-5d and band-to-band excitation / G.Stryganyuk, D.Trots, A.Voloshinovskii et al. // J. of Lumin. – 2008. – Vol. 128. – P. 355.

B5 Spectral-kinetic characteristics of Pr^{3+} luminescence in $LiLuF_4$ host upon excitation in the UV–VUV range / G.Stryganyuk, G.Zimmerer, N.Shiran et al. // J. of Lumin. – 2008. – doi:10.1016/j.jlumin.2008.06.003.

B6 Charge transfer luminescence of Yb^{3+} ions in $LiY_{1-x}Yb_xP_4O_{12}$ phosphates / G.Stryganyuk, S.Zazubovich, A.Voloshinovskii et al. // J. Phys.: Condens. Matter. – 2007. – Vol. 19. – P. 036202.

B7 Luminescence of Ce doped $LaCl_3$ microcrystals incorporated into a single-crystalline NaCl host / V.V.Vistovskyy, P.V.Savchyn, G.B.Stryganyuk et al. // J. Phys.: Condens. Matter. – 2008. – Vol. 20. – P. 325218.

B8 Peculiarities of core-valence luminescence of CsBr and $Rb_{1-x}Cs_xBr$ crystals / A. Voloshinovskii, I. Pashuk, Ya. Chornodol'skyi et al. // Phys. Stat. Sol. (B). – 2004. – Vol. 241 (11). –P. 2613.

B9 Luminescence of YbP_3O_9 upon excitation in the UV–VUV range / G. Stryganyuk, D. Trots, I. Berezovskaya et al. // J. Phys.: Condens. Matter. – 2007. – Vol. 19. – P. 346236.

Додаток Е

Список статей, використаних для авторської атрибуції наукових текстів

№	Умовне позначення	***Pieter Dorenbos***
1	*D1*	Dorenbos P. The 5d level positions of the trivalent lanthanides in inorganic compounds / P. Dorenbos // Journal of Luminescence. - Elsevier Ltd., 2000 - Vol. 91. - pp. 155-176.
2	*D2*	Dorenbos P. 5*d*-level energies of Ce^{3+} and the crystalline environment. III. Oxides containing ionic complexes / P. Dorenbos // Physical Review B. - The American Physical Society, 2001. - Vol. 64. - pp. 125117 - 125117-12.
3	*D3*	Dorenbos P. Systematic behaviour in trivalent lanthanide charge transfer energies / P. Dorenbos // Journal of Physics: Condensed Matter. - Institute of Physics Publishing, 2003. - Vol. 15. - pp. 8417–8434.
4	*D4*	Dorenbos P. Exchange and crystal field effects on the $4f_{n-1}5d$ levels of Tb^{3+}/ P. Dorenbos // Journal of Physics: Condensed Matter. - Institute of Physics Publishing, 2003. Vol. 15. - pp. 6249–6268.
5	*D5*	Dorenbos P. 5*d*-level energies of Ce^{3+} and the crystalline environment. I. Fluoride compounds / P. Dorenbos // Physical Review B. - The American Physical Society, 2000. - Vol. 62, N. 23. - pp. 15640-15649.
6	*D6*	Dorenbos P 5*d*-level energies of Ce^{3+} and the crystalline environment. II. Chloride, Bromide, and Iodide compounds / P. Dorenbos // Physical Review B. - The American Physical Society, 2000. - Vol. 62, N. 23. - P. 15650- 15659.

7 *D7* Influence of the anion on the spectroscopy and scintillation mechanism in pure and Ce^{3+}-doped K_2LaX_5 and LaX_3 (*X*=Cl, Br, I) / E. van Loef, P. Dorenbos, C. van Eijk [et al.] // Physical Review B. – 2003. - Vol. 68. P. 0451080-0451089.

8 *D8* Lanthanide 4*f*-level location in lanthanide doped and cerium-anthanide codoped NaLaF4 by photo- and thermoluminescence / A. Krumpel, E. van der Kolk, D. Zeelenberg [et al.] // Journal of Applied Physics. – 2008. – Vol. 104, P. 073505-0-073505-10.

9 *D9* Scintillation and anomalous emission in elpasolite $Cs_2LiLuCl_6:Ce_{3+}$/ A. Bessière, P. Dorenbos, C. Van Eijk [et al.] // Journal of Luminescence. 2006. – Vol. 117. – P. 87–198.

10 *D10* Dorenbos P. Anomalous 10-ns emission in Ce^{3+}-doped Cs_3LuCl_6 / P. Dorenbos, E.V.D. van Loef, and C.W.E. van Eijk // PHYSICAL REVIEW B. – 2003.- Vol. 68. P. 125108 - 12511

11 *DA1* Dorenbos P. Predictability of 5d level positions of the triply ionized lanthanides in halogenides and chalcogenides / P. Dorenbos // Journal of Luminescence. –2000. – Vol. 87-89. – P. 970-972.

12 *DA2* Dorenbos P. The $4f^n \rightarrow 4f^{n-1}5d$ transitions of the trivalent lanthanides in halogenides and chalcogenides / P. Dorenbos // Journal of Luminescence. – 2000. – Vol. 91. – P. 91-106.

13 *DA3* Dorenbos P. 5*d*-level energies of $Ce^{3+}$and the crystalline environment. II. Chloride, bromide, and iodide compounds / P. Dorenbos // Physical Review B. –2000. – Vol. 62 (23). – P. 15 650- 15 659.

14 *DA4* Dorenbos P. Relating the energy of the $[Xe]5d^1$ configuration of Ce^{3+} in inorganic compounds with anion polarizability and cation electronegativity / P. Dorenbos // Physical Review B. – 2002. – Vol. 65 (23) . – P. 235110 – 235116.

15 *DA5* Dorenbos P. 5d-level energies of Ce^{3+} and the crystalline

environment.IV. Aluminates and ‘‘simple’’ oxides / P. Dorenbos // Journal of Luminescence – 2002. – Vol. 99. – P. 283–299.

16 *DA6* Dorenbos P. Energy of the first $4f^7$ -$4f^6$5d transition of Eu^{2+} in inorganic compounds / P. Dorenbos // Journal of Luminescence – 2003. – Vol. 104. – P. 239–260.

17 *DA7* Dorenbos P. Calculation of the energy of the 5d barycenter of $La_3F_3[Si3O9]:Ce^{3+}$ / P. Dorenbos // Journal of Luminescence – 2003. – Vol. 105. – P. 117–119.

18 *DA8* Dorenbos P. Locating lanthanide impurity levels in the forbidden band of host crystals / P. Dorenbos // Journal of Luminescence – 2004. – Vol. 108. – P. 301–305.

19 *DA9* Dorenbos P. The Eu^{3+} charge transfer energy andthe relation with the band gap of compounds/ P. Dorenbos // Journal of Luminescence – 2005. – Vol. 111. – P. 89–104.

20 *DA10* Dorenbos P. Absolute location of lanthanide energy levels and the performance of phosphors / P. Dorenbos // Journal of Luminescence – 2007. – Vol. 122-123. – P. 315–317.

21 *DA11* Dorenbos P. Energy of the Eu^{2+} 5d state relative to the conduction band in compounds / P. Dorenbos // Journal of Luminescence – 2008. – Vol. 128. – P. 578–582.

22 *DA12* Dorenbos P. Lanthanide charge transfer energies and related luminescence, charge carrier trapping, and redox phenomena / P. Dorenbos // Journal of Alloys and Compounds – 2009. – Vol. 488. – P. 568–573.

№ Умовне позначення

Andries Meijerink

1 *M1* $4f^n \rightarrow 4f^{n-1} \rightarrow 5d$ transitions of the light lanthanides: Experiment and theory / L. van Pieterson, M. Reid, R. Wegh [et al.] // Physical Review B. – 2002. – Vol. 65. – pp. 045113-1 - 045113-16.

2 *M2* $4f^n \rightarrow 4f^{n-1} \rightarrow 5d$ transitions of the heavy lanthanides: Experiment

and theory / L. van Pieterson, M. F. Reid, G. W. Burdick [et al.] // Physical Review B. – 2002. – Vol. 65. – pp. 045114-1 - 045114-13.

3 *M3* Quantum cutting by cooperative energy transfer in $Yb_xY_{1-x}PO4:Tb^{3+}$ / P. Vergeer, T. Vlugt, M. Kox [et al.] // Physical Review B. – 2005. – Vol. 71. – pp. 014119-1 - 014119-11.

4 *M4* Wegh R., Meijerink A. Spin-allowed and spin-forbidden $4f^n \rightarrow 4f^{n-1} \rightarrow 5d$ transitions for heavy lanthanides in fluoride hosts / R. Wegh, A. Meijerink // Physical Review B. – 1999. – Vol. 60(15). – pp. 10 820-10 830.

5 *M5* Visible quantum cutting in Eu^{3+}-doped gadolinium fuorides via downconversion / R. Wegh, H. Donker, K. Oskam, A. Meijerink // Journal of Luminescence. – 1999. – Vol. 82. – pp. 93-104.

6 *M6* High-resolution measurements of the vacuum ultraviolet energy levels of trivalent gadolinium by excited state excitation , P. Peijzel, P. Vermeulen, W. J. M. Schrama [et al.] // Physical Review B. – 2005. – Vol. 71. – P. 125126-1 - 125126-10.

7 *M7* $4f^{n-1} \rightarrow 5d \rightarrow 4f$ *n* emission of Ce^{3+}, Pr^{3+}, Nd^{3+}, Er^{3+}, and Tm^{3+} in $LiYF_4$ and YPO_4 / P. Peijzel, P. Vergeer, A. Meijerink [et al.] // Physical Review B. – 2005.- Vol. 71. – P. 045116-1 - 045116-9.

8 *M8* Vacuum-ultraviolet spectroscopy and quantum cutting for Gd^{3+} in $LiYF_4$ / R. Wegh, H. Donker, and A. Meijerink, [et al.] // Physical Review B. – 1997. - Vol. 56, N 21. – P. 13841-13848.

9 *M9* Oskam K. Site selective 4f5d spectroscopy of CaF_2 : Pr^{3+} / K. Oskam, A. Houtepen, A. Meijerink, // Journal of Luminescence. – 2002. – Vol. 97. – P. 107–114.

10 *M10* Wegh R. High-energy $^2G(2)_{9/2}$ emission and $4f^25d \rightarrow 4f^3$ multiphonon relaxation for Nd^{3+} in orthoborates and orthophosphates / R. Wegh,W. van Klinken, A. Meijerink // Physical Review B. 2001. – Vol. 64. – P. 045115-1 - 045115-5.

№	Умовне позначення	*Gregory Stryganyuk*
1	*S1*	Charge transfer luminescence of Yb^{3+} ions in $LiY_{1-x}Yb_xP_4O12$ phosphates / G. Stryganyuk, S. Zazubovich, A. Voloshinovskii [et al.] // Journal of Physics: Condensed Matter. – 2007. – Vol. 19. – P. 036202-1 - 036202-12.
2	*S2*	Photon cascade luminescence from Pr^{3+} ions in $LiPrP_4O_{12}$ polyphosphate / T. Shalapska, G. Stryganyuk, Yu. Romanyshyn [et al.] // Journal of Physics D: Applied Physics. – 2010. – P. 405404-1 - 405404-8.
3	*S3*	Luminescence of YbP_3O_9 upon excitation in the UV–VUV range / G Stryganyuk, D Trots, I. Berezovskaya [et al.] // Journal of Physics: Condensed Matter. 2007. – Vol. 19. – P. 346236-1 – 346236-11.
4	*S4*	Luminescence of Pr^{3+} doped K_2LaCl_5 microcrystals encapsulated in KCl host / G. Stryganyuk, P. Savchyn, Z. Khapko [et al.] // Optical Materials. – 2009. – Vol. 31. – P. 619–623.
5	*S5*	Spectral-kinetic characteristics of Pr^{3+} luminescence in $LiLuF_4$ host upon excitation in the UV–VUV range / G. Stryganyuk, G. Zimmerer, N.Shiran [et al.] // Journal of Luminescence. – 2008. - P. 1937-1941.
6	*S6*	Peculiarities of core-valence luminescence of CsBr and $Rb_{1-x}Cs_xBr$ crystals / A. Voloshinovskii, I. Pashuk, Ya. Chornodol'skyi G. Stryganyuk, G. Zimmerer, N.Shiran [et al.] // Phys. Stat. Sol. (b). – 2004. – Vol. 241 № 11. – P. 2613.
7	*S7*	Luminescence of Ce doped $LaCl_3$ microcrystals incorporated into a single-crystalline NaCl host / V. Vistovskyy, P. Savchyn, G. Stryganyuk [et al.] // J. Phys.: Condens. Matter. – 2008. – Vol. 20. – P. 325218.
8	*S8*	Luminescence of Ce^{3+} doped $LaPO_4$ nanophosphors upon Ce^{3+}

4f→5d and band-to-band excitation / G. Stryganyuk, D.Trots, A. Voloshinovskii [et al.] // Journal of Luminescence. – 2008. – Vol. 128. – P. 355–360.

9 *S9* Spectral-kinetic study of $LaCl_3$:Ce crystals/ A.S. Voloshinovskii, O.T. Antonyak, G.B. Stryganyuk [et al.] // Ukr. J. of Phys. Optics. – 2002. – Vol. 3 (3). – P. 194 – 199.

10 *S10* Luminescent characteristics of pure and Ce doped K_2LaCl_5 phase in KCl host A. Voloshinovskii, G. Stryganyuk, G. Zimmerer [et al.] // Phys. Stat. Sol. (a). 2005. – Vol. 202, № 9, Rapid Research Letters R101–R103.

№ Умовне позначення

Georg Zimmerer

1 *Z1* Stryganyuk G., Zimmerer G. RE^{3+} VUV $d\rightarrow f$ Luminescence Investigated by Synchrotron Radiation Excitation at HASYLAB / G. Stryganyuk, G. Zimmerer // Proceedings of the XIII FEOFILOV SYMPOSIUM “Spectroscopy of crystals doped by rare-earth and transition-metal ions”, Physics of the Solid State, Vol. 50, No. 9, Pleiades Publishing, Ltd., 2008, pp. 1631–1638.

2 *Z2* Zimmerer G Status report on luminescence investigations with synchrotron radiation at HASYLAB / G. Zimmerer // Nuclear Instruments and Methods in Physics Research A. – 1991. – Vol. 308. – P. 178-186.

3 *Z3* Zimmerer G. Excitons in rare-gas solids: Exotic or model-like? / G. Zimmerer // Journal of Luminescence. – 2007. – Vol. 125. – P. 287–293.

4 *Z4* Zimmerer G. Luminescence spectroscopy with synchrotron radiation: History, highlights, future / G. Zimmerer // Journal of Luminescence. – 2006. – Vol. 119–120. – P. 1–7.

5 *Z5* G. Zimmerer, SUPERLUMI:Aunique setup for luminescence spectroscopy with synchrotron radiation, Radiation Measurements.

– 2007. – Vol. 42. – P. 859 – 864.

6 *Z6* High-resolution vacuum ultraviolet spectroscopy of 5*d*-4*f* transitions in Gd and Lu fluorides / M. Kirm, J. Krupa, V. Makhov [et al.] // PHYSICAL REVIEW B. –2004. – Vol.**70**, Rapid Communications. – P. 241101(R).

7 *Z7* V.N. Makhov, N.M. Khaidukov, D. Lo , M. Kirm, G. Zimmerer, Spectroscopic properties of Pr^{3+} luminescence in complex fluoride crystals, Journal of Luminescence. –2003. – Vol. 102–103. – P. 638–643.

8 *Z8* V.N. Makhov, N.M. Khaidukov, N.Yu. Kirikova, M. Kirm, J.C. Krupa, T.V. Ouvarova, G. Zimmerer, VUV spectroscopy of wide band-gap crystals doped with rare earth ions, Nuclear Instruments and Methods in Physics Research A. 2001. – Vol. 470. – P. 290–294.

9 *Z9* G. Zimmerer, Luminescence properties of rare gas solids, Journal of Luminescence. – 1979. – Vol. 18/19. – P. 875–881.

10 *Z10* V.N. Makhov, N.M. Khaidukov, N.Yu. Kirikova, M. Kirm, J.C. Krupa, T.V. Ouvarova, G. Zimmerer, VUV emission of rare-earth ions doped into fluoride crystals, Journal of Luminescence. – 2000. – Vol. 87/89. – P. 1005–1007.

Додаток Є

Список “функціональних слів” [256]

a
about
above
accordingly
across
after
afterwards
again
against
albeit
all
allow
allowable
allowed
allows
allowing
almost
alone
along
already
also
although
always
am
among
amongst
an
and
another
any
anybody
anyhow
anyone
anything
anywhere
apart
appear
appears
appropriate
are
aren
around
as
at
away
be
became
because
become
becomes
been
before
beforehand
behind
below
beside
besides
between
beyond
big
both
but
by
considering
cannot
co
consequently
consider
considerable
considered
can
considers
contain
containing
contains
corresponding
could
currently
did
didn
do
does
doesn
doing
done
down
downwards
dozen
during
each
either
else
enough
entire
entirely
et
etc
even
ever
every
ex
example
except
exclusive
exclusively
far
few
first
firstly
for
former
forth
found
from
further
furthermore

get
given
go
gone
got
had
hadni
half
hardly
has
have
having
hence
here
hereafter
hereby
herein
hereupon
hitherto
how
howbeit
however
hundred
ie
if
immediate
in
inasmuch
inc
include
included
includes
including
indeed
indicate
indicated
indicates
inner
insofar
instead
into
inward
is
isn
it
its
itself
just
last
latter
latterly
least
less
lest
like
little
many
may
mean
meaning
means
meanwhile
might
missed
misses
missing
more
moreover
most
mostly
much
must
name
namely
near
necessary
neither
never
nevertheless
new
next
no
nobody
none
noone
nor
normally
not
note
notes
nothing
now
nowhere
of
off
often
oh
old
on
once
one
only
onto
or
other
others
otherwise
ought
out
outside
over
overall
own
particular
particularly
per
perhaps
placed
please
plus
possible
probably
provides
questionable
quite
rather
really
relatively

respectively
right
said
same
secondly
see
seem
seemed
seeming
seems
self
sensible
sent
serious
several
shall
should
shouldn
since
so
some
somebody
somehow
someone
something
sometime
sometimes
somewhat
somewhere
specified
specify
specifying
still
sub
such
sup
taken
than
that
the
then
thence
there
thereafter
thereby
therefore
therein
thereupon
these
this
thorough
thoroughly
those
though
thousand
through
throughout
thus
to
together
too
toward
towards
under
unless
until
unto
up
upon
use
used
useful
uses
using
usually
value
various
very
via
viz
vs
want
was
wasn
way
well
went
were
weren
what
whatever
when
whence
whenever
where
whereafter
whereas
whereby
wherein
whereupon
wherever
whether
which
while
whither
who
whoever
whole
whom
whose
why
will
with
within
without
would
wouldn
yet

Додаток Ж

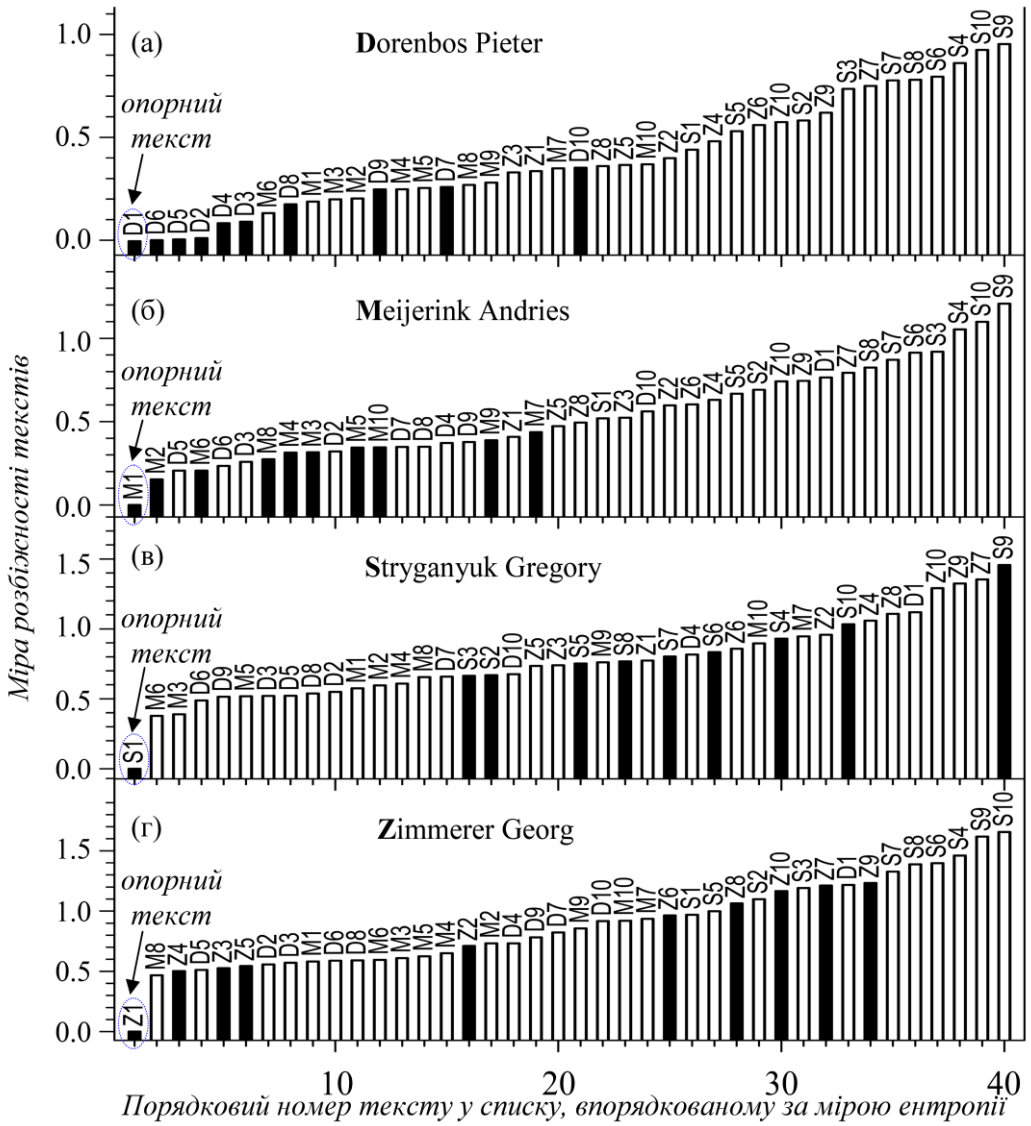


Рис. 1. Розподіл текстів за мірою розбіжності, обчисленою методом ентропії, беручи за опорний текст словник “функціональних слів”.

Зафарбовані прямокутники відповідають текстам автора, які розпізнаються. Тексти відсортовані по осі абсцис в міру зростання розбіжності між досліджуваними текстами та опорним текстом. Уздовж осі ординат відображено дивергенцію між текстами.

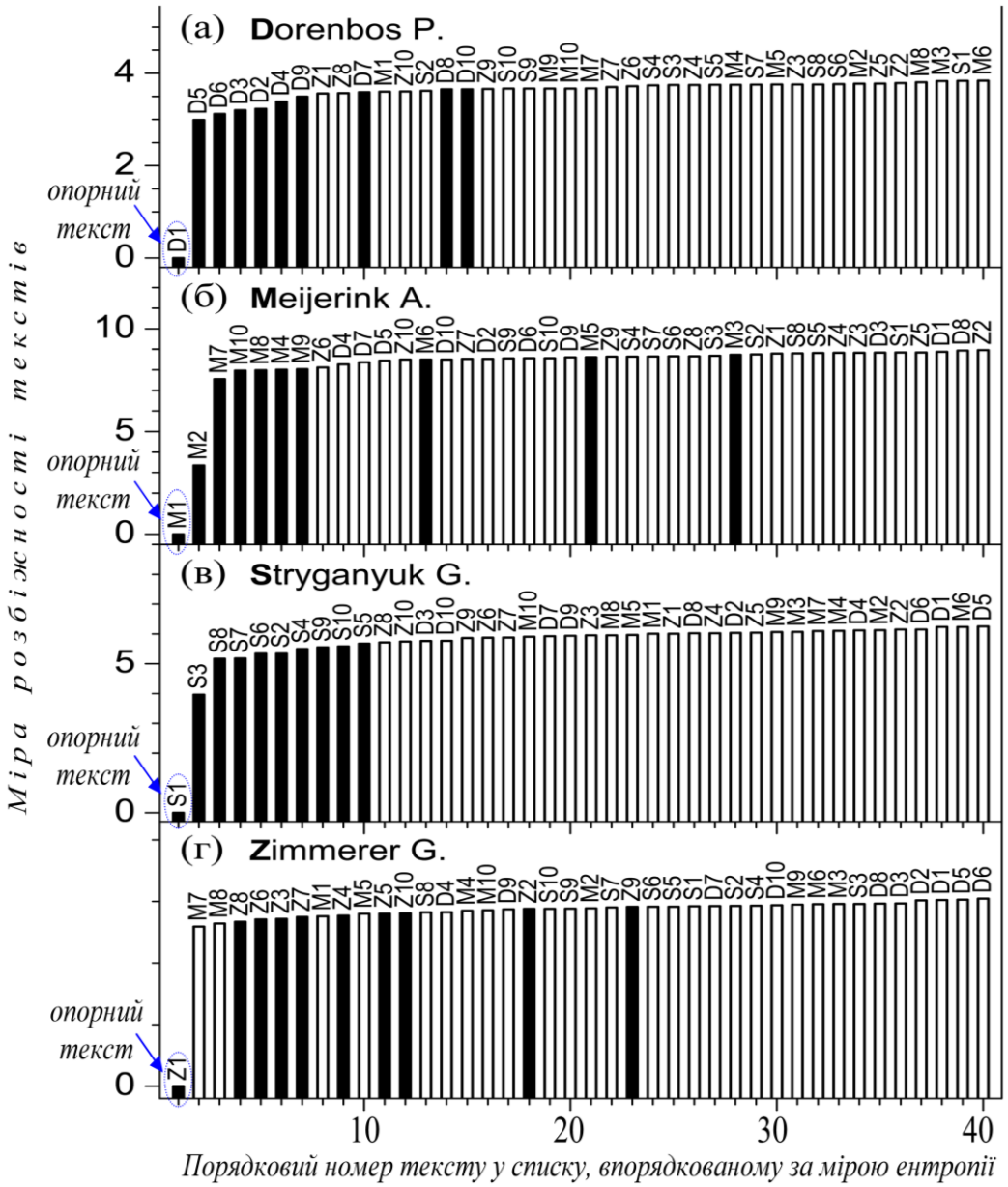


Рис. 2. Розподіл текстів за мірою розбіжності, обчисленою методом ентропії для послідовності чотирьох слів із урахуванням найкращого опорного тексту.