УДК 81'1:81'3:81'13:81'38(811.11-811.16)

Волошиновська І. А.,
Національний університет «Львівська політехніка», м. Львів

ЗІСТАВЛЕННЯ РАНГОВО-ЧАСТОТНОГО РОЗПОДІЛУ СЛІВ В АНГЛО-, НІМЕЦЬКО- ТА УКРАЇНОМОВНИХ НАУКОВИХ І ХУДОЖНІХ ТЕКСТАХ

У статті запропоновано новий підхід щодо атрибуції текстів різних функціональних стилів. Для цього використано модифіковану формулу Лавалетті. Увагу зосереджено на порівнянні англо-, німецько– та україномовних наукових та художніх текстів XXI століття.

Ключові слова: *Закон Ципфа, модифікована функція Лавалетті, рангово-частотний розподіл слів, інтерквантильний інтервал.*

В статье предложен новый подход к атрибуции текстов разных функциональных стилей. Для этого использовано модифицированную функцию Лавалетти. Внимание сосредоточено на англо-, немецко– и украиноязычных научных и художественных текстах XXI века.

Ключевые слова: *Закон Ципфа, модифицированная функция Лавалетти, рангово-частотное распределение слов, интерквантильный интервал.*

A new approach for text atrribution of different functional styles using modified Lavalettie's function is proposed. The attention is turned on the sceintific and belletristics texts of XXI century written in English, German and Ukrainian languages.

Key words: *Zipf law, modified Lavalette's function, words rank-frequenty distribution, interquartile distance.*

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, І. Popescu, J. Rudman) тощо. Метою дослідження є виявлення закономірностей і відмінностей стильової атрибуції англо-, німецько- та україномовних наукових і художніх текстів. Для цього проведено комплексний аналізу взаємозв'язку мовних та математичних підходів, що і визначає актуальність даної статті.

У цій статті, для зіставлення текстів за науковим та художнім стилем, запропоновано використовувати модифіковану формулу Лавалетті, яка є однією із сучасних модифікацій закону Ципфа [1, с. 9]:

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

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

На відміну від інших модифікацій закону Ципфа [2; 5; 7; 8], ця формула з високою точністю описує всі ділянки рангового розподілу слів у тексті, а саме: а) 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 $n[k+q]/[n-(k+q)+1]$. Для визначення ступеня якості апроксимації між розрахунковими та експериментальними значеннями ймовірності появи i-го слова у досліджуваному тексті використовують коефіцієнт детермінації R^2 [6, p. 58; 8, p. 718]. Коефіцієнт детермінації може набувати значень у таких межах: $0 \le R^2 \le 1$. Близькі до 1 значення коефіцієнта детермінації R^2 відповідають високій якості відтворення експериментальних даних апроксимаційною функцією.

Матеріалом дослідження є: а) *наукові* англомовні дисертаційні праці (R. Wegh, M. True, D. Talapin, L. Pieterson, Y. Kuzminykh, Ch. Bostedt), журнал «Physical Rewiew B», а також вибірка наукових статей чотирьох авторів проф. д-р. Pieter Dorenbos, проф. д-р. Andries Meijerink, д-р. Gregory Stryganyuk та проф. д-р. Georg Zimmerer; німецькомовні книжки (Wolfgang W. Osterhage Studium Generale Physik. Ein Rundflug von der klassischen bis zur modernen Physik, M. Komma Moderne Physik mit Maple: von Newton zu Feynman, R. Scharf Ausgezeichnete Physik); дисертаційні праці (A. Guesmann, T. Latz, C. Granzow, C. Rotsch), україномовні журнали («Український фізичний журнал», «Вісник ЛНУ, серія Фізична», «Фізика конденсованих високомолекулярних систем»), дисертаційні праці (В. Вістовський, А. Пушак, Г. Стриганюк, П. Савчин); б) *художні* тексти XXI століття: англомовні (M. Albom, H. Fielding, J. Harries, S. King, J. Rowling); німецькомовні (A. Friedrich, K. Gier, F. Shätzing, P. Süskind); україномовні (Л. Дереш, О. Забужко, Ю. Покальчук, В. Шкляр).

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 [4]. 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 [3, p. 18]. У цій статті текст нульового стилю не формувався, оскільки зіставлялися лише два стилі між собою. У таких випадках створення нульового стилю є недоцільним [3, p. 17].

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. Комплексний аналіз англо-, німецько– та україномовних наукових та художніх текстів показав, що математичний параметр *s* відповідає за функціональний стиль.

Для зіставлюваних мов середнє значення параметра *s* збільшується у напрямку українська – німецька – англійська мови. Параметр *s* є найбільший для англомовних наукових текстів (*s=1.00*) та найменший – для україномовних наукових текстів (*s=0.92*). Ця ж тенденція щодо зменшення параметра s у напрямку англійська (*s=1.12*) – німецька (*s=1.08*) – українська (*s=0.97*) мови спостерігається і для художніх текстів великого обсягу. На рис. 1 та 2 представлено криві розподілу слів в англо-, німецько– та україномовних текстах наукової та художньої літератури, апроксимованих модифікованою функцією Лавалетті для загального масиву всіх аналізованих текстів для кожної з розглянутих мов.

Метод довірчих інтервалів використано, щоб оцінити статистичну значимість отриманих параметрів апроксимації. Для оцінки розкиду параметра s кожного автора використаємо межі $S_{середнє}$ 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.

Для кожної з досліджуваних мов визначено межі інтерквантильних інтервалів параметра s рангово-ймовірнісного розподілу слів для стильової атрибуції наукових і художніх текстів (табл. 1). Межі інтерквантильних інтервалів не перетинаються для англомовних наукових текстів [0.85 – 1.01] та художні текстів [1.04 – 1.12]. Отже, між текстами існує значима статистична різниця, і 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.

Таблиця 1

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]

Отже, якщо параметр s будь-якого тексту відповідає значенню, яке вкладається в межі, характерні для наукового або художнього стилів у зіставлюваних мовах, то цей текст є або науковим, або художнім. Ці узагальнюючі дані значення параметра s для англо-, німецько та україномовних наукових та художніх текстів можна використовувати для стильової атрибуції текстів.

Проаналізовано рангово-частотні розподіли слів англо-, німецько– та україномовних наукових і художніх текстів. Зіставлення перших найчастіше вживаних 30 слів у досліджуваних мовах дозволило виявити:

1) 11 спільних для англо-, німецько- та україномовних наукових текстів найчастіше вживаних слів: i*n – 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 досліджуваних мов: *and, in, on, with, as – und, in, auf, als – і, в, на, з, як*.

Загальна тенденція для проаналізованих наукових англо-, німецько- та україномовних текстів проявляється у наявності серед перших 300 найчастіше вживаних слів загальнонаукових слів (*структура, метод, величина, система, параметр, результат*), дієслів мислення (*думати, вважати*) та пізнання (*вивчати, досліджувати, одержувати*), відсутності назв об'єктів дослідження (*хімічних речовин*).

У художньому тексті для трьох досліджуваних мов серед виділених перших 300 найчастіше вживаних слів переважають слова для означення частин тіла (наприклад, *рука, нога, очі*) та обставин часу (*день, ніч, рік*), іменники (*мати, батько, людина, Бог*), дієслова сприйняття (*бачити, дивитись, чути*).

Підсумовуючи все вище сказане, слід зазначити, що модифіковані формули, що забезпечують апроксимацію частотного розподілу слів у тексті, виявляють нові можливості у проведенні стильової атрибуції текстів. Запропонована модифікація формули Лавалетті $f(k; q; s; n)$ передбачає два параметри (*q* та *s*) для апроксимації рангово-ймовірнісного розподілу слів, які набувають характерних значень, залежно від стилю тексту – наукового або художнього. Показано, що для наукової літератури параметр *s* є меншим у порівнянні з параметром *s* художньої літератури.

Перспективами подальшого дослідження є: порівняння зміни значень апроксимаційних параметрів запропонованої модифікованої функції Лавалетті для одного наукового, або художнього тексту, який перекладений різними мовами; визначення та зіставлення меж зміни інтерквантильних інтервалів англо-, німецько- та україномовних текстів інших функціональних стилів; аналіз зміни кількісних показників параметра s залежно від століття написання текстів. Отже, такі комплексні філолого-математичні дослідження розкривають нові підходи щодо визначення функціонального стилю та є важливими для розв'язку задач інформаційного пошуку документів.

Література:

1. 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 а. – Вип. 16. – С. 334-339.
2. Орлов Ю. К. Модель частотной структуры лексики / Ю. К. Орлов // Исследования в области вычислительной лингвистики и лингвостатистики. – М. : МГУ, 1978. – С. 59-118.
3. 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.
4. Разинкина Н. М. Функциональная стилистика английского языка / Нина Марковна Разинкина. – М. : Высш. шк., 1989. – 182 с.
5. Тулдава Ю. Проблемы и методы квантитативно-системного исследования лексики / Юхан Тулдава. – Таллин : Валгус, Тартуский государственный Университет, 1987. – 204 с.
6. Kelih E. Graphemhäufigkeiten in Slawischen Sprachen: Stetige Modelle / E. Kelih // Glottometrics. – 2009. – Vol. 18. – P. 52-68.
7. Popescu Ioan-Iovitz. On a Zipf's law Extension to Impact Factors / Ioan-Iovitz Popescu // Glottometrics. – 2003. – Vol. 6. – P. 83-93.
8. Popescu I.-I. Zipf's law – another view / I.-I. Popescu, G. Altmann, R. Köhler // Quality and Quantity. – 2010. – Vol. 44. – P. 713-731/
9. 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.