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

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

Автореферат дисертації на здобуття наукового ступеня
кандидата філологічних наук

Київ – 2013

Дисертацією є рукопис

Робота виконана в Національному педагогічному університеті імені М. П. Драгоманова, Міністерство освіти і науки України

Науковий керівник доктор філологічних наук, доцент
Толчеєва Тетяна Станіславівна,
Національний педагогічний університет
імені М. П. Драгоманова,
кафедра загального мовознавства і германістики,
завідувач кафедри

Офіційні опоненти: доктор філологічних наук, старший науковий співробітник
Радзієвська Тетяна Вадимівна,
Інститут мовознавства ім. О. О. Потебні НАН України,
відділ загального мовознавства,
провідний науковий співробітник

кандидат філологічних наук, доцент
Терехова Діана Іванівна,
Київський національний лінгвістичний університет,
кафедра теоретичної і прикладної лінгвістики
та новогрецької філології, доцент

Захист відбудеться "17" вересня 2013 р. о 13^{30} годині на засіданні спеціалізованої вченої ради К 26.053.15 у Національному педагогічному університеті імені М. П. Драгоманова за адресою: 01601, м. Київ-30, вул. Пирогова, 9

З дисертацією можна ознайомитись у бібліотеці Національного педагогічного університету імені М. П. Драгоманова за адресою: 01601, Київ-30, вул. Пирогова, 9

Автореферат розісланий "15" серпня 2013 р.

Учений секретар
спеціалізованої вченої ради Л. В. Кравець

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

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 ймовірну 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. Шевельов, S. Argamon, J. Binongo, M. Koppel), фізики (Ю. Головач, А. Ровенчак, І. 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), і 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 за допомогою 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- й 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- й україномовних наукових і художніх текстів;

– виявити критерії розмежування англо-, німецько- й 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;

– схарактеризувати 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- й україномовних художніх текстів.

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

Предметом аналізу є стильова, тематична й 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 стилів; *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 одержаних результатів визначається тим, що в роботі *вперше*: 1) *розроблено* комплексну методику здійснення стильової, тематичної та авторської атрибуції англо-, німецько- й 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* найчастотніші слова у наукових і художніх текстах досліджуваних мов і проаналізовано загальні тенденції їх вживання (серед найчастотніших 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 виконання 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 найчастотніших 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), “Шевченківська весна: 2013” (Київ 2013), “Научная дискуссия: вопросы филологии, искуствоведения и культурологии” (Москва 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 і 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, додатки займають 34 сторінки.

ОСНОВНИЙ ЗМІСТ

У **вступі** обґрунтовано актуальність теми дослідження, сформульовано мету, визначено завдання, об'єкт, предмет, методи дослідження, наукову новизну та практичне значення одержаних результатів, схарактеризовано фактичний матеріал, указано форми апробації та представлено структуру роботи.

У **першому** розділі **"Теоретичні засади вивчення атрибуції текстів різних функціональних стилів"** критично проаналізовано визначення терміна "атрибуція" та суміжних термінів у сучасному мовознавстві, охарактеризовано лінгвістичні параметри атрибуції наукових та художніх текстів, а також текстів інших функціональних стилів, особливу увагу приділено застосуванню лінгвістичного параметра послідовності вживання літер/слів для атрибуції текстів.

Термін "атрибуція" у сучасному мовознавстві не має і дотепер однозначного витлумачення. Так, в Енциклопедії української мови терміни "атрибуція" й "авторизація" визначаються як синоніми: *атрибуція* – встановлення авторства тексту на основі композиції, способів текстотворення, почерку, мови змісту і позатекстових відомостей про його походження та історію. Атрибуція здійснюється шляхом зіставлення неавторизованого твору з авторизованим, щоб на підставі подібності (відмінності) довести припущення про авторство. Останнє часто є причиною того, що атрибуцію називають ще й *авторизацією*. В інших словниках ці терміни розрізняють: а) *атрибуція* – визначення достовірності, аутентичності художнього твору, його автора, місця й часу створення; *авторизація* – підтвердження авторства, авторського права (Великий тлумачний словник української мови); б) *атрибуція* – приписування анонімного художнього твору певному авторові; *авторизація* – надання автором уповноважень, згоди в будь-якій справі (Словник чужомовних слів). У словнику "Thinkmap visual thesaurus" тлумачення терміна "атрибуція" виходить за межі визначення автора тексту, а як синоніми до нього подано: 1) *ascription* (приписування певної якості, характеристики людині або предмету); 2) *categorization, classification* (категоризація, класифікація, групування людей, предметів у подібні класи, категорії); 3) *sorting* (сортування, упорядкування об'єктів відповідно до певного критерію).

У контексті такої термінологічної невизначеності в мовознавстві сформувалися два підходи до змістового об'єму поняття "атрибуція". Одні науковці вважають, що атрибуція може стосуватися тільки питань визначення автора тексту, інші ж – розглядають її у ширшому розумінні та, окрім визначення автора тексту, застосовують для визначення стилю, тематики, хронологічних особливостей текстів. У рамках цих підходів вирізняють: 1) *авторську атрибуцію* – встановлення автора

тексту (А. Баранов, М. Пещак); 2) *не авторську атрибуцію* – віднесення тексту до певної мови, стилю, періоду часу, літературної школи, літературного напрямку і т. ін. (С. Бук, П. Вашак, Г. Мартиненко, H. Baayen, M. Koppel, Y. Yang). Завдання як *авторської*, так і *не авторської* атрибуції полягає у віднесенні тексту до наперед зафіксованої множини критеріїв групування текстів на основі подібності лінгвостилістичних характеристик (Г. Мартиненко).

У дисертаційній праці прийнято погляд, згідно з яким атрибуцію не можна зводити лише до визначення автора тексту, а слід сприймати як розподіл текстів за групами відповідно до певного критерію. Таким критерієм може бути довільна ознака (тематика, стиль, жанр, мова, століття написання тексту тощо) відповідно до поставленого дослідницького завдання. На основі цієї позиції пропонуємо робоче визначення ***атрибуції*** як процесу упорядкування довільних текстових документів у групи за функціональним стилем, тематикою або автором за наперед заданими критеріями або визначеними під час цього процесу.

Аналіз наукових праць, присвячених вивченню атрибуції текстів різних функціональних стилів, свідчить про різні аспекти здійснення атрибуції художнього стилю, зокрема *авторської атрибуції* художніх текстів (П. Берков, П. Вашак, Н. Дарчук, О. Кукушкіна, М. Марусенко, М. Пещак, І. Севбо, Г. Хетсо, Д. Хмельов, D. Holmes, P. Juola, M. Koppel, J. Rudman). Окремий інтерес становить питання вибору оптимальних лінгвістичних параметрів для розмежування текстів за *стилем* (В. Критська, С. Сушко, О. Шевельов, P. Grzybek, Y. Yang). Останнім часом спостерігається інтерес науковців і до атрибуції електронних повідомлень, листів, передусім дослідники аналізують помилки щодо повторення літер, заміни літер, інверсії літер, пропущення літер, злиття слів (I. Biskub, M. Koopel), структурне оформлення та форматування текстів (K. Calix, O. de Vel).

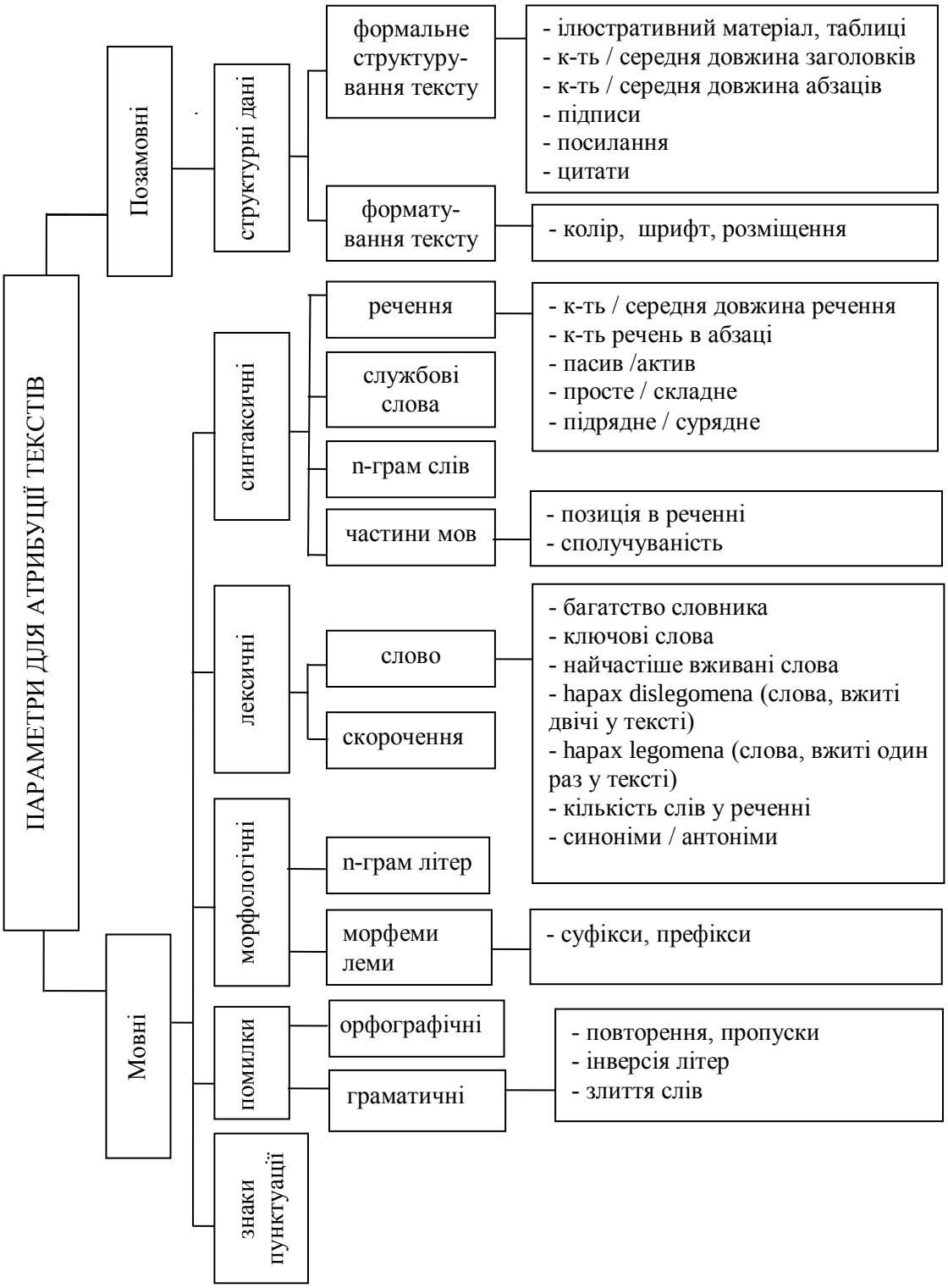
Помітною є тенденція і до вивчення *атрибуції наукових текстів* (А. Коваль, І. Колегаєва, Т. Радзієвська, Н. Разінкіна, Т. Яхонтова, I. Posner). Здебільшого аналізу підлягають відмінності між жанрами *наукового стилю* (J. Swales); особливості структури *наукової статті* (Т. Яхонтова); закономірності *вживання різних частин мови* (A. Reimerink); лексичний склад, зокрема *терміни, ключові слова* (H. Glasman-Deal). Питання ж авторської атрибуції наукових статей досліджено спорадично (I. Posner), адже ця процедура має свою специфіку. Наукова стаття не завжди є одноосібною, а арсенал мовних засобів, на відміну від художнього твору, прямує до уніфікації. До того ж у науковому стилі будь-який вияв індивідуальності є допустимим, проте це не органічна якість цього стилю, елементи індивідуального хоча і "вкраплюються" у текст наукового викладу, та залишаються все ж таки чужостильовими (А. Коваль, Н. Разінкіна).

Важливою проблемою також є виявлення лінгвістичних параметрів, придатних для авторської атрибуції текстів. На сьогодні ще не запропоновано усталеного набору параметрів, на основі яких можна було б стандартизувати процеси атрибуції текстів за автором (М. Штокмар, M. Jockers, J. Rudman). У дисертації розроблено класифікацію оптимальних параметрів для здійснення атрибуції текстів, які скласифіковано на *мовні* (синтаксичні, лексичні, морфологічні, знаки пунктуації, помилки) та *позамовні* (структурні критерії) (табл. 1).

Ефективною є процедура атрибуції текстів у разі використання *лінгвістичного параметра послідовності вживання літер або слів* (відомий як "*n*-грам" (W. Cavnar).

Таблиця 1

Параметри для здійснення атрибуції текстів



За допомогою послідовності вживання однієї і більше літер здійснено спроби *стильової* (С. Сушко), *тематичної* (J. Bellegarda, M. Mahajan) та *авторської* (Д. Хмельов, J. Grieve, V. Kešelj, F. Peng) атрибуції текстів.

Другий розділ **“Методологічна база аналізу атрибуції текстів різних функціональних стилів”** присвячено огляду методів атрибуції текстів різних функціональних стилів, серед яких виділено метод моніторингу групування текстів та відповідних їм слів, проаналізовано програмні системи для стильової та авторської атрибуції художніх текстів різних мов, запропоновано комплексну методику стильової, тематичної й авторської атрибуції текстів.

Зародження статистичного аналізу тексту датується кінцем XIX – поч. XX ст. (П. Струве). Для виконання мовознавчих завдань (С. Айвазян, С. Бук, В. Глинський, Б. Головін, В. Левицький, В. Перебийніс, Р. Піотровский) застосовують різні математико-статистичні методи. З-поміж методів атрибуції можна виділити: метод машини опорних векторів, штучні нейронні мережі, дерева рішень, метод ентропії. Існує низка програмних продуктів для атрибуції текстів: “Авторовед”, “Атрибутор”, “ВААЛ”, “Лингвоанализатор”, “СМАЛТ”, “Delta”, “JGAAP”. Більшість із цих програм не передбачають зміни параметрів опрацювання даних, що не дозволяє використовувати їх як інструмент наукового дослідження. Такі програми працюють із текстами художньої літератури. Можливість атрибуції наукових текстів у них не розглядається. Це зумовлює необхідність розробки програм для атрибуції наукових текстів, що і було зроблено у цій роботі.

Перший етап передбачає створення різних вибірок текстів залежно від типу атрибуції. Для *стильової* атрибуції текстів: 1) наукового стилю підібрано англо-, німецько- й україномовні тексти різних жанрів (монографія, стаття, дисертація) з галузі фізики, щоб простежити за рангово-частотним розподілом слів залежно від жанру текстів; 2) художнього стилю підібрано як аутентичні англо-, німецько-, україномовні твори класиків світової художньої літератури XIX–XXI ст., так і перекладені англійською мовою (з цих творів створено вибірки, що містили тексти: а) написані носіями англійської, німецької, української мов; б) перекладені з російської мови англійською; в) написані носіями англійської мови та перекладені англійською мовою; г) написані у XIX ст. або XXI ст.). Таке розмаїття вибірок створено для аналізу залежності рангово-частотного розподілу слів від згаданих відмінностей наукових і художніх текстів, а також для виявлення характерних для кожного зі стилів найчастіше вживаних слів. Для *тематичної* атрибуції обрано праці наукової конференції, оскільки тематика конференційних секцій є заздалегідь визначеною, а це полегшує процедуру визначення ефективності групування текстів. Для *авторської* атрибуції наукових текстів залучено вузькоспеціалізовані наукові статті в галузі спектроскопії та матеріалознавства (підбір текстів з однаковою тематикою та експериментальними методами дослідження передбачає, що розмежування цих текстів буде швидше авторським, ніж тематичним). З метою встановлення авторства художніх текстів обрано тексти авторів XXI ст.

На *другому етапі* за допомогою описового методу вибрано оптимальний параметр для проведення атрибуції текстів – лінгвістичний параметр послідовності вживання слів (*n*-грам слів).

Третій етап пов'язаний із визначенням частот цих параметрів за допомогою розробленої у цій дисертаційній праці програми “Lexical Content Searcher”, яка дає змогу отримати частотні словники вживання одного і більше слів для кожного тексту, а також узагальнений частотний словник слів для всього масиву текстів.

На *четвертому етапі* здійснено підбір відповідних методів та алгоритмів, які забезпечують успішну атрибуцію текстів. *Метод частотного розподілу слів* (закон Ципфа) застосовано для стильової атрибуції наукових і художніх текстів, а на основі *зіставно-типологічного* методу виявлено характерні для кожного зі стилів найчастіше вживані слова. *Метод одночасного моніторингу групування текстів та відповідних їм слів* використано для тематичної та авторської атрибуції, у результаті чого тексти були згруповані за тематикою конференційних секцій, за автором наукового та художнього текстів, за ключовими словами тематичних секцій конференції, а також проаналізовані послідовності вживання одного та більше слів для авторів наукових і художніх текстів. Запропоновано *ентропійний метод* (дивергенція Кульбака-Лайблера) для авторської атрибуції текстів. Для цього розроблено програму "LinearAnalisys", яка розраховує "розбіжність" текстів за ключовими словами або довільним текстом, з яким порівнюють інші тексти. Метод одночасного моніторингу групування текстів та відповідних їм слів виокремлено серед інших, оскільки він передбачає можливість одночасного проведення атрибуції текстів та виділення семантичних груп слів і їх зіставлення із ознаками текстів, що ними описуються (Е. Браверман, К. Есбенсен, I. Jolliffe). Вивчення текстів різних авторів за допомогою цього методу дало змогу виявити зміну частоти вживання слів при переході від одного тексту до іншого. Слова групуються за закономірністю зміни частоти їх вживання у текстах: якщо група слів, що вживалася дуже часто в одному тексті, відсутня в іншому, то можна припустити, що характеристика, описана цією групою слів, не притаманна другому тексту. Такою характеристикою може виступати тематика тексту або ж авторський стиль. Перевага цього методу полягає в можливості атрибуції текстів без наперед заданих критеріїв. Розглянутий метод застосовують для авторської атрибуції художніх текстів (H. Baayen, J. Binongo, J. Burrows, D. Holmes, P. Juola), проте він не був апробований для авторської атрибуції наукових текстів.

П'ятий етап передбачав візуалізацію результатів атрибуції через відображення розподілу текстів графічно у просторі головних компонент. Результати оцінки параметрів рангово-ймовірнісного розподілу Ципфа для наукової і художньої літератури представлено за допомогою програмного пакета Origin-8.

Третій розділ **"Стильова атрибуція наукових і художніх англо-, німецько- та україномовних текстів"** присвячено дослідженню закономірностей частотного розподілу слів у наукових і художніх англо-, німецько- й україномовних текстах, зіставленню рангово-частотного розподілу слів у наукових і художніх текстах досліджуваних мов.

Основним критерієм розмежування англо-, німецько- й україномовних текстів наукового і художнього стилів є залежність рангово-частотного розподілу слів у тексті. Для стильової атрибуції текстів проаналізовано закон Ципфа (описує взаємозв'язок частоти слова в тексті та його рангу у списку слів) і його модифікації. Ранговий розподіл Ципфа можна представити трьома зонами: перша стосується слів з високою частотою появи у тексті; друга – прямолінійна, де розташовані слова з середньою частотою вживання; третя відповідає за низькочастотні слова. Закон Ципфа, як правило, дає змогу описати лише слова з середньою частотою появи в тексті. Аналіз широкого вибору модифікацій закону Ципфа дозволив виділити:

а) закон Мандельброта, який покращує результати опису в діапазоні високочастотних слів, та б) закон Юла-Саймона, що сприяє опису низькочастотних слів. Адаптовано параметр q з функції Мандельброта до базової функції Лавалетті. Запропонована у дисертації модифікація функції Лавалетті $f(k;q,s,n)$ передбачає два параметри (q та s) для опису рангово-ймовірнісного розподілу слів:

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

де f – ймовірність появи слова у тексті; k – порядковий номер (ранг) слова у списку; q – параметр, що описує розподіл високочастотних слів; s – параметр, що описує розподіл слів із середньою частотою вживання; n – обсяг словника.

Комплексний аналіз англо-, німецько- й україномовних наукових і художніх текстів показав, що математичний параметр s модифікованої функції Лавалетті відповідає за функціональний стиль. Стрімкість спаду ймовірності появи слова виявилася меншою у науковій літературі, а отже, і кількісні показники параметра s для наукових текстів є меншими порівняно із художніми текстами. Для кожної з досліджуваних мов визначено межі інтерквантильних інтервалів параметра s для наукових і художніх текстів. Межі інтерквантильних інтервалів не перетинаються для англомовних наукових [0,85 – 1,01] і художніх [1,04 – 1,12] текстів, а тому між текстами існує значима статистична різниця: вони належать до різних функціональних стилів. Аналогічні результати отримано і для німецькомовних наукових [0,86 – 0,94] і художніх [0,97 – 1,13] текстів та україномовних наукових [0,82 – 0.86] і художніх [0,89 – 0,97] текстів. Якщо кількісний показник параметра s певного тексту входить у вищевстановлені межі інтерквантильного інтервалу, то цей текст є науковим або художнім. Щоб перевірити ефективність запропонованої методики проаналізовано: 1) наукові тексти різного жанру (у випадку одного жанру наукової літератури, а саме: наукові статті з галузі фізики, закономірності рангово-частотного розподілу слів (s=0,984) близькі до розподілу слів у наукових текстах різних жанрів (s=1,001)); 2) художні тексти різних століть та мови написання твору (рангово-частотні розподіли слів є близькими і перебувають в інтервалі [1,12 – 1,17], не виходячи за межі 95% інтерквантильного інтервалу [1,06 – 1,22], знайденого для вибірки однорідної щодо носія мови та часу написання текстів. Це свідчить про те, що відмінності між параметром s для цих текстів несуттєві, вибірки можна вважати однорідними і вони належать до одного стилю).

Спільним у науковому тексті для трьох мов є наявність у першій тридцятці найчастіше вживаних службових частин мови та дієслова *бути*, відсутність займенників, іменників. У художньому тексті – висока частота вживання займенників (“*I*”, “*Ich*”, “*Я*” має подібну частоту вживання у зіставлюваних мовах). Для англо-, німецько- й 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).

Загальна тенденція для проаналізованих наукових текстів простежується у наявності серед перших 300 найчастіше вживаних слів загальнонаукових термінів

(*data, figure, function, model, methods, parametrs, result, values; das System, der Parametr, die Methode, das Ergebnise, die Berechnung, die Gleichung; структура, метод, величина, система, параметр, результат*). У художньому тексті для трьох досліджуваних мов переважають слова на позначення частин тіла (*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 (*people, father, God, mother, die Mutter, der Gott, der Vater, die Menschen, мати, батько, людина, Бог*).

У **четвертому** розділі **"Тематична та авторська атрибуція англомовних наукових текстів"** проаналізовано тематичну атрибуцію наукових текстів та можливості виділення тематичних напрямків за заголовками, анотаціями та тезами наукових доповідей, запропоновано інтегровану методику аналізу авторської атрибуції наукових текстів методом одночасного моніторингу групування текстів та відповідних їм слів із залученням послідовності вживання слів та розглянуто ентропію як метод авторської атрибуції наукових текстів.

Для тематичної атрибуції наукових праць Міжнародної конференції LUMDETR-2006 (Люмінесцентні Детектори та Перетворювачі Іонізуючого Випромінювання) ефективною виявилась послідовність одного слова. Виділено 5 тематичних секцій конференції із 9. За допомогою зіставлення розподілів текстів із розподілом словоформ виокремлено слова, показові для виділених секцій. Наприклад, для секції "Дозиметричні матеріали" характерним є вживання термінів, що описують особливості параметрів дозиметричних матеріалів (*glow, peak, dose*) чи методів дослідження (*radiation, irradiation, heating*). Терміни *storage, X-ray, PSL (photostimulated luminescence), phospors* та формули хімічних сполук $CsBr$, $CsEuBr_3$, $EuAl_2O_4$ є найбільш уживаними для текстів секції "Запасаючі та інші фосфори". Для секції "Домішки, дефекти, пастки" – *TSL, temperature, deffects,* $PbWO_4$.

Якщо відсутній доступ до повних текстів наукових статей, то інформацію про них можна отримати, аналізуючи менші за розміром від статей їх заголовки, анотації або тези. Підставою для цього є близьке просторове групування пар "теза-стаття" та тріад "заголовок-анотація-стаття". Значна відстань між текстами деяких із розглянутих пар "теза-стаття" може відображати зміни, внесені у статтю, порівняно з текстом тез. Природно, що автори можуть виправляти початкову ідею досліджень та переглядати зроблені висновки у період між реєстрацією тез доповіді та редагуванням остаточної версії статті. Редакторські виправлення слід також урахувати.

Авторська атрибуція ускладнюється для наукових статей, де автор обмежений у стилі та засобах вираження своїх думок. Проаналізовано 16 статей, що є результатом колективних досліджень у співавторстві з Г. Стриганюком. Їх поділено на: групу А (Г. Стриганюк виконував один із фрагментів дослідження); групу В (остаточне редагування статей проводив Г. Стриганюк). Найбільш чітко групування статей за автором досягнуто для послідовності чотирьох слів, при якій тексти групи А та В згурпувались у дві окремі області. Виділено характерні для цього автора послідовності чотирьох слів, які виражають: припущення (*one can expect that, may correspond to the, may be caused by*); описують спостережувані об'єкти (*have been found, is revealed for, appears due to*); представляють та порівнюють результати (*it is concluded that, is estimated to be, is evaluated to be*).

Валідні результати отримано і для одночасної атрибуції статей P. Dorenbos, A. Meijerink, G. Stryganyuk та G. Zimmerer, де оптимальний розмір параметра послідовності вживання слів також був чотири слова. Статті цих авторів згрупувались у чотири області, які було використано для розпізнавання невідомих текстів. До текстів G. Stryganyuk додано ще одну його статтю. Після повторного групування текстів вона увійшла саме в область статей цього автора. Кількість спільних послідовностей чотирьох слів для авторів наукових текстів мінімальна. Серед перших 70 найчастіше вживаних послідовностей чотирьох слів виявлено лише одну спільну послідовність *the excitation spectrum of.* Для статей P. Dorenbos, G. Stryganyuk і G. Zimmerer спільною є послідовність *in the case of*. Зменшення спільних послідовностей чотирьох слів між авторами сприяє покращенню авторської атрибуції текстів. Варто зазначити, що і для одного автора ймовірність появи однакової послідовності чотирьох слів у декількох текстах дуже мала. Найменша кількість послідовностей чотирьох слів, що повторюються у декількох текстах, є у G. Zimmerer, а найбільша – у A. Meijerink. Найчастіше вживаними послідовностями чотирьох слів є*: the energy of the* (P. Dorenbos); *the intensity of the (*A. Meijerink); *in the range of* (G. Stryganyuk); *at the superlumi station* (G. Zimmerer). Спільними найчастіше вживаними моделями послідовностей чотирьох слів у наукових текстах є: article + noun + preposition + article (*the size of the, the transition of the*) та preposition + article + noun + preposition (*on the basis of, for the formation of*). Наявні також групи послідовностей чотирьох слів, які вживаються з подібною частотою, і об'єднавши їх, можна відтворити частину речення. Це відбувається переважно тоді, коли автор копіює речення з одного тексту в інший і частково його змінює. Наприклад, *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* A. Meijerink використовує у тому ж контексті (*The authors are grateful to Dr. P. Gürtler from HASYLAB for the opportunity to use the excellent facilities for VUV spectroscopy at the DESY synchrotron).*

Авторська атрибуція текстів була також проведена методом ентропії (дивергенція Кульбака-Лайблера). Ефективності методів ентропії й одночасного моніторингу групування текстів і відповідних їм слів є співмірними. Проте застосування методу одночасного моніторингу групування текстів і відповідних їм слів дає змогу здійснювати моніторинг лексики, що визначає характеристики аналізованих текстів, та дає детальну інформацію про роль цих одиниць у тексті. Метод ентропії Кульбака-Лайблера забезпечує одномірне представлення результатів атрибуції, вимагає для порівняння вибору опорного тексту та не передбачає аналізу вкладу лексичних одиниць.

У **п'ятому** розділі **"Авторська атрибуція англо-, німецько- та україномовних художніх текстів"** використано метод одночасного моніторингу групування текстів та відповідних їм слів і показано, що для англо-, німецько- й україномовних художніх текстів оптимальна ідентифікація за авторським стилем досягається, аналізуючи послідовність трьох слів. Зі збільшенням розміру послідовності слів до чотирьох, п'яти або ж його зменшення до двох ефективність атрибуції обраним методом суттєво зменшується.

Найуживанішими послідовностями трьох слів є: а) *there was a, for a moment, out of the, it was a* (англомовний художній текст); б) *Ich weiß nicht, hin und her, es war ein, nach einer Weile* (німецькомовний художній текст); в) *так і не, та ще й, раз у раз, що все знають* (україномовний художній текст).

Аналіз найчастіше вживаних послідовностей трьох слів у проаналізованих текстах дав змогу встановити послідовності, характерні для всіх текстів трьох мов. Серед перших 35 найчастіше вживаних послідовностей трьох слів спільними виявились чотири послідовності для англійської (*there was a, out of the, one of the, the back of*), чотири – для німецької (*Ich weiß nicht, es war ein, an der Wand, schüttelte den Kopf*) та дві – для української (*так і не, що в нього*) мов.

Для англомовних художніх текстів серед найчастіше вживаних 35 послідовностей трьох слів найбільш уживаними виявилися послідовності, побудовані за моделлю: 1) article+noun+preposition (*the end of, the back of*); 2) preposition + article + noun (*for a moment, on the floor*); 3) conjunction + pronoun + verb (*and I was, and he was, but it was*). У німецькомовних художніх текстах найчастотнішою послідовністю трьох слів є *Ich weiß nicht*, яка відповідає моделі pronoun + verb + particle; характерною виявилась також модель preposition + article + + noun (*in der Hand, auf den Tisch*). Типовими моделями для послідовності трьох слів в україномовних художніх текстах є: 1) сполучник + займенник + частка (*і я не, що я не*); 2) займенник + частка + частка (*я вже не, це ж не*); 3) сполучник + прийменник + + займенник (*а в нас*). Спільною найчастотнішою моделлю для послідовності трьох слів з-поміж проаналізованих англо- та німецькомовних художніх текстів є "прийменник + артикль + іменник" (*in the dark, in der Nacht*). Важко простежити наявність спільної моделі для англо-, німецько- й україномовних художніх текстів серед перших 35 найчастіше вживаних послідовностей трьох слів. Імовірність появи однакових моделей послідовностей з трьох слів серед перших найчастіше вживаних 35 послідовностей в україномовних авторів є меншою, ніж в англо- та німецькомовних авторів.

Виявлено притаманні для кожного з авторів послідовності трьох слів: N. Gaimann *there was a*, J. Harris *for a moment*, M. Albom *the end of*, J. Rowling *out of the*, A. Friedrich *in der Nähe*, K. Gier *in der Zeit*, F. Schätzing *schüttelte den Kopf*, P. Süskind *und wenn er*, Ю. Андрухович *до того ж*, О. Забужко *так і не*, Л. Костенко *це ж не*, Ю. Покальчук *і я не*. З-поміж аналізованих авторів виділяється Ю. Покальчук, який використовує послідовності трьох слів, що можна розглядати як прості речення: *все буде добре, що з ним, що з тобою* (наприклад, "*Що з тобою? Ти такий страшний зараз!*" "Озерний вітер"). Виділяється послідовність *тато з мамою* (О. Забужко), яка не притаманна для англо-, німецько- й україномовних авторів. За умов відмінності авторських стилів написання твору, слід очікувати різний набір послідовностей уживання слів, які входять до числа перших, найчастіше вживаних. Малоймовірна поява найчастіше вживаних слів одного автора у списку найчастіше вживаних слів загального масиву текстів при зіставленні перших десятків найчастіше вживаних послідовностей трьох слів. Важливими для ідентифікації авторського стилю є слова (послідовності слів), які притаманні одному автору та відсутні у творах інших авторів, або ж слова, частота вживання яких суттєво відрізняється під час аналізу творів різних авторів.

Загальні результати дослідження дають змогу зробити такі **висновки**:

Теоретико-методологічні засади дисертаційного дослідження ґрунтуються на: 1) визначенні атрибуції як процесу упорядкування довільних текстових документів у групи за функціональним стилем, тематикою або автором за наперед заданими критеріями або такими, що визначаються під час цього процесу; 2) особливостях атрибуції наукових текстів, які є результатом роботи колективу авторів, що ставить високі вимоги до вибору оптимальних параметрів та методів атрибуції тексту; 3) класифікації вихідних параметрів атрибуції текстів на мовні (синтаксичні, лексичні, морфологічні, помилки) та позамовні (структурні дані); 4) виборі оптимальних методів опрацювання текстів; 5) застосуванні статистико-математичних методів, які забезпечують об'єктивність результатів атрибуції. Практична основа виконання завдань дисертації – розробка методів та програм для: а) опису рангово-частотного розподілу слів у текстах; б) підрахунку частот появи послідовності вживання одного і більше слів; в) обчислення дивергенції Кульбака-Лайблера.

Запропонована комплексна методика, яка передбачає формування репрезентативної вибірки текстів, вибір оптимальних параметрів тексту, визначення їх абсолютних частот, обробку статистичних даних, виявилась ефективною для здійснення стильової, тематичної та авторської атрибуції англо-, німецько- й україномовних наукових і художніх текстів.

Визначення функціонального стилю зіставлюваних текстів відбувалося на основі аналізу рангово-частотного розподілу слів. Запропонована модифікація функції Лавалетті дала змогу відтворити рангово-частотний розподіл слів у тексті, а її параметр *s* відповідав за розмежування текстів за функціональним стилем. Межі визначених інтерквантильних інтервалів параметра *s* для художніх і наукових текстів не перетинаються. Стрімкість спаду ймовірності появи слова у наукових текстах є меншою, ніж у художніх. Зіставлення рангово-частотного розподілу слів в англо-, німецько- й україномовних наукових та художніх текстах показало, що серед найчастіше вживаних слів переважають: 1) службові частини мови, загальнонаукові терміни (науковий текст); 2) службові частини мови та займенники, слова на позначення частин тіла та періодів дня (художній текст).

Із залученням методу одночасного моніторингу групування текстів і відповідних їм слів та послідовності вживання з одного слова успішно здійснено тематичну атрибуцію статей, що були опубліковані за матеріалами міжнародної конференції. Визначено ключові слова, притаманні кожному з тематичних розділів конференції. Показано, що одночасний аналіз статей та відповідних їм тез можна використати для оцінки відмінності текстів опублікованої статті від заявлених тез.

Аналіз вживання авторами послідовності з чотирьох слів із залученням методу одночасного моніторингу групування текстів та їх слів є оптимальним для авторської атрибуції наукових текстів. Зіставлення словників послідовностей чотирьох слів різних авторів показало, що у них дуже мала кількість спільних послідовностей чотирьох слів. Установлено, що стиль автора виявляється у вживанні характерних послідовностей чотирьох слів для: 1) опису спостережуваних об'єктів; 2) вираження припущень; 3) представлення та порівняння результатів.

Для авторської атрибуції художніх текстів визначення автора методом одночасного моніторингу групування текстів і відповідних їм слів сягає максимальної ефективності при аналізі послідовності трьох слів. В англо- та німецькомовних художніх текстах ймовірність вживання однакових послідовностей трьох слів є більша, ніж в україномовних художніх текстах. Виявлено спільну модель побудови послідовностей трьох слів у художніх англо- та німецькомовних текстах, що не є не властивою для україномовних текстів.

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 ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. 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.

2. 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. Сер. : Філологічні науки. – Луцьк : РВВ "Вежа" ВНУ імені Лесі Українки, 2008. – № 5. – С. 375–379.

3. 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.

4. 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.

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

6. 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.

7. 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.

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

9. 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 междунар. заочной науч.-практ. конф., (Москва, 05 марта 2013 г.) – М. : Изд-во “Международный центр науки и образования”, 2013. – С. 104–109.

10. 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.

11. 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.

АНОТАЦІЯ

Волошиновська І. А. Стильова, тематична й авторська атрибуція наукових і художніх текстів (на матеріалі англійської, німецької та української мов). – На правах рукопису.

Дисертація на здобуття наукового ступеня кандидата філологічних наук за спеціальністю 10.02.15 – загальне мовознавство. – Національний педагогічний університет імені М. П. Драгоманова. – Київ, 2013.

У дисертаційній праці запропоновано новий підхід до зіставно-типологічного вивчення стильової, тематичної та авторської атрибуції англо-, німецько- й україномовних наукових і художніх текстів; виявлено критерії розмежування англо-, німецько- й україномовних текстів наукового і художнього стилів на основі методу рангово-частотного розподілу слів; здійснено тематичну атрибуцію англомовних наукових текстів із залученням методу одночасного моніторингу групування текстів і відповідних їм слів та частоти вживання одного й більше слів; апробовано 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; установлено оптимальний розмір послідовності вживання одного та більше слів для авторської атрибуції англо-, німецько- й україномовних художніх текстів.

Ключові слова: атрибуція, науковий і художній стилі, рангово-частотний розподіл слів, метод одночасного моніторингу групування текстів і відповідних їм слів, *n*-грам слів.

АННОТАЦИЯ

Волошиновская И. А. Стилевая, тематическая и авторская атрибуция научных и художественных текстов (на материале английского, немецкого, украинского языков). – На правах рукописи.

Диссертация на соискание ученой степени кандидата филологических наук по специальности 10.02.15 – общее языкознание. – Национальный педагогический университет имени М. П. Драгоманова. – Киев, 2013.

В диссертационном исследовании разработана комплексная методика анализа стилевой, тематической и авторской атрибуции англо-, немецко- и украиноязычных научных и художественных текстов; выявлены критерии разграничения англо-, немецко- и украиноязычных текстов научного и художественного стилей; осуществлена тематическая атрибуция англоязычных научных текстов с привлечением метода одновременного мониторинга группирования текстов и соответствующих им слов; определена специфика авторской атрибуции научных и художественных текстов с привлечением параметра последовательности употребления одного и больше слов; разработаны программы для подсчета абсолютных и относительных частот параметра последовательности употребления одного и больше слов, вычисления дивергенции Кульбака-Лайблера.

Предложена модификация закона Ципфа для рангово-частотного распределения слов в тексте. Модифицированная функция Лавалетти позволяет описать распределение высокочастотных, среднечастотных и низкочастотных слов в тексте. Проведена стилевая атрибуция художественных и научных текстов на основе анализа числовых параметров модифицированной функции Лавалетти. Установлено, что различия между параметром *s*, который отвечает за художественный и научный стили, статистически значимы в пределах 2σ доверительного интервала. Определена общая тенденция, проявляющеяся в том, что стремительность спада вероятности появления слова в научных текстах является меньшей, чем в художественных текстах. В научных текстах среди чаще всего употребляемых слов преобладают служебные части речи, общенаучные термины; в художественных текстах – служебные части речи и местоимения, слова, обозначающие части тела и периоды дня.

Проведена тематическая и авторская атрибуция научных текстов с привлечением метода одновременного мониторинга группирования текстов и соответствующих им слов (метода анализа главных компонент) и параметра последовательности употребления слов (*n*-грамм слов). Этот метод позволяет пространственно разделить тексты в соответствии с их тематикой или автором, осуществить мониторинг соответствующей лексики. Тематическая и авторская атрибуция осуществляется путем перераспределения тематических и авторских признаков, изменяя размер *n* последовательности слов.

При анализе последовательности слов с n=1 можно осуществить тематическую атрибуцию текстов. Благодаря одновременному мониторингу группирования текстов и соответствующих им слов удалось выделить ключевые слова, которые характерны для разных тематических секций конференции. В диссертации показано, что одновременный анализ статей и соответствующих им

тезисов, аннотаций и заглавий как отдельных элементов выборки способствует лучшему разделению основных характеристик текстового массива и может быть использован для оценки отличия текстов опубликованной статьи от изначально заявленных тезисов.

Оптимальная эффективность авторской атрибуции достигнута для такого размера последовательности слов, когда объем словаря достигает своего максимального значения. Признаки авторского стиля больше проявляются, если увеличить размер последовательности слов до 4-х для научных текстов и до 3-х для художественных текстов. Для исследуемых авторов составлены словари наиболее употребляемых последовательностей четырех (научный текст) и трех слов (художественный текст). Показано, что стиль автора научного текста проявляется в употреблении характерных последовательностей из четырех слов для: 1) описания наблюдаемых объектов; 2) выражения предположений; 3) представления и сравнения результатов исследования. Сравнение словарей последовательностей четырех слов разных авторов в научных текстах показало, что у авторов научных текстов, по сравнению с авторами художественных текстов, очень малое количество общих последовательностей из четырех слов. Вероятность появления одинаковых последовательностей из трех слов среди первых чаще всего употребляемых последовательностей украиноязычных авторов является меньшей, чем в художественных текстах англо- и немецкоязычных авторов.

Ключевые слова: атрибуция, научный и художественный стили, рангово-частотное распределение слов, метод одновременного мониторинга группирования текстов и соответствующих им слов, *n*-грамм слов.

SUMMARY

Voloshynovska I. A. Stylistics, Thematic and Author Attribution of Scientific and Belletristic texts (based on the material of English, German, Ukrainian languages). – Manuscript.

Thesis for the Candidate degree in Philology, Speciality 10.02.15 – General Linguistics. – National Pedagogical Dragomanov University – Kyiv, 2013.

A new approach for the comparative typological studies of style, thematic and author attribution of English, German and Ukrainian scientific and belletristic texts is proposed at the present thesis; criteria for the determination of scientific and belletristic style in English, German and Ukrainian languages are depicted on the basis of rank-frequency word distribution method; thematic attribution of English scientific texts is performed using method of simultaneous monitoring grouping of texts and their corresponding words (principal component analysis) together with the parameter of n-gram words frequency; procedure of author attribution of English scientific texts combining the principal component analysis method with the parameter of *n*-gram words is tested; the optimal value of the *n*-gram words order is determined for author attribution of English, German and Ukrainian belletristic texts.

Key words: attribution, scientific and belletristic styles, rank-frequency word distribution, method of simultaneous monitoring grouping of texts and their corresponding words, *n*-gram words.

Підписано до друку 12.08.2013 р. Формат 60х90/16.
Ум. друк. арк. 0,9. Обл.-вид. арк. 0,9.
Тираж 100. Зам. 39.

«Видавництво “Науковий світ”»®
Свідоцтво ДК № 249 від 16.11.2000 р.
м. Київ, вул. Боженка, 23, оф. 414.
200-87-13, 200-87-15, 050-525-88-77
E-mail: nsvit@mail.ru