ISSN 2224-087X. Електроніка та інформаційні технології. 2017. Випуск 7. С. 125–137
Electronics and information technologies. 2017. Issue 7. P. 125–137

УДК 004.6, 004.9, 538.9

ЗІСТАВНИЙ АНАЛІЗ СТАТИСТИЧНИХ ВЛАСТИВОСТЕЙ СХІДНОСЛОВ'ЯНСЬКИХ ТЕКСТІВ

О. Кушнір[1], Т. Стрипко[1], В. Таранець[2],
Л. Кушнір[1], С. Вельгош[1]

[1]*Львівський національний університет імені Івана Франка,*
вул. Ген. Тарнавського, 107, 79017, м. Львів, Україна
o_kushnir@franko.lviv.ua

[2]*Міжнародний гуманітарний університет,*
вул. Фонтанська Дорога, 33, 65009, м. Одеса, Україна

Розроблено програмне забезпечення для дослідження низки статистичних властивостей текстів, написаних східнослов'янськими мовами, найперше українською та російською. Реалізовано комп'ютеризовану методику фонетичного аналізу та складоподілу. Виконано статистичну обробку текстів із трьох корпусів художньої літератури, одержано відповідні рангові залежності та описано їх теоретичними виразами. Шляхом зіставного аналізу з'ясовано схожі та відмінні риси фонетики і фоніки української та російської мов, а також їхній історичний розвиток.

Ключові слова: статистична лінгвістика, компаративна лінгвістика, корпусна лінгвістика, східнослов'янські мови, українська мова, російська мова.

Вступ. Статистичні властивості текстів природними мовами найчастіше вивчають на лінгвістичних рівнях графем [1–5], лексем [6, 7] або відповідних n-грам [8], іноді – на рівнях речень [7, 9, 10]. Водночас, статистику фонем і силабем у комп'ютерній лінгвістиці досліджено менше (див. [11–13]). Одна з причин – ускладненість, неоднозначність і значні технічні труднощі реалізації базових лінгвістичних алгоритмів виділення фонем і силабем. З іншого боку, інтерес до таких досліджень пов'язаний хоча би з тим, що відкриті склади історично відігравали значну роль у давніх слов'янських мовах, де домінувала тенденція, яку умовно називають «законом відкритого складу». Додатковим наслідком цієї тенденції була порівняно висока «прозорість» мови, або низький консонантний коефіцієнт $k = F_c/F_v$, де F_c і F_v – відповідно кількості (абсолютні частоти) приголосних «*c*» і голосних «*v*» у тексті (див. [14–16]). Тому дослідження фоностатистики та статистики силабем за стандартними методами корпусної лінгвістики можуть виявитися корисними в плані з'ясування закономірностей еволюції природних мов і їхньої топологічної близькості.

Мета цієї роботи – статистичні дослідження частот звуків, характеру і структури складів, а також консонантного коефіцієнта для корпусів українських, російських і давньоруських текстів. На підставі відповідних статистичних даних ми порівнювали близькість сучасних української та російської мов до давньоруської мови.

Методика досліджень. Українські, російські та давньоруські корпуси налічували відповідно 35, 37 і 5 художніх текстів. Сукупні обсяги корпусів текстів, виражені в кількості букв, наведено в табл. 1. Із давньоруських ми аналізували тексти, автентичність і орієнтовний час написання яких не викликав сумнівів у широких колах науковців. Найвідоміші серед них – «Повість временних літ» і «Слово о полку Ігоревім». Для цих текстів, що датуються XI століттям або дещо пізнішим історичним періодом, іноді вживають категоричніший (хоча й не прийнятий усіма дослідниками) термін «давньоукраїнські».

Першою складовою нашої методики є фонетичний аналіз, за яким кожну графему або сполучення графем заміняють на їхні фонемні відповідники. Наприклад слово «*моя*» в режимі відображення фонетики представляють як «*moja*», де «*j*» – це умовне позначення для т. зв. слабкого звуку «*й*» (див. табл. 2). Відповідні фонетичні правила мають свої особливості для кожної з трьох досліджених мов. Найскладнішим є фонетичний аналіз російської мови, оскільки вона вирізняється наймасштабнішими редукційними та асиміляційними змінами та значно відхиляється від простого «алфавітного» принципу вимови, переважно притаманного сучасній українській та багатьом іншим слов'янським мовам. З іншого боку, помірніше виражені процеси асиміляції та редукції притаманні також українській і давньоруській мовам.

Таблиця 1

Деякі статистичні параметри корпусів текстів досліджених мов (С.К.В. –середньоквадратичне відхилення параметра по корпусу; див. також пояснення в тексті)

Мова	Українська		Російська		Давньоруська	
Параметр	Середнє	С.К.В.	Середнє	С.К.В.	Середнє	С.К.В.
Сумарна довжина текстів L_l	$1{,}174\times10^7$	–	$1{,}150\times10^7$	–	$5{,}828\times10^5$	–
Сумарна кількість складів N_s	$5{,}094\times10^6$	–	$4{,}912\times10^6$	–	$2{,}728\times10^5$	–
Кількість букв на один склад $N_{ls} = L_l/N_s$	2,305	–	2,342	–	2,136	–
Відносна частота приголосних f_c	0,557	0,003	0,576	0,004	0,530	0,005
Відносна частота голосних f_v	0,443	0,003	0,424	0,004	0,470	0,006
Консонантний коефіцієнт $k = f_c/f_v$	1,257	0,014	1,358	0,021	1,128	0,025
Відносна частота відкритих складів f_{op}	0,771	0,009	0,731	0,016	0,901	0,021
Відносна частота закритих складів f_{cl}	0,229	0,009	0,269	0,016	0,099	0,021
Відношення частот відкритих і закритих складів $K = f_{op}/f_{cl}$	3,367	0,179	2,717	0,202	9,101	1,693

Оскільки найпершим завданням цього дослідження були розрахунки консонантного коефіцієнта та характеристик складів, ми не враховували тих фонетичних редукцій і асиміляцій, які не приводили до змін сумарної кількості голосних і приголосних, а лише до перерозподілу між кількостями різних голосних або різних приголосних. Скажімо, редукція «*солнце–сонце*» в російській вимові в цьому плані принципова, оскільки її не-

врахування призвело би до появи «зайвого» приголосного звуку, а фонетична заміна «*сосна–сасна*» в російській мові менш принципова, оскільки тут заміна одного голосного на інший не впливає на консонантний коефіцієнт. Зрештою, повномасштабний аналіз усіх фонетичних явищ вкрай утруднений на практиці. Він потребує враховувати, як мінімум, позицію букви в слові, наявність і властивості кількох попередніх і наступних букв, наголос у слові, приналежність букви до кореня, афікса або закінчення, з'ясування, якою частиною мови є дане слово, і навіть темпу та «сили» мовлення. Відповідно, автоматизований фонетичний розбір потребує строгої алгоритмізації принаймні морфеміки, словотвору, морфології, синтаксису та акцентології. До того ж, відповідні правила часто сформульовані не строго. Отже, маємо викличне завдання з переднього краю галузі обробки природних мов, на повне розв'язання якого ми не претендували. Принаймні автори не натрапляли в літературі на повідомлення про комп'ютерні програми, здатні виконувати повний фонетичний розбір в українській або російській мовах.

Окрім самостійного інтересу до статистики фонем, фонетичний аналіз є необхідним кроком на шляху реалізації наступного етапу складоподілу. Принципи складоподілу базуються на акустичних або артикуляційних властивостях приголосних – їхній «силі» (див. [17]). За теорією сонорності, гучність звуків спадає в ряді голосних (найгучніших), сонорних приголосних, а далі дзвінких і глухих приголосних. Голосні формують вершину складу. Якщо в слові трапляється послідовність кількох приголосних, то межу між складами проводять там, де гучність звуків у складі перестає зростати. У рамках якісно схожого, проте дещо відмінного «енергетичного» підходу до теорії мовлення, замість акустичних характеристик приголосних аналізують їхню «власну енергетичність» або «напруженість» вимови [17] (див. табл. 2).

Таблиця 2

Рангова таблиця умовної сили приголосних звуків української мови в рамках енергетичного підходу

№ з/п	Звук і його умовне позначення	Сила звуку
1	u (сильний «*в*»), Y (сильний «*й*»)	7
2	р, л, w (середній «*в*»)	6
3	м, н	5
4	б, д, ґ	4
5	п, т, к	3
6	v (слабкий «*в*»), j (слабкий «*й*»), г, з, ж, d («*дз*»), z («*дж*»)	2
7	ф, с, х, ц, ч, ш	1

Коротко підсумуємо алгоритм розбивання слів на склади:

- на початку аналізу визначаємо кількість голосних звуків у слові; якщо голосний єдиний, то слово односкладове;
- якщо в слові більше ніж однин голосний звук, то можуть трапитися такі випадки: між голосними немає жодного приголосного (а), між ними є один приголосний (б), два приголосні (в) або три чи більше приголосних (г);
- у випадках (а) або (б) слово ділять на склади відповідно між голосними або перед єдиним приголосним між ними;
- якщо між голосними є два або більше приголосних, слід шукати висхідну за силою послідовність приголосних; фонеми, які формують таку послідовність, відносять до того ж складу, а якщо приголосні формують низхідну або «горизонтальну» послідов-

ність за силою, то їх відносять до різних складів;

- зокрема, у випадку (в) два сусідні приголосні відносять до різних складів, якщо вони не формують висхідної послідовності або приписують їх до наступного складу, якщо вони формують таку послідовність;
- у найскладнішому випадку (г) знаходимо першу висхідну послідовність, що складається принаймні з двох звуків; тоді склад розділяють перед початком цієї послідовності;
- нарешті, якщо в послідовності трьох або більше приголосних у випадку (г) немає жодної висхідної послідовності, то склад обриваємо або після першого з цих приголосних, або перед останнім із них.

На додаток до стандартної частотної таблиці складів, ми будували також частотні таблиці структури складів. Замість конкретних фонем, у структурному представленні складів фігурують лише умовні позначення голосних і приголосних *v* і *c*. Наприклад, у фонемному і структурному представленні поділ слова «*молоко*» на склади має вигляд відповідно «*мо-ло-ко*» і «*cv-cv-cv*».

У середовищі Visual Studio 2017 мовою програмування C# було створено програму, яка на виході генерує файл із фоно- та силабостатистичними даними для кожної текстової бази. Розроблена програма передбачала такі етапи роботи: зчитування вхідних текстових файлів; їхню попередню обробку (очищення текстів од латиниці, цифр і символів тощо); перетворення графем на фонеми згідно з правилами для обраної мови; розбиття слів на склади; розрахунки статистичних показників графем, фонем і силабем; формування вихідних даних і їхнє експортування. Для підвищення швидкодії опрацювання текстових файлів здійснювали паралельно – кожен в окремому потоці. Максимальна кількість одночасно оброблюваних файлів залежить од кількості ядер процесора комп'ютера, на якому виконується застосунок.

Програма передбачала роботу для трьох досліджуваних мов у режимах абсолютних (*F*) або відносних (*f*) частот, а також в режимі підрахунку букв або в фонетичному режимі. Було передбачено можливість вибору альтернативний правил на основі енергетичного (див. табл. 2) або сонорного підходу до поділу на склади. Програма давала змогу редагувати і зберігати списки голосних і приголосних звуків, чисельні значення їхніх сил, а також списки букв, які не позначають звуки, і символів, які використовують у визначенні звуків і/або поділі на склади (наприклад, буква «ь» і апостроф в українській мові).

Для кожного з текстів ми розраховували такі статистичні показники (див. табл. 1): довжини текстів у словах, складах або буквах чи звуках; частоти окремих букв і звуків; сумарні частоти всіх голосних і приголосних і консонантний коефіцієнт $k = f_c/f_v$; частоти усіх складів і типів (або структур) складів; частоти перших складів слів і їхніх структур; сумарні частоти відкритих (f_{op}) та закритих (f_{cl}) складів і їхнє відношення $K = f_{op}/f_{cl}$; частоти подвоєнь звуків і т. ін.

Окрім фоно- і силабостатистичних параметрів для кожного тексту, ми визначали також їхні зважені середні значення по корпусу та середньоквадратичні відхилення (С.К.В.) від цих середніх. Для вибірки $\{ x_{ij} \}$ деякого *j*-го параметра *i*-го тексту середнє значення $\bar{x}_j$ по корпусу і відповідне С.К.В. Δx_j знаходили за формулами

$$\bar{x}_j = \sum_i x_{ij} w_i \,,\ \Delta x_j = \left(\overline{x_j^2} - (\bar{x}_j)^2 \right)^{1/2} \,,\ \overline{x_j^2} = \sum_i x_{ij}^2 w_i \,, \tag{1}$$

де $w_i = L_i \Big/ \sum_i L_i$ та L_i –ваговий коефіцієнт і довжина *i*-го тексту, відповідно.

Результати та обговорення. 1. Фоностатистика. Основні результати, здобуті для фонетичного режиму аналізу корпусів, наведено в табл. 3. Зазначимо, що в давньоруській мові звуки «*ь*», «*ъ*» і «*ѣ*» – нередуковані або лише частково редуковані голосні; решта умовних позначень для української та російської мов наведено в табл. 2. Коректна інтерпретація даних табл. 3 передбачає врахування таких фактів: 1) ті ж самі звуки в різних мовах іноді мають різні позначення (наприклад, український звук «*і*» та російський «*и*») [*]; 2) ми не враховували таких тонких фонетичних явищ як, наприклад, вимова ненаголошеного «*о*» як «*а*» або асиміляція «*в*» до «*ф*» у російській мові. З цих же причин статистичні дані табл. 3 слід вважати наближеними.

Незважаючи на істотно менший обсяг статистичних вибірок українських і російських текстів, порівняно з аналізом частот графем у роботі [18], результати цього аналізу загалом добре корелюють з даними табл. 3. Попри іноді різні рангові позиції, частоти багатьох фонем на зразок високочастотних «*а*», «*о*», «*і*» або фонем «*р*», «*м*», «*б*», «*т*», «*д*» і «*г*» із середніми або низькими частотами досить схожі для всіх досліджених мов [**]. Це засвідчує близькість української, російської та давньоруської мов. Зокрема, з урахуванням вимови ненаголошеного «*о*» як «*а*» в російській мові частоти та рангові позиції перших двох найчастотніших звуків «*а*» і «*о*» майже ідентичні.

Особливістю української фонетики є несподівано високий, порівняно з російською та давньоруською мовами, ранг звуку «*и*». Щоправда, цей висновок не остаточний і потребує додаткової перевірки, зокрема додаткового уточнення властивостей звуків «*ь*», «*ъ*» і «*ѣ*» у давньоруській мові. Російська мова більш схожа до давньоруської за частотами звуків «*е*», «*у*», «*с*» і «*з*», а українська – за частотами звуків «*н*», «*т*» і «*л*». Нарешті, за частотою звуків «*к*», «*х*» і «*ж*» українська та російська мови істотно ближчі одна до одної, ніж давньоруська.

Наступний важливий для порівняння мов фактор – це залежності частоти *f* звуків від їхнього рангу *r* (див. рис. 1). Згідно з даними праць [1–5], статистичні розподіли частот графем добре описуються функцією Вейбуля, що приводить до наближено логарифмічних рангових залежностей:

$$f(r) \approx a - b \lg r \,, \qquad (2)$$

де *a* і *b* – постійні. У працях [11–13] показано, що ті ж закономірності виконуються для рангових залежностей фонем. Дані рис. 1 підтверджують цей факт емпірично. Параметри апроксимації *a* і *b*, похибки їхнього визначення Δa і Δb, а також коефіцієнти кореляції *R*, які описують якість графічної лінійної апроксимації в масштабі $f(\lg r)$, наведено в табл. 4. Нижчий коефіцієнт *R* для давньоруської мови природно пов'язати з істотно меншим обсягом текстів (див. перший рядок у табл. 1). Дані табл. 4 засвідчують, що за формальними характеристиками рангової залежності $f(r)$ українська мова істотно ближча до давньоруської, аніж російська мова.

[*] Тут і надалі використовуємо позначення звуків, притаманні українській мові.

[**] Тут ми нехтуємо відмінностями вимови «*г*» в українській і російській мовах, а точну вимову різних звуків у давньоруській мові відтворити практично неможливо.

Таблиця 3

Середні значення $\bar{f}$ і С.К.В. Δf відносних частот фонем для українського, російського та давньоруського корпусів текстів

Українська мова			Російська мова			Давньоруська мова		
Звук	$\bar{f}$	Δf	Звук	$\bar{f}$	Δf	Звук	$\bar{f}$	Δf
а	0,108	0,007	*о*	0,111	0,005	*а*	0,096	0,004
о	0,096	0,002	*а*	0,101	0,006	*о*	0,093	0,005
и	0,066	0,003	*э*	0,086	0,005	*е*	0,079	0,023
і	0,061	0,005	*и*	0,067	0,003	*и*	0,072	0,026
н	0,059	0,003	*н*	0,064	0,003	*с*	0,052	0,002
е	0,053	0,002	*т*	0,058	0,004	*т*	0,046	0,004
т	0,046	0,003	*с*	0,052	0,002	*н*	0,044	0,004
р	0,043	0,003	*л*	0,050	0,004	*р*	0,043	0,002
у	0,042	0,002	*в*	0,044	0,002	*j*	0,039	0,004
с	0,040	0,002	*р*	0,043	0,003	*у*	0,035	0,004
л	0,040	0,003	*у*	0,035	0,003	*л*	0,034	0,007
д	0,034	0,002	*к*	0,033	0,003	*м*	0,033	0,002
к	0,034	0,004	*д*	0,032	0,003	*д*	0,032	0,003
м	0,031	0,002	*м*	0,032	0,002	*в*	0,028	0,005
п	0,029	0,002	*j*	0,027	0,003	*ъ*	0,025	0,011
в	0,027	0,002	*п*	0,027	0,002	*к*	0,025	0,002
и	0,025	0,002	*г*	0,019	0,001	*п*	0,024	0,003
з	0,023	0,001	*ы*	0,019	0,001	*ь*	0,024	0,002
j	0,022	0,003	*ч*	0,018	0,002	*ш*	0,020	0,004
ч	0,021	0,002	*б*	0,018	0,001	*б*	0,019	0,001
б	0,019	0,002	*з*	0,017	0,001	*г*	0,017	0,003
г	0,018	0,002	*ш*	0,012	0,001	*ж*	0,015	0,002
ш	0,015	0,001	*Y*	0,011	0,002	*ч*	0,015	0,002
Y	0,012	0,002	*ж*	0,010	0,001	*ы*	0,015	0,002
х	0,012	0,001	*х*	0,009	0,001	*з*	0,013	0,002
ж	0,009	0,001	*ц*	0,004	0,001	*ѣ*	0,020	0,011
ц	0,007	0,002	*ф*	0,002	0,001	*и*	0,011	0,001
w	0,006	0,001				*х*	0,001	0,002
ф	0,002	0,001				*w*	0,001	0,001
r	0,001	0,001				*ц*	0,001	0,002
d	0,0004	0,0001				*Y*	0,0004	0,001
z	0,0002	0,0001				*ф*	0,0004	0,0004

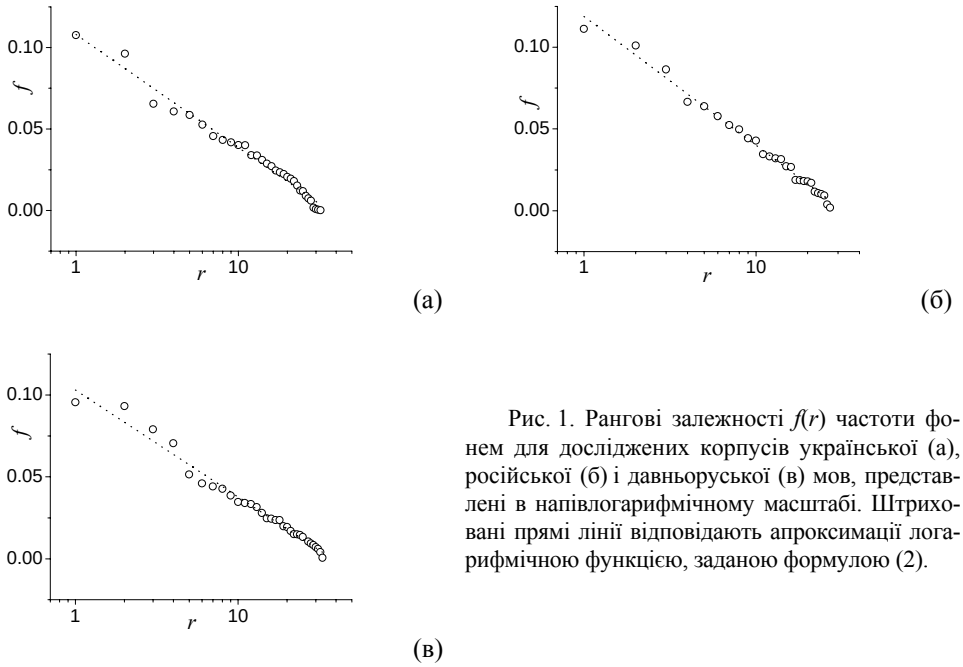


Рис. 1. Рангові залежності $f(r)$ частоти фонем для досліджених корпусів української (а), російської (б) і давньоруської (в) мов, представлені в напівлогарифмічному масштабі. Штриховані прямі лінії відповідають апроксимації логарифмічною функцією, заданою формулою (2).

Таблиця 4

Параметри апроксимації рангових залежностей для фонем логарифмічною функцією (див. текст і формулу (2))

Мова	a	Δa	b	Δb	R
Українська	0,108	0,002	0,069	0,002	0,980
Російська	0,119	0,002	0,079	0,002	0,989
Давньоруська	0,103	0,002	0,065	0,002	0,979

Згідно з даними табл. 1, консонантний коефіцієнт для українського корпусу складає $k \approx 1{,}26$. Зазначимо, що загалом ця величина втрапляє до широкого інтервалу даних, одержаних різними авторами (від 1,2 до 1,36 – див. [16]), перебуваючи поблизу нижньої межі цього інтервалу. З іншого боку, опрацювання текстів у більшості минулих досліджень, швидше за все, здійснювалося вручну, а оброблені вибірки текстів, відповідно, були порівняно малими. Крім того, переважно не було описано й точної методики одержання даних. Зокрема, ми припускаємо, що в літературі могли не враховувати всіх редукцій приголосних звуків, не аналізувати подовжених та подвоєних звуків, не розрізняти слабкі звуки «*ŭ*», «*ѥ*» із їхніми сильними «напівголосними» аналогами тощо. Всі ці, а також інші схожі процедури вкрай важко виконати без помилок, не маючи повністю комп'ютеризованої методики. Більше того, сама дефініція фонем і їхнього списку (наприклад, окремий розгляд палаталізованих фонем) не є однозначними та потребують додаткового аналізу та обговорення (див. [12]). Не виключено також, що аналіз прозорості мов і частот окремих звуків у багатьох працях насправді зводився до підрахунку кі-

лькостей букв, які відповідають тим чи іншим звукам, навіть без спроб врахувати правила фонетики. Майже напевно можемо припустити, що такий спрощений підхід застосовували в одночасних дослідженнях фонетики десятків мов, на зразок праці [11]. На додаток, похибки розрахунків частот будь-яких лінгвістичних елементів дуже серйозно зростають зі зменшенням обсягу оброблених текстів (див., наприклад, обговорення в праці [19]). Це добре ілюструють і помітно вищі С.К.В. частот, одержані для найменшого за обсягом давньоруського корпусу (див. дані табл. 3 для частот звуків і дані наведеної нижче табл. 5 для частот різних типів складів). Як наслідок, ми схильні вважати попередні літературні дані менш надійними, аніж наші емпіричні результати.

Із даних табл. 1 випливає, що значення $k \approx 1{,}26$ для української мови є, по-перше, істотно нижчим за відповідний параметр для російського корпусу (1,36), а, по-друге, воно помітно ближче до значення $k \approx 1{,}13$, притаманного давньоруській мові. Перший із перерахованих результатів вже описаний в літературі (див. [14, 16]) і відповідає добре відомій евфонії або милозвучності української мови. Дослідження лінгвістичної діахронії довели високу частотність відкритих складів у давніх слов'янських мовах і поступову втрату цієї властивості під час подальшої еволюції, а одним із очевидних наслідків закону відкритого складу є висока прозорість давніх мов. Тому наш другий результат засвідчує, що сучасна українська мова успадкувала та помітно краще зберегла цю прозорість давньоруської мови, якщо порівнювати з сучасною російською. Отже, на підставі знайдених нами об'єктивних фактів природно припустити, що за цією рисою українську мову слід вважати більш давньою, аніж російську.

На додаток, високий консонантний коефіцієнт української мови прямо заперечує псевдонаукові підходи, за якими цю мову вважають «діалектом» або «відгалуженням» російської, сформованим під впливом західнослов'янських мов (найперше польської) на давньоруську мову, ототожнену з російською. Якби ці припущення були коректними, то консонантний коефіцієнт російської як «правонаступниці» давньоруської мови був би низьким і близьким до останньої, а відповідний параметр для української мови мав би бути вищим через вплив менш прозорих західнослов'янських мов (наприклад, для польської мови за даними Ю. Тулдави маємо $k \approx 1{,}43$ – див. [16]). Можливість взаємного впливу близьких за географічним ареалом розповсюдження мов, хай навіть віддалених топологічно, загалом є добре відомим з лінгвістичної літератури. Наприклад, нижча прозорість сучасної польської мови, мабуть, таки пов'язана із впливом германських мов її західних сусідів, консонантний коефіцієнт яких ще вищий ($k \approx 1{,}5$ [16]). Проте припущення про вплив польської мови на прозорість української прямо заперечується емпіричними фактами. Прозорість мови найвища на теренах сучасної України, яка географічно збігається з ареалом давньоруської мови, а мови і західних, і східних її сусідів виявляють нижчу прозорість. Тому більше підстав має припущення про те, що втрата прозорості сучасною російською мовою зумовлена впливом угро-фінських мов.

Результати та обговорення. 2. Силабостатистика. Перейдемо до аналізу статистики складів у трьох досліджуваних мовах. Найперше звернемо увагу на те, що середня довжина складу в українському корпусі ($N_{ls} \approx 2{,}31$) дещо менша за довжину складу в російському корпусі (2,34) і помітно перевищує відповідний параметр (2,14) для давньоруського корпусу (див. табл. 1). Перший із цих фактів узгоджується із тим, що середня довжина українських слів менша, ніж російських [19], і корелює зі знайденим вище фактом меншої частки приголосних фонем в українській мові.

За браком місця ми не представлятимемо всіх результатів дослідження статистики складів, обмежившись лише даними для структури складів, які проілюстровано в табл. 5. Примітно, що рангове впорядкування типів складів для української та російської мов майже однакове, але відмінне від давньоруської мови. Найпомітніша специфіка українського корпусу – це істотніша роль складів типу «*c*», порівняно з типом «*vc*». Українська та давньоруська мови ближчі, зокрема, за статистикою складів «*cv*», «*cvc*», «*ccvc*» і «*cvcc*», російська та давньоруська мови ближчі за статистикою «*v*» і «*cccv*»; нарешті, українська та російська майже «рівновіддалені» від давньоруської мови за статистикою «*vc*» і «*ccv*».

Таблиця 5

Середні значення $\bar{f}$ і С.К.В. Δf відносних частот перших десяти за рангом структур силабем для українського, російського та давньоруського корпусів текстів

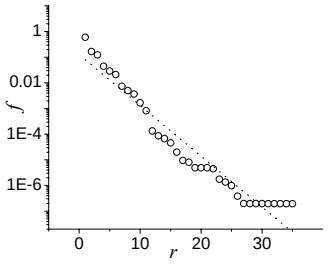
Українська мова			Російська мова			Давньоруська мова		
Склад	$\bar{f}$	Δf	Склад	$\bar{f}$	Δf	Склад	$\bar{f}$	Δf
cv	0,599	0,012	*cv*	0,516	0,016	*cv*	0,634	0,029
cvc	0,165	0,008	*cvc*	0,184	0,011	*ccv*	0,128	0,009
ccv	0,122	0,005	*ccv*	0,133	0,004	*v*	0,120	0,015
v	0,045	0,004	*v*	0,071	0,005	*cvc*	0,055	0,013
ccvc	0,029	0,003	*ccvc*	0,034	0,005	*cccv*	0,018	0,002
c	0,021	0,007	*vc*	0,021	0,002	*ccvc*	0,014	0,004
vc	0,007	0,002	*c*	0,020	0,002	*vc*	0,013	0,003
cccv	0,005	0,001	*cccv*	0,009	0,001	*c*	0,011	0,004
cvcc	0,004	0,001	*cvcc*	0,005	0,001	*cccvc*	0,002	0,001
cccvc	0,002	0,0003	*cccvc*	0,003	0,001	*cvcc*	0,002	0,002

Аналіз відносних частот відкритих (f_{op}) і закритих (f_{cl}) складів (див. табл. 1) засвідчує, що параметр $f_{op} \approx 90\%$ для давньоруської мови помітно більший за відповідні дані для сучасних української (77%) і російської (73%) мов. Усе ж українська мова за показниками f_{op} і K ближча до давньоруської. Знову ж таки, з точки зору діахронічної лінгвістики ці статистичні дані схиляють до висновку про тривалішу історію розвитку української мови, порівняно з російською.

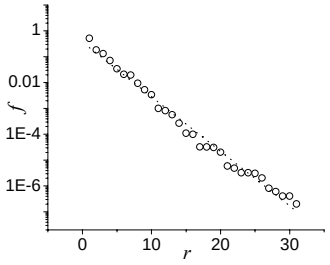
Цікаві результати одержуємо, аналізуючи повні рангові залежності $f(r)$ для частоти структур силабем у досліджуваних корпусах. На рис. 2 ці залежності, представлені в масштабі $\lg f(r)$, майже лінійні, виявляючи лише незначні ознаки вгнутості. У будь-якому разі, спроби лінійної апроксимації шляхом представлення графіків в інших масштабах, які відповідають альтернативним лінійно спадній, степеневій або логарифмічній функціям, безуспішні: графіки $f(r)$ у цих масштабах виявляють дуже істотні відхилення від прямих ліній. Отже, залежності частота–ранг для типів силабем наближено описуються спадною експоненційною функцією:

$$f(r) \approx A\exp(-Br), \tag{3}$$

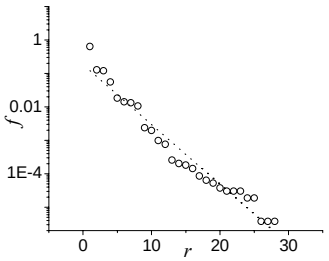
де A і B – постійні. Наскільки відомо авторам, цей емпіричний факт встановлено вперше. В табл. 6 представлено параметри лінійної апроксимації $a = \lg A$ і $b = B$, відповідні похибки Δa і Δb, а також параметр якості апроксимації R.



(а)



(б)



(в)

Рис. 2. Рангові залежності $f(r)$ частоти типів силабем для досліджених корпусів української (а), російської (б) і давньоруської (в) мов, представлені в напівлогарифмічному масштабі. Штриховані прямі лінії відповідають апроксимації експоненційною функцією, заданою формулою (3).

Таблиця 6

Параметри апроксимації рангових залежностей для структур складів експоненційною функцією (див. текст і формулу (3))

Мова	a	Δa	b	Δb	R
Українська	–0,900	0,178	0,198	0,009	0,939
Російська	–0,434	0,081	0,211	0,004	0,987
Давньоруська	–0,740	0,119	0,179	0,007	0,958

Для ліпшого розуміння одержаних результатів скористаємося стандартними «співвідношеннями універсальності», які пов'язують показники α, β і θ степеневих рангової та частотної залежностей і залежності розмірів словника V від розмірів тексту L, відповідно (див. наприклад, [6, 20]):

$$\beta = 1 + 1/\alpha\,,\ \beta = 1 + \theta\,. \tag{4}$$

Якщо формально представити експоненційну функцію (3) як частковий випадок степеневої із показником степеня $\alpha \to \infty$, то на підставі формули (4) одержимо $\theta \to 0$, тобто логарифмічну залежність словника $V(L)$. Отже, експоненційна рангова залежність структурних типів силабем означає асимптотично повільніше, якщо порівнювати зі стандартною степеневою функцією, логарифмічне зростання «словника» типів силабем. Недавно такі ж результати було одержано для рангової залежності слів і залежності словника $V(L)$ окремих східних мов (наприклад, китайської або корейської) з обмеженим словником [21, 22] (див. також аналіз поведінки деяких інших лінгвістичних елементів [23, 24]). Ми вбачаємо в цьому непогану аналогію, оскільки, по-перше, «словник» типів силабем також обмежений і, по-друге, мови східних народів в деякому сенсі можна вважати силабічними; ієрогліф відповідає силабемі та переносить семантику, а тому йому можна поставити у відповідність слово в європейських мовах.

З іншого боку, експоненційний характер рангової залежності не слід вважати тривіальним наслідком обмеженості «словника» складів ($V_{\max} \sim 30-40$). Скажімо, «словник» графем або фонем теж обмежений $V_{\max} \sim 30$, проте рангова залежність для них логарифмічна, а не експоненційна. На підставі зв'язків (4) це описується випадком $\alpha \to 0$ і $\theta \to \infty$, до певної міри протилежним до випадку силабем або китайських слів. Відповідно, «словник» графем зростає експоненційно швидко ($V = V_{\max}[1-\exp(-L/L_c)]$, де L_c – константа) і насичується вже на довжинах текстів $L \sim 100-1000$ [25].

На завершення зупинимося на порівнянні даних табл. 6 для різних корпусів. Як і для рангових залежностей фонем, параметри рангових залежностей типів силабем в українській і давньоруській мові ближчі, порівняно з російською мовою. Фактично за всіма проаналізованими вище основними параметрами фоно- і силабостатистики сучасна українська мова виявляється ближчою до давньоруської, ніж сучасна російська мова.

Висновки. Отже, в цій роботі розроблено програмне забезпечення для всебічного фонетичного та силабічного аналізу текстових баз українською, російською та давньоруською мовами. Одержано низку емпіричних фоностатистичних та силабостатистичних даних для відповідних мов, зокрема рангові залежності для фонем і силабем. Підтверджено відомий з літератури результат про наближену логарифмічну рангову залежність для фонем. Вперше встановлено, що рангова залежність для структурних типів силабем у всіх досліджених мовах описується експоненційною функцією, схоже до статистики слів у деяких східних мовах.

Фактично всі найголовніші параметри фоно- і силабостатистики, наведені в табл. 1, 4 і 6 для українського та російського корпусів, відрізняються в статистично вагомих межах – на величини, типово більші за одне або три стандартних відхилення. Тому за фонетичними та силабічними властивостями українська та російська мова є істотно відмінними в статистичному плані лінгвістичними системами, а не різними реалізаціями єдиної лінгвістичної системи. Інакше, припущення про те, що ці мови можна трактувати як різновиди або «діалекти» єдиної мови, принципово неправильні.

Незважаючи на цей факт, серед трьох проаналізованих нами мов сучасні українська та російська все ж формують своєрідний кластер, у межах якого багато рис фоно- і силабостатистики близькі між собою і, водночас, помітно віддалені від рис давньоруської мови. Така кластеризація української та російської, очевидно, зумовлена близькістю ареалів і тривалими взаємодіями та взаємовпливами. Відмінності ж давньоруської мови від сучасних української та російської підкреслюють визначальний характер мовної еволюції, яка відбувалася протягом близько 800 років.

На підставі одержаних нами статистичних даних для фонем і силабем показано, що українська мова є ближчою до давньоруської, аніж російська мова, помітно переважаючи останню за параметром прозорості та відносним внеском відкритих складів. Ці результати слугують суто лінгвістичними аргументами на користь тези про те, що українська мова має давнішу історію, порівняно з російською мовою.

Серед перспектив цього дослідження ми вбачаємо перевірку одержаних статистичних закономірностей для більших корпусів сучасних мов, а також вивчення статистики силабем безвідносно до їхніх структурних типів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. *Гусейн-Заде С. М.* О распределении букв русского языка по частоте встречаемости // Пробл. передачи информ. – 1988. – Т. 24. – С. 102–107.
2. *Kelih E.* Graphemhäufigkeiten in slawischen Sprachen: Stetige Modelle // Glottometrics. – 2009. – Vol. 18. – P. 52–68.
3. *Li W.*, *Miramontes P.*, *Cocho G.* Fitting ranked linguistic data with two-parameter functions // Entropy. – 2010. – Vol. 12. – P. 1743–1764.
4. *Li W.*, *Miramontes P.* Fitting ranked English and Spanish letter frequency distribution in US and Mexican presidential speeches // J. Quant. Linguist. – 2011. – Vol. 18. – P. 359–380.
5. *Pande H.*, *Dhami H. S.* Mathematical modelling of occurrence of letters and word's initials in texts of Hindi language // Int. J. Math. & Sci. Comput. – 2013. – Vol. 3. – P. 19–38.
6. *Newman M. E. J.* Power laws, Pareto distributions and Zipf's law // Contemp. Phys. – 2005. – Vol. 46. – P. 323–351.
7. *Altmann E. G.*, *Gerlach M.* Statistical laws in linguistics // Proc. Flow Machines Workshop: Creativity and Universality in Language (Paris, 2014). – arXiv:1502.03296 (2015).
8. *Damashek M.* Gauging similarity with n-grams: language-independent categorization of text // Science. – 1995. – Vol. 267. – P. 843–848.
9. *Sigurd B., Eeg-Olofsson M., van de Weijer J.* Word length, sentence length and frequency – Zipf revisited // Studia Linguistica. – 2004. – Vol. 58. – P. 37–52.
10. *Grzybek P.* Close and distant relatives of the sentence: Some results from Russian // In: "Methods and Applications of Quantitative Linguistics", Selected papers of the 8th International Conference on Quantitative Linguistics (QUALICO). – Ed. by I. Obradović, E. Kelih, R. Köhler. – Belgrade, Serbia, April 26–29, 2012. – P. 44–58.
11. *Tambovtsev Yu.*, *Martindale C.* Phoneme frequencies follow a Yule distribution (The form of the phonemic distribution in world languages) // SKASE J. Theor. Linguist. – 2007. – Vol. 4. – P. 1–11.
12. *Buk S.*, *Mačutek J.*, *Rovenchak A.* Some properties of the Ukrainian writing system // Glottometrics. – 2008. – Vol. 16. – P. 63–79.
13. *Grzybek P.* Letter, grapheme and (allo-)phone frequencies: the case of Slovak // Glottotheory. – 2009. – No 2. – P. 30–48.
14. *Мосенкіс Ю. Л.* Проблема милозвучності української мови: Теоретичні й методичні аспекти // Наукові записки НаУКМА. Філологічні науки. – 2002. – Т. 20. – С. 23–25.
15. *Grzybek P.* Historical remarks on the consonant–vowel proportion – from cryptoanalysis to linguistic typology. The concept of phonological stoichiometry (Francis Lieber, 1800–1872) // Glottometrics. – 2013. – Vol. 26. – P. 96–103.
16. *Прокопова Л.* Ще раз про основний параметр милозвучності української мови // Українська мова. – 2010. – № 2. – С. 76–80.
17. *Таранец В. Г.* Энергетическая теория речи. – Одесса : Печатный дом, 2014. – 188 с.
18. *Кушнір О. С.*, *Байовський А. М.*, *Іваніцький Л. Б.*, *Рихлюк С. В.* Флуктуації частоти літер і знаків в українських і російських текстах // Матер. VII Українсько-Польської наук.-практ. конф. «Електроніка та інформаційні технології» (Львів–Чинадієво, Україна). – Львів : Видавн. Львів. ун-ту, 2015. – С. 76–79.
19. *Кушнір О. С.*, *Брик О. С.*, *Дзіковський В. Є.*, *Іваніцький Л. Б.*, *Катеринчук І. М.*, *Кісь Я. П.* Статистичний розподіл і флуктуації довжин речень в українському, росій-

ському і англійському корпусах // Вісн. нац. ун-ту «Львівська політехніка». Серія «Інформаційні системи та мережі». – 2016. – №854. – С. 228–239.
20. *Ferrer i Cancho R.*, *Solé R. V.* Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited // J. Quant. Linguist. – 2001. – Vol. 8. – P. 165–173.
21. *Lü L.*, *Zhang Z.-K.*, *Zhou T.* Deviation of Zipf's and Heaps' laws in human languages with limited dictionary sizes // Sci. Rep. – 2013. – Vol. 3. – 1082 (7 p.).
22. *Deng W.*, *Allahverdyan A. E.*, *Li B.*, *Wang Q. A.* Rank-frequency relation for Chinese characters // Eur. Phys. J. B. – 2014. – Vol. 87. – P. 47–66.
23. *Kushnir O. S.*, *Maksysko M. Ya.*, *Ivanitskyi L. B.*, *Rykhlyuk S. V.* Rank dependences and lexical frequency spectra for the subgroups of different-length words in texts // Електроніка та інформаційні технології. – 2015. – Вип. 5. – С. 167–174.
24. *Кушнір О. С.*, *Мацюняк О. І.* Залежності ранг–частота для символьних N-грам у природних текстах // Матер. VII Українсько-Польської наук.-практ. конф. «Електроніка та інформаційні технології» (Львів–Чинадієво, Україна). – Львів : Видавн. Львів. ун-ту, 2015. – С. 84–86.
25. *Kushnir O. S.*, *Ivanitskyi L. B.*, *Rykhlyuk S. V.* New text-length scaling effects in statistics of natural texts // Матер. VII Українсько-Польської наук.-практ. конф. «Електроніка та інформаційні технології» (Львів–Чинадієво, Україна). – Львів : Видавн. Львів. ун-ту, 2015. – P. 80–83.

Стаття: надійшла до редакції 08.05.2017,
доопрацьована 15.05.2017,
прийнята до друку 16.05.2017.

COMPARATIVE ANALYSIS OF STATISTICAL PROPERTIES OF EAST SLAVIC TEXTS

O. Kushnir[1], T. Srtypko[1], V. Taranets[2], L. Kushnir[1], S. Velgosh[1]

[1]*Ivan Franko National University of Lviv,*
107 Tarnavsky Street, UA-79017 Lviv, Ukraine,
o_kushnir@franko.lviv.ua

[2]*International Humanitarian University,*
33 Fontanska Doroga Street, UA-65009 Odessa, Ukraine

In this work we have developed software for studying a number of statistical properties of the texts written in East Slavic languages, primarily Ukrainian and Russian ones. A computerized technique for the phonetic analysis and syllabication has been worked out. Statistical processing of texts taken from three corpora of fiction has been carried out. The corresponding rank dependences have been obtained and their description given in terms of theoretical relationships. Basing on the technique for comparative analysis, we have revealed similar and distinctive features of the phonetics and phonics of Ukrainian and Russian languages, as well as their historical development.

Key words: statistical linguistics, comparative linguistics, corpus linguistics, East Slavic languages, Ukrainian language, Russian language.