(people are thankful for god Himself). The analysis of the semantic features of the opposition «God – people» showed the positive attitude towards God and God's mercy and the negative attitude to human sin: *God is faithful and merciful; people are sinners who forget to thank God.* The research of key messages that can be derived from the homilies using the intent-analysis shows that the protestant priests pay a lot of attention to God's love and the fulfillment of people's needs; people, on the other hand, are sinners and often tend not to thank God for His deeds.

Key words: discourse, discourse analysis, religious communication, the English protestant sermon, sermon preached on Thanksgiving Day, intent analysis, thanksgiving to God.

Стаття надійшла до редколегії
18.04.2016 р.

УДК 811.161.2'33:519.25

Ігор Кульчицький,
Ігор Ліхнякевич

Ентропія одно- та двограм символів в україномовних текстах

У статті досліджено один з аспектів лінгвостатистики для україномовних текстів на семіотичному рівні. Оскільки мова як системний феномен виявляє свої об'єктивні властивості і в кількісних, і в якісних ознаках, то пояснити функціонування низки мовних категорій без застосування статистичних методів неможливо. У дослідженнях ідіолекту письменника статистичними методами важливою є проблема визначення авторського інваріанта, під яким розуміємо таку кількісну характеристику літературного тексту, яка однозначно характеризує одного автора або невелику групу «близьких» авторів та приймає суттєво відмінні значення для творів різних авторів. Для дослідження інваріантом вибрано ентропію одно- й двограм символів розширеної української абетки. Експерименти проведено на творах Василя Стефаника, Марка Черемшини, Софії Яблонської, текстах журналу «Літературно-науковий вісник» та газети «Сільський господар». Виявилося, що така ентропія інваріантом бути не може, тому потрібно продовжувати вивчати цей аспект.

Ключові слова: квантитативна лінгвістика, ентропія, статистичні методи, ідіолект, одинограми, двограми, статистичний інваріант, атрибуція текстів.

Постановка наукової проблеми та її значення. Центральна ідея сучасного мовознавства – системність мовних фактів як на фонетичному та граматичному, так і на лексичному рівнях [13, с. 6]. Оскільки мова виявляє свої об'єктивні властивості і в кількісних, і в якісних ознаках, то пояснити функціонування низки мовних категорій без застосування статистичних методів неможливо [13, с. 94]. Ще І. О. Бодуен де Куртене зазначав: «Потрібно частіше використовувати в мовознавстві кількісне, математичне мислення, і таким чином наближаючи його все більше до точних наук» [2]. Якщо експериментальна психолінгвістика, фонетика, соціолінгвістика вже тривалий час застосовують методи точних наук, то семантика, синтаксис і морфологія почали використовувати статистичні методи лише в останні десятиліття XX ст. Стрімкий розвиток застосування статистичних методів у лінгвістиці зумовили такі чинники: опрацювання емпіричних даних вимагало більшого використання статистичних інструментів; з'явилися нові погляди на дослідження мовних явищ, які залучають теорію ймовірностей, теорію інформації, статистичне моделювання; розвиток комп'ютерної й корпусної лінгвістики. Як результат – кількісні методи в лінгвістиці стали стійкою методологічною основою [25].

Застосування останніх у лінгвістиці значно розширює та модифікує наші знання як про саму мовну систему, так і про можливості її функціонування [9]. Є три основні напрями їх використання [12]: отримання різноманітних кількісних відомостей; побудова лінгвістичних моделей із використанням методів теорії ймовірностей, зазвичай із залученням даних попереднього напряму; статистична перевірка гіпотез про ті чи інші явища.

Аналіз досліджень цієї проблеми. Першими спробами застосування статистики в лінгвістичних дослідженнях були частотні словники, які подають списки слів із частотою їх уживання в певному тексті. Першим таким словником уважають словник Ф. Кединга – «Частотний словник німецької мови», (1898 р.) [13, с. 97].

Прикладом застосування лінгвостатистичних методів є дослідження лексики певного автора, твору чи жанру, адже кількісні характеристики тексту дають змогу встановити не лише склад лексики, а й співвідношення використання її різних пластів, співвідношення слів, які трапляються рідко та часто й т. ін. Статистичні методи дослідження уможливлюють дослідження всього складу, так би мовити, нейтральної лексики, яка є показником різноманітності чи одноманітності словника письменника [13, с. 96]. Сюди відносять і проблеми атрибуції текстів, методи якої дають змогу дослідити текст на п'яти рівнях [1]:

- символьний (семіотичний) – досліджує частотність символів тексту, особливості застосування пунктуаційних знаків тощо;
- орфографічний – виявляє характерні помилки в написанні слів;
- синтаксичний – установлює особливості побудови речень, перевагу тих чи інших мовних конструкцій тощо;
- лексико-фразеологічний – визначає словниковий запас автора, особливості вживання слів та виразів, схильність до застосування рідковживаних та іншомовних слів, діалектизмів, архаїзмів, неологізмів, професіоналізмів, арготизмів, звичок використання фразеологізмів, паремій тощо;
- стилістичний – виявляє жанр, загальну структуру тексту, для літературних творів – сюжет, характерні зображальні засоби (метафора, іронія, алегорія, гіпербола, порівняння), стилістичні фігури (градація, антитеза тощо), інші характерні мовні прийоми.

Хоча використання математичних методів у лінгвістиці передбачили ще Ф. де Соссюр та І. О. Бодуен де Куртене, фактично застосовувати їх почали із середини минулого століття [9, с. 95]. Дослідження в цьому напрямі проводили [3; 4] такі зарубіжні вчені, як Ґабріель Альтман (Gabriel Altmann), Рейнгард Кйолер (Reinhard Köhler) (Німеччина); Петер Ґжибек (Peter Grzybek), Еммеріх Келіх (Emmerich Kelih) (Австрія); Гейза Віммер (Geiza Wimmer) (Словаччина); Адам Павловскі (Adam Pawłowski), Ядвіґа Самбор (Jadwiga Sambor) (Польща); Юхан Тулдава (Естонія); Раймунд Пйотровський, Анатолій Шайкевич (Росія) та ін. В Україні квантитативними дослідженнями мовних явищ займаються Володимир Широков [20], Максим Кригін [10], Валентина Перебийніс [17], Соломія Бук [3; 4] й ін.

Мета та завдання статті. Під час дослідження ідіолекту письменників статистичними методами важливою є проблема визначення авторських інваріантів. Дотримуючись принципів, поданих у [18], під авторським інваріантом розумітимемо таку кількісну характеристику (параметр) літературного тексту, який:

- однозначно характеризує своєю поведінкою одного автора або невелику групу «близьких» авторів;
- приймає суттєво різні значення для творів різних авторів.

Такий інваріант, за словами науковців [18], повинен мати такі характеристики:

- він повинен бути таким, щоби автор не міг його сильно контролювати;
- він обов'язково зберігає практично постійне значення для творів одного автора;
- для різних груп авторів він має бути суттєво різним.

Ураховуючи думки, висловлені у [14; 15], таким інваріантом може слугувати ентропія символів та двограм у творах письменників. Під ентропією символа чи двограми розуміємо кількість інформації, яку містить відповідно, символ чи двограма [15]. **Мета дослідження** – з'ясувати, чи справді ентропія одно- та двограм символів україномовних текстів може бути інваріантом авторського стилю.

Організація та матеріали дослідження. Матеріалом дослідження обрано художні твори Василя Стефаника [5; 6; 7; 8], Марка Черемшини [19], Софії Яблонської [21; 22; 23; 24], журнал «Літературно-науковий вісник» [11] та газету «Сільський господар» [16]. Це зумовлено наявністю вичитаних електронних форм вищевказаних текстів. Перед дослідженням усі тексти за допомогою програм мовою Python перевірено на коректність кодів символів і здійснено необхідне коректування.

Виклад основного матеріалу та обґрунтування отриманих результатів дослідження. Для дослідження обрані тексти розділено за автором чи належністю до періодичного видання на п'ять масивів текстів: твори В. Стефаника, М. Черемшини, С. Яблонської, журнал «Літературно-науковий вісник» і газета «Сільський господар». За принципами, поданими у [20], у текстах залишено лише букви української абетки, пробіл, апостроф і дефіс (розширений алфавіт української мови). Для зручності обрахунків текст розділено на абзаци розміром у 500 символів із точністю до слова. Загальну характеристику масивів текстів відображено в табл. 1.

Таблиця 1

Загальна характеристика масивів творів

Об'єкт аналізу	Кількість абзаців	Кількість символів
Твори С. Яблонської	1672	841 071
Твори Марка Черемшини	928	466 642
Твори ЛНВ	1488	748 349
Твори Василя Стефаника	607	305 228
Газета «Сільський господар»	1824	918 013

Ентропію обраховували за формулою, поданою в [14; 8]:

$$H = -\sum_{i=1}^{36} p_i \log_2 p_i .$$

Дослідження складалося з таких кроків.

Крок 1. Для кожного масиву тексту обраховано кількість символів розширеної абетки, їх частоту та ентропію.

Крок 2. У кожному масиві обрано послідовно десять сегментів з інтервалом в 0,02 довжини тексту. Для кожного з них обраховано кількість символів розширеної абетки, їх частоту та ентропію.

Крок 3. У кожному масиві обрано десять випадкових сегментів, які співмірні за розміром із послідовними сегментами. Для кожного сегмента обраховано кількість символів розширеної абетки, їх частоту та ентропію.

Узагальнені результати подано в табл. 2–6.

Таблиця 2

Дослідження розподілу символів у творах Василя Стефаника

	Кількість		Ентропія середня
	абзаців	символів	
Весь текст	607	305 228	4,407033000
Послідовний сегмент (середнє)	48	24 123	4,406714000
Випадковий сегмент (середнє)	60	30 162	4,405421000

Таблиця 3

Дослідження розподілу символів у творах Марка Черемшини

	Кількість		Ентропія середня
	абзаців	символів	
Весь текст	928	466 642	4,476512000
Послідовний сегмент (середнє)	74	37 217	4,478810000
Випадковий сегмент (середнє)	92	46 259	4,475264000

Таблиця 4

Дослідження розподілу символів у творах Софії Яблонської

	Кількість		Ентропія середня
	абзаців	символів	
Весь текст	1672	841 071	4,489471171
Послідовний сегмент (середнє)	134	67 397	4,490578646
Випадковий сегмент (середнє)	167	84 013	4,488519944

Таблиця 5

Дослідження розподілу символів у літературно-науковому віснику

	Кількість		Ентропія середня
	абзаців	**символів**	
Весь текст	1488	748 349	4,507923068
Послідовний сегмент (середнє)	119	59 857	4,504355627
Випадковий сегмент (середнє)	148	74 389	4,506743073

Таблиця 6

Дослідження розподілу символів у газеті «Сільський господар»

	Кількість		Ентропія середня
	абзаців	**символів**	
Весь текст	1824	918 013	4,502749000
Послідовний сегмент (середнє)	146	73 480	4,500282000
Випадковий сегмент (середнє)	182	91 575	4,502952000

Крок 4. Кроки 2 і 3 повторено для двограм символів усіх масивів творів. Узагальнені результати відображено в табл. 7–11.

Таблиця 7

Дослідження розподілу двограм у творах Василя Стефаника

	Кількість		Ентропія середня
	абзаців	**Символів**	
Весь текст	607	305 227	7,908782556
Послідовний сегмент (середнє)	48	23 791	7,842817124
Випадковий сегмент (середнє)	60	29 824	7,857341808

Таблиця 8

Дослідження розподілу двограм у творах Марка Черемшини

	Кількість		Ентропія середня
	абзаців	**Символів**	
Весь текст	928	466 641	8,061698
Послідовний сегмент (середнє)	74	36 859	8,022614
Випадковий сегмент (середнє)	92	45 910	8,025142

Таблиця 9

Дослідження розподілу двограм у творах Софії Яблонської

	Кількість		Ентропія середня
	абзаців	**символів**	
Весь текст	1672	841 070	8,050741
Послідовний сегмент (середнє)	134	67 047	8,027955
Випадковий сегмент (середнє)	167	83 622	8,027907

Таблиця 10

Дослідження розподілу двограм у літературно-науковому віснику

	Кількість		Ентропія середня
	абзаців	символів	
Весь текст	1488	748 348	8,146356
Послідовний сегмент (середнє)	119	59 465	8,104411
Випадковий сегмент (середнє)	148	74 024	8,120853

Таблиця 11

Дослідження розподілу двограм у газеті «Сільський господар»

	Кількість		Ентропія середня
	абзаців	символів	
Весь текст	1824	918 012	8,082847
Послідовний сегмент (середнє)	146	73 128	8,053996
Випадковий сегмент (середнє)	182	91 242	8,062806

Зведені результати обрахунку ентропії одно- та двограм символів у сегментах масивів тексту відображено на рис. 1–2. Сегменти 1–10 – це послідовні сегменти, а 11–20 – випадково вибрані.

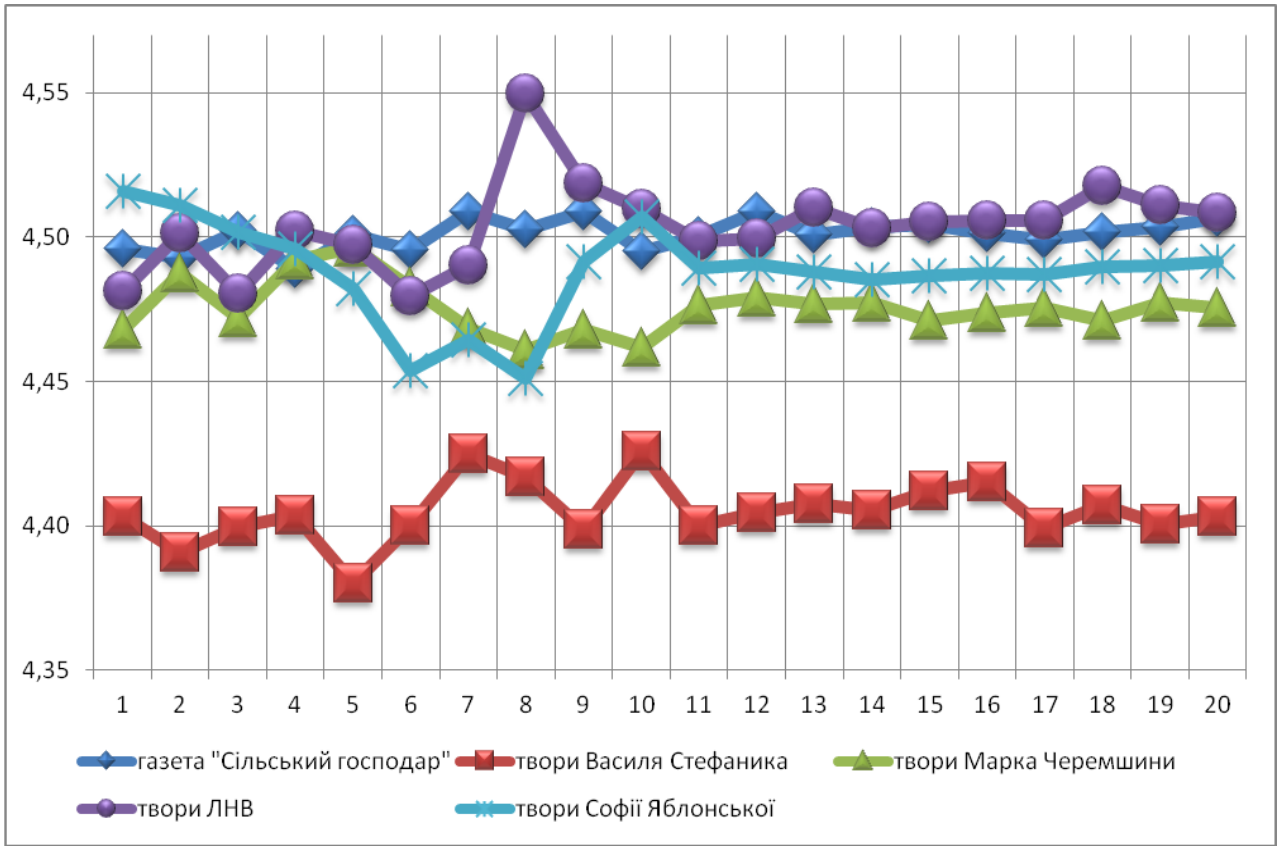

Рис. 1. *Значення ентропії символів у масивах текстів*

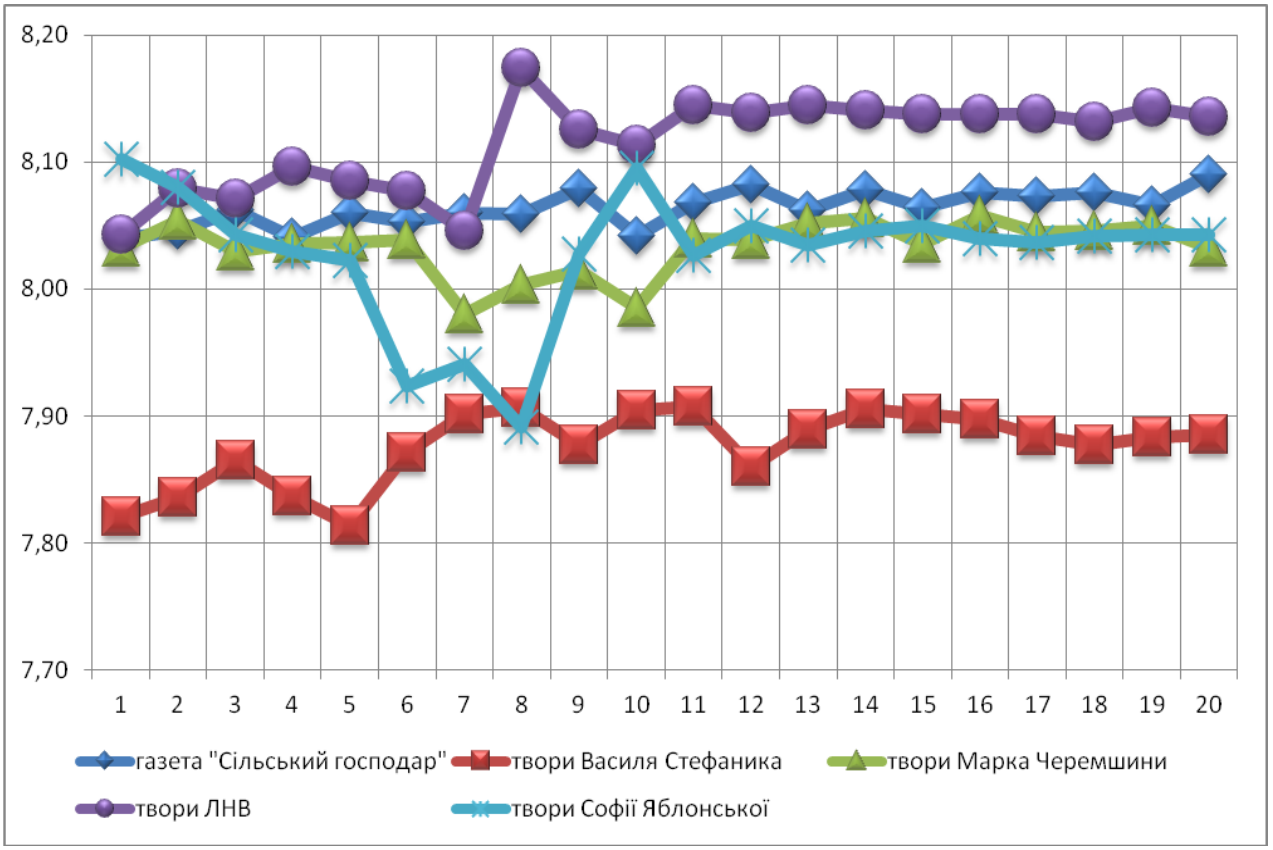

Рис. 2. *Значення ентропії двограм символів у масивах текстів*

Аналізуючи отримані результати, бачимо, що ентропія символів у всіх масивах текстів, за винятком творів Василя Стефаника, має приблизно однакове значення. Для випадково обраних сегментів значення ентропії достатньо стабільне.

Висновки та перспективи подальших досліджень. У межах вибраних масивів творів ні ентропія символів, ні ентропія двограм символів не може бути авторським інваріантом.

Надійна відмінність значень ентропії у творах Василя Стефаника зайвий раз підтверджує його унікальність як письменника.

Під час квантитативних досліджень краще робити вибірки випадковим, а не послідовним чином.

Дослідження проведено лише на множині символів розширеного українського алфавіту та на множині його двограм. Їх потрібно продовжити для триграм і т. ін.

Отримані результати стосуються лише творів Василя Стефаника, Леся Мартовича, Марка Черемшини, Софії Яблонської, журналу «Літературно-науковий вісник» та газети «Сільський господар». Для узагальнених висновків стосовно всієї української мови потрібні подальші дослідження на матеріалі творів інших письменників.

Джерела та література

1. Батура Т. В. Формальные методы определения авторства текстов / Т. В. Батура // Вестник НГУ. – Серия : «Информационные технологии». – Т. 10, вып. 4. – С. 81–94.
2. Бодуэн де Куртенэ И. А. Избранные труды по общему языкознанию : в 2. т. / И. А. Бодуэн де Куртенэ. – М. : [б. и.], 1963.
3. Бук Соломія Лінгвостатистичний опис «Не спитавши броду» Івана Франка / Соломія Бук [Електронний ресурс]. – Режим доступу : http://www.lnu.edu.ua/faculty/ Philol/www/visnyk/55_2011/55_2011_Buk.pdf
4. Бук Соломія Сучасні методи дослідження мови письменника у слов'янознавстві / Соломія Бук [Електронний ресурс]. – Режим доступу : http://www.lnu.edu.ua/page/n61/010.pdf
5. Стефаник В. Межа / В. Стефаник // «Літературно-науковий вісник». – Т. 92, кн. 2. – Львів, 1927. – С. 97–98
6. Стефаник В. Твори / передм. В. Коряка. До друку виготував Ів. Лизанівський. – 3-тє вид. // ДВУ. – 1929. – С. 94–95.
7. Стефаник В. Твори / Василь Стефаник ; з дереворитами В. Касіяна і М. Бутовича. – Львів : Друк. Вид. спілки «Діло», 1933. – 222 с.
8. Стефаник В. Шкільник / В. Стефаник // «Рідна школа» – № 1. – Львів, 1932. – С. 2–4.

9. Гладкий А. В. «Математические методы изучения естественных языков». Математическая логика, теория алгоритмов и теория множеств : сб. работ, посвящ. акад. Петру Сергеевичу Новикову к его семидесятилетию / А. В. Гладкий // Труды МИАН СССР. – 133. – 1973. – С. 95–108.
10. Кригін М. Ю. Дослідження інформаційно-статистичних властивостей українського тексту / М. Ю. Кригін, В. А. Широков // Математические машины и системы. – ІПММС НАНУ. – 2000. – № 1. – С. 120–127.
11. Літературно-науковий вісник. – LXXXII : Кн. I–IV. – 1924. – 368 с.
12. Марков А. А. Пример статистического исследования над текстом «Евгения Онегина», иллюстрирующий связь испытаний в цеп / А. А. Марков // Известия. Имп. акад. наук. – Серия 6. – 1913. – № 3. – С. 153–162.
13. Методы изучения лексики / под ред. А. Е. Супруна. – Минск : Изд-во БГУ им. В. И. Ленина, 1975. – 232 с.
14. Пиотровский Р. Г. Информационные измерения язика / Р. Г. Пиотровский.– Л. : [б. и.], 1968. – 116 с.
15. Пиотровский Р. Г. Математическая лингвистика / Р. Г. Пиотровский, К. Б. Бектаев, А. А. Пиотровская.– М. : [б. и.], 1977.– 383 с.
16. Сільський господар: орган кураєвого товариства господарського «Сільський господар».– Львів, 1926.– Число 1–12.
17. Статистичні параметри стилів / за ред. Перебийніс В. С. – К. : Наук. думка, 1967. – 260 с.
18. Фоменко В. П. Авторский инвариант русских литературных текстов / В. П. Фоменко, Т. Г. Фоменко // Новая хронология Греции: Античность в Средневековье. – М. : МГУ, 1995. – 422 с.
19. Черемшина Марко. Твори : у 3 т. / Марко Черемшина ; за ред. Є. Пеленського. – Львів : Ізмарагд, 1937.– Т. 1. – 206 с. – Т. 2. – 245 с. – Т. 3. –231 с.
20. Широков В. А. Інформаційна теорія лексикографічних систем / В. А. Широков. – Киев : Довіра, 1998. – 331 с.
21. Яблонська С. І. Далекі обрії. – Т. 1 / Софія Іванівна Яблонська. – Львів : [б. в.], 1939. – 183 с. – (Бібліотека «Діла»).
22. Яблонська С. І. Далекі обрії. – Т. 2 / Софія Іванівна Яблонська. – Львів : [б. в.], 1939. – 169 с. – (Бібліотека «Діла»).
23. Яблонська С. І. Книга про батька / Софія Іванівна Яблонська. – Едмонтон ; Париж : Слово, 1977. – 237 с. – (Об'єднання українських письменників у Канаді).
24. Яблонська С. І. Чар Марока / Софія Іванівна Яблонська. – Львів : Книгарня наук. т-ва ім. Шевченка, 1932. – 83 с.
25. Gries Th. S. Quantitative methods in linguistics / Th. S. Gries. – CA : University of California, 1973. – 13 p.

Кульчицкий Игорь, Лихнякевич Игорь. Энтропия одно- и двограмм символов в украиноязычных текстах. В статье исследуется один из аспектов лингвостатистики для украиноязычных текстов на семиотическом уровне. Поскольку язык как системный феномен выявляет свои объективные свойства, как в количественных, так и в качественных признаках, то объяснить функционирование ряда языковых категорий без использования статистических методов невозможно. При исследовании идиолекта писателей статистическими методами важнейшей является проблема определения авторских инвариантов, под которыми подразумеваем такую количественную характеристику литературного текста, которая однозначно характеризирует своим поведением одного автора или небольшую группу «близких» авторов и принимает различные значения для произведений других авторов. Для исследования инварианта избрана энтропия одно- и двограмм символов расширенного украинского алфавита. Эксперименты проведены на произведениях Василия Стефаника, Марко Черемшины, Софии Яблонской, текстах журнала «Литературно-науковый висник» и газеты «Сильськый господар». Оказалось, что такая энтропия инвариантом быть не может, поэтому необходимо продолжить изучение этого аспекта.

Ключевые слова: квантитативная лингвистика, энтропия, статистические методы, идиолект, одногаммы, двограммы, статистический инвариант, атрибуция текстов.

Kulchytskyy Ihor, Likhniakevych Ihor. Entropy of Unigram and Bigram Symbols in Ukrainian-language Texts. The article is devoted to one of the aspects of lingual statistics for Ukrainian-language texts on the semiotic level. Language is a systemic phenomenon, which reveals its objective features of both quantitative and qualitative characteristics. It is impossible to explain the functioning of several speech categories without the use of statistical methods. The major problem of the writers' idiolect study applying the statistical methods is to determine the writers' invariants that are defined as such quantitative characterization of literary text which uniquely characterizes the behavior of one author or a small group of authors and takes significantly different values for the works of other authors. For this study the invariant of the entropy of unigram and bigram symbols of the extended Ukrainian alphabet has been chosen. The linguistic researches have been conducted on the works of Vasyl Stefanyk, Marko Cheremshyna, Sofia Yablonska, *Literary and Scientific Journal* articles and the newspaper *Silskiy Gospodar*. It turned out that this entropy could not serve as an invariant. Thus, further study of this aspect is required.

Key words: quantitative linguistics, entropy, statistical methods, idiolect, unigram, bigram, statistical invariant, text attribution.

Стаття надійшла до редколегії
25.03.2016 р.