УДК 658.012.011.56

Г.В. ГЛАДКІВСЬКА, фахівець 1-ої категорії Науково-дослідного центру правової інформатики Академії правових наук України

ДО ПИТАННЯ ДОСЛІДЖЕННЯ СТАТИСТИЧНИХ ХАРАКТЕРИСТИК УКРАЇНСЬКОЇ МОВИ

Анотація. *Досліджуються інформаційно-статистичні властивості законодавчих текстів на державній мові.*

Дана стаття є продовженням дослідження, розпочатого в [1], і охоплює деякі лінгвістичні питання української мови. В засобах масової інформації часто обговорюють питання затвердження російської мови як державної в Україні, та зазвичай розглядаються історичні чи демографічні аспекти питання. Не менш важливими є дослідження статистичних характеристик обох мов. В цьому аспекті широко відомі результати в основному щодо російської мови, наприклад [2, 3]. Тому була поставлена задача визначення інформаційно-статистичних параметрів саме української мови. Дослідження таких властивостей природної мови важливе для оптимізації людино-машинних інтерфейсів, що оперують з мовою. Вивчення цього питання допоможе зрозуміти, наприклад, який обсяг ресурсів є оптимальним для збереження текстів законів на державній мові за створення баз даних правової інформації, а також дослідити окремі фактори, що впливають на термінологічну ентропію текстів законодавчих актів. Як вказано в [4], саме термінологічна ентропія законодавчого тексту призводить до можливого неадекватного сприйняття зацікавленими особами “букви закону”.

Слід зазначити, що ґрунтовні дослідження інформаційно-статистичних властивостей українського тексту та короткий огляд аналогічних робіт наведено в [5]. В даній роботі досліджуються тексти літературних творів, а не законодавчих документів, і за модель тексту обрано ланцюги Маркова.

Наведемо необхідні для дослідження поняття теорії ймовірності та інформації. Однією з основних статистичних характеристик мови є ентропія – чисельна міра невизначеності ситуації. Цей показник вимірює інформацію, що міститься в певному тексті. Її обчислюють згідно з формулою [2]:

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i , \qquad (1)$$

де n – кількість можливих результатів випробування, p_i – імовірність результату випробування. В даному дослідженні будемо знаходити значення ентропій H_0, H_1, H_2, H_3, де $H_0 = \log_2 N$ (N – кількість літер алфавіту), H_1 дорівнює ентропії сукупності літер алфавіту, H_2 – ентропії сукупності всіх пар літер, H_3 – ентропії сукупності всіх трійок літер, утворених з алфавіту. Для української мови приймемо $N=32$ (аналогічно, як для російської мови в [2]), за умови, що приймемо за одну букви “Г” і “Ґ” та “І” і “Ї” і вважаємо пропуск буквою.

Очевидно, що невизначеність досягає найбільшого значення, коли результати випробування рівноймовірні. Наприклад, невизначеність при досліді, що полягає у виборі наугад однієї з букв українського алфавіту, дорівнює $H_0 = \log_2 32 = 5$.

Для знаходження значення ентропії H_1 для мови необхідно обчислити частоти появи кожної букви в певному тексті. Нижче вказано відомі дані про ентропію різних мов [2].

Мова	Англійська	Російська	Німецька	Французька	Іспанська
H_1	4,03	4,35	4,10	3,96	3,98

Для обчислення умовних ентропій H_2 , H_3 використано методику, аналогічну як в [2]. Нехай маємо дві системи X і Y, які в загальному випадку залежні. Припустимо, що система X набула значення x_i. Позначимо через $P(y_j / x_i)$ умовну ймовірність того, що система Y набуде стану y_j за умови, що система X перебуває у стані x_i:

$$P(y_j / x_i) = P(Y = y_j / X = x_i).$$

Визначимо умовну ентропію системи Y за умови, що система X перебуває у стані x_i:

$$H(y_j / x_i) = -\sum_j P(y / x_i) \log P(y_j / x_i) = M_{x_i}[-\log P(y / x_i)],$$

де M_{x_i} – оператор умовного математичного сподівання величини, що міститься в дужках, за умови $X \sim x_i$.

Умовна ентропія залежить від того стану x_i, якого набула система X; для одних станів вона більша, для інших менша. Визначимо середню або повну ентропію системи Y, ураховуючи, що система може набувати будь-яких значень. Для цього кожну умовну ентропію помножимо на ймовірність відповідного стану P_i, а далі сумуємо всі такі добутки:

$$H(Y / X) = \sum p_i H(Y / x_i). \qquad (2)$$

Значення ентропії H_0 і H_1, а також умовної ентропії 2-го та 3-го порядків (H_2 та H_3 відповідно) для російської та англійської мов наступні [2]:

Ентропія	H_0	H_1	H_2	H_3
Англійська мова	4,76	4,03	3,32	3,10
Російська мова	5,00	4,35	3,52	3,01

Такі результати були отримані при обчисленні ентропії для текстів, мова яких близька до “середньо літературної”.

Проведені дослідження показали, що ентропія наукових та законодавчих текстів дещо відрізняється від наведеної вище. Зокрема, було досліджено тексти, написані “мовою законів”: Конституція України [6], Конституція Російської Федерації [7], Кодекс адміністративного судочинства України [8], Закон України “Про Основні засади розвитку інформаційного суспільства в Україні на 2007 – 2015 роки” українською мовою та його офіційний переклад російською мову [9]. Вихідні тексти були піддані попередній обробці, а саме: було видалено спеціальні символи, знаки пунктуації, абзаци, залишено тільки по одному пропуску. Одержані результати дослідження на ентропію текстів вказаних документів наведено в таблиці нижче.

Назва тексту	H_1	H_2	H_3
Українська мова			
Кодекс адміністративного судочинства України	4,375	3,340	2,249
Конституція України	4,310	3,379	2,254
Закон України “Про Основні засади розвитку інформаційного суспільства в Україні на 2007 – 2015 роки”	4,473	3,365	2,168

Російська мова			
Конституция Российской Федерации	4,383	3,255	2,132
Закон Украины “Об Основных принципах развития информационного общества в Украине на 2007–2015 годы”	4,409	3,308	2,185

Розроблено алгоритми та відповідні програми для обчислення ентропії (звичайної та умовної) для текстів українською та російською мовами на мові C++ Builder. Зокрема, кожній букві присвоювався порядковий номер, розпізнавалась кожна буква, розраховувалась їх загальна кількість та імовірність появи кожної букви в тексті, на основі цього розраховувалась ентропія H_1 за формулою (1). Що стосується розрахунків ентропії H_2 та H_3, то для кожного тексту спершу обчислювались частоти дво- та трибуквених комбінацій і далі використовувалась формула (2). Для прикладу наведено скріншоти програми для визначення ентропії H_1 (Рис. 1).

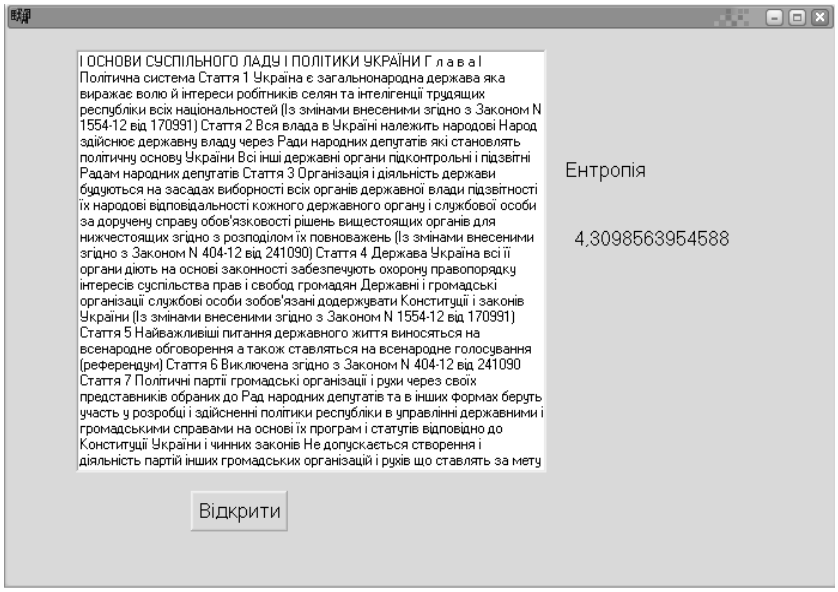

Рис. 1 – Фрагмент програми визначення ентропії H_1.

Ентропія наукових текстів російською мовою H_1 є дещо вищою від аналогічного показника для “середньо літературної” мови. Теж саме характерне для більшості текстів українською мовою. Це означає більший ступінь невизначеності наукових текстів, що можна пояснити частим вживанням поряд зі звичайними словами наукових термінів, які містять букви, що не властиві звичайній мові.

Умовна ентропія H_2 для досліджуваних текстів російською мовою є нижчою від аналогічного показника для текстів, мова яких близька до “середньо літературної”. Це свідчить про нижчий ступінь невизначеності, що можна пояснити наявністю широко поширених спеціальних термінів і виразів у спеціалізованих текстах. Умовна ентропія H_2 текстів українською мовою вища, ніж для російської мови. Це означає більшу невизначеність українських текстів, тобто тексти українською мовою містять більше інформації, ніж російські тексти, однакові за об’ємом.

Значення умовної ентропії H_3 для текстів, що написані мовою законів, є значно нижчим, ніж для “середньо літературної” мови, що свідчить про надмірність спеціалізованих текстів. При порівнянні цього показника для текстів українською та російською

мовами виявилось, що ентропія текстів українською мовою в середньому вища. Це свідчить про більшу інформативність текстів українською мовою.

При дослідженні надмірності мови важливо визначити, наскільки відрізняються по об’єму тексти, перекладені з однієї мови на іншу. При підрахунку кількості букв у тексті Закону України “Про Основні засади розвитку інформаційного суспільства в Україні на 2007-2015 роки” українською мовою (рис. 2) та офіційному перекладі цього тексту на російську мову (рис. 3) виявилось, що обсяг російського тексту більший на 4,9 %.

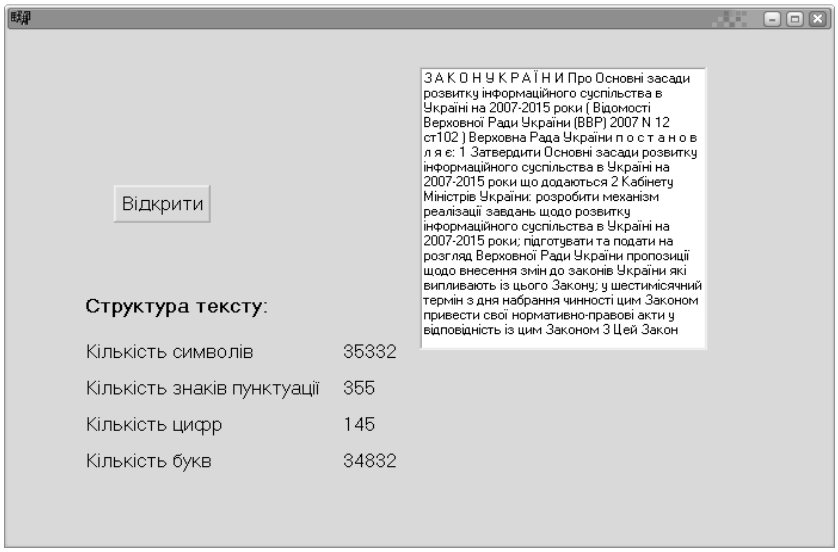

Рис. 2 – Фрагмент програми визначення структури українського тексту.

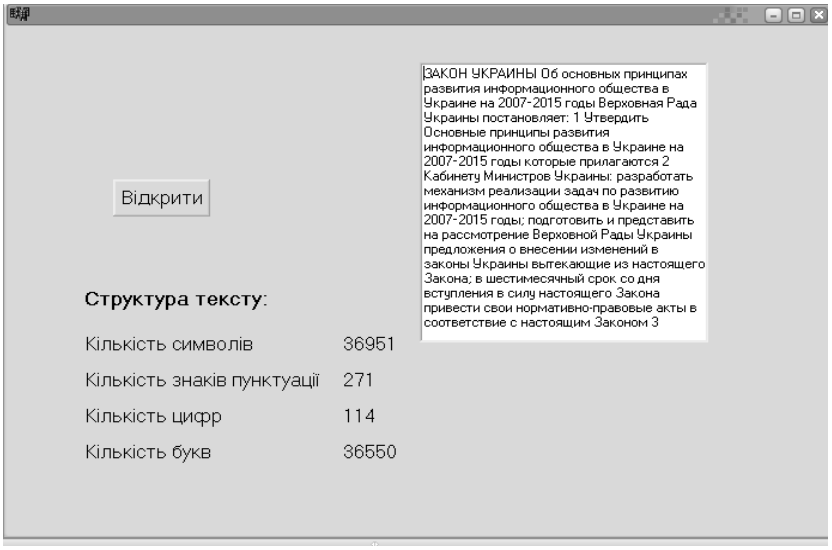

Рис. 3 – Фрагмент програми визначення структури російського тексту.

Можна припустити, що досліджені тексти містять однакову інформацію. Тому менший обсяг одного тексту порівняно з іншим свідчить про більшу економність української мови порівняно з російською. Це значить, що для зберігання на комп’ютері і друку текстів українською мовою буде витрачено менше ресурсів, а для їх перегляду при однаковому рівні знання обох мов витрачатиметься менше часу.

Безумовно, для вагоміших висновків та конкретних рекомендацій слід дослідити значно більше російських та українських текстів. Подальші дослідження у даному на-

прямі полягають в обчисленні умовних ентропій вищих порядків та у збільшенні обсягу вихідних законодавчих текстів, тоді вони будуть представляти значний інтерес як для лінгвістів, так і для юристів.

Використана література

1. Швець М. Я. До проекту Концепції державної мовної політики в Україні [Текст] // Правова інформатика. – 2007. – № 1(13). – С. 3-4.

2. Яглом И. М. Теория информации и лингвистика [Текст] / И. М. Яглом, Р. Л. Добрушин, А. М. Яглом // Вопросы языкознания. – 1960. – № 1. – С. 100-110.

3. Пиотровский Р. Г. Информационные измерения языка [Текст] / Р. Г. Пиотровский. – Л. : Наука, 1968. – С. 17-81.

4. Туранин В. Ю. Терминологическая энтропия законодательного текста [Текст] // Современное право. – 2005. – № 11. – С. 39-42.

5. Кригін М. Ю. Дослідження інформаційно-статистичних властивостей українського тексту / М. Ю. Кригін, В. А. Широков [Текст] // Математичні машини і системи. – № 1. – С. 120-127.

6. Конституція України : [від 28.06.1996 р. № 254к/96-ВР] // Відомості Верховної Ради України. – 1996. – № 30. – Ст. 141.

7. Конституция Российской Федерации. – Режим доступу : //www.constitution.ru

8. Кодекс адміністративного судочинства України : Закон України : [від 06.07.2005 р. № 2747-IV] // Відомості Верховної Ради України. – 2005. – № 35-36, № 37. – Ст. 446.

9. Про Основні засади розвитку інформаційного суспільства в Україні на 2007 – 2015 роки [Текст] : Закон України : [від 09.01.2007 р. № 537-V]. – К. : Видання Інтернет Асоціації України, 2008. – 57 с.

~~~~~~~~~~~~~ * * * ~~~~~~~~~~~~~
~~~~~~~~~~~~~