УДК 681.3
DOI: 10.20535/SRIT.2308-8893.2019.2.05

ОЦІНЮВАННЯ КОНТРОЛЮВАЛЬНИХ І КОРИГУВАЛЬНИХ ВЛАСТИВОСТЕЙ РЕФЕРЕНТНОГО СЛОВНИКА СИСТЕМИ ПЕРЕВІРКИ І ВИПРАВЛЕННЯ ОРФОГРАФІЇ

В.А. ЛИТВИНОВ, С.Я. МАЙСТРЕНКО, К.В. ХУРЦИЛАВА, С.В. КОСТЕНКО

Анотація. Розглянуто моделі оцінювання властивостей референтного орфографічного словника (РОС) системи перевірки і виправлення орфографії. Контролювальні властивості РОС визначаються ймовірністю невиявлення типової помилки і ймовірністю хибного сигналу про помилку. Поставлено завдання оптимізації РОС за Парето, запропоновано покроковий алгоритм його розв'язання, наведено дані експериментальної оцінки результативності алгоритму для обраних словників російської й української мов. Коригувальні властивості визначаються ймовірностями правильного і неправильного коригування типових помилок. Запропоновано моделі оцінювання, наведено результати моделювання для обраних словників. Показано, що РОС, оптимізований за контролювальними властивостями, має і кращі коригувальні властивості. Отримані результати можуть бути покладені в основу інструменту порівняльної оцінки, вибору і поліпшення потенційних властивостей конкретного РОС для заданої предметної галузі.

Ключові слова: помилка тайпінгу, перевірка орфографії, контролювальні властивості, коригувальні властивості.

ВСТУП

Натепер системи перевірки орфографії (СПО) є як затребуваним самостійним продуктом (ОРФО, Language Tool та ін.), так і обов'язковим компонентом текстових редакторів, пошукових систем, поштових клієнтів, електронних словників та ін. Центральний елемент таких систем — референтний орфографічний словник, що містить коректні слова деякої предметної галузі, з якими порівнюються слова, що перевіряються. Наявна в доступних джерелах інформація відображає передусім орієнтовані на користувача відомості про інструменти СПО — функціонал, технологію роботи, особливості використання тощо.

Разом з виявленням орфографічних помилок багато загальних і спеціалізованих текстових редакторів та інших програм оброблення текстів пропонують функцію автоматичного і напівавтоматичного виправлення помилок. Короткий огляд основних методів автоматичного виправлення і нечіткого пошуку (fuzzy string search) на основі оцінки відстаней Левенштейна і Дамерау–Левенштейна наведено у працях [1, 2]. Типовим рішенням щодо вибору алгоритмів для конкретної реалізації є використання фонетичних алгоритмів [3, 4]. Дослідження різних модифікацій таких алгоритмів та альтернативних алгоритмічних підходів і систем (наприклад, [5, 6]) відображені у ряді публікацій в пострадянських і зарубіжних джерелах.

Загалом публікації за темою присвячені користувацьким та алгоритмічним аспектам проблематики СПО. Питанням же кількісної оцінки і можливого поліпшення потенційних контролювальних і коригувальних властивостей самих РОС мало приділяється уваги.

Запропонована робота спрямована на часткове заповнення відзначеного пропуску щодо типових помилок тайпінгу.

Надалі використаємо загальні позначення:

A_j — слово РОС $(j = 1..N)$;

$\overline{A}_j$ — слово РОС, спотворене помилкою;

q — алфавіт символів РОС.

Поділимо різноманітні помилки $\overline{A}_j$ на дві групи: ансамбль специфічних помилок K , що підлягають коригуванню, та інші помилки. Традиційно (та у відповідності зі складовими показника відстані Дамерау–Левенштейна) віднесемо до ансамблю K типові орфографічні помилки тайпінгу — однократні транскрипції E_1, уставки E_2 і випадіння E_3 символу та суміжні транспозиції E_4. Властивості РОС стосовно різноманітних довільних помилок принципово не обліковують і не аналізують, оскільки, якщо не накладати обмежень на характер цих помилок, можна стверджувати, що для кожного слова A_j існують помилки, що переводять його у будь-яке інше слово. Тому контролювальні і коригувальні властивості РОС надалі оцінюватимемо щодо здатності виявляти і виправляти саме базові помилки ансамблю K .

КОНТРОЛЮВАЛЬНІ ВЛАСТИВОСТІ РОС

Дисфункція референтного словника

Контролювальні властивості РОС визначаються ймовірністю невиявлення помилки в результаті випадкового збігу спотвореного слова з деяким стороннім допустимим словом. Груба оцінка значення $Q_{\text{нв}}^{(0)}$ такої ймовірності може ґрунтуватися на припущенні щодо випадкового характеру спотворень слова A_j та зіставленні потужностей $Q_{\text{з}}$ заборонених (відсутніх у РОС) комбінацій символів і Q_p допустимих [7]:

$$Q_{\text{нв}}^{(0)} \approx \frac{Q_{\text{Д}}}{Q_{\text{Д}} + Q_{\text{з}}} \approx \frac{N}{q^n}, \qquad (1)$$

де n — середня кількість символів у слові.

Для РОС із цілеспрямовано введеною надмірністю і відносно рівномірним (випадковим) розподілом N реальних слів серед q^n різних значень комбінацій n символів в алфавіті q — зокрема для кодових довідників — оцінка (1) може бути достатньо близькою до істини. Для природномовних слів (слів у текстовому редакторі, ключового слова в пошуковій системі і т.ін.) і специфічних спотворень, викликаних типовими помилками користувача, припущення про випадковий характер розподілів значень слів та їх

можливих спотворень не виконуються. Тут найбільш імовірні прості спотворення можуть дати значно більшу кількість неправдивих збігів з реально існуючими словами і, відповідно, набагато гіршу результативність контролю.

Реальні значення Q і $Q_{\text{нв}}$ можуть бути оцінені безпосередньо за допомогою моделі, що імітує процес спотворення кожного зі слів РОС помилками ансамблю K та виявлення цих помилок шляхом пошуку збігів помилкових слів із словами РОС.

Повновибіркове $(j = 1...N)$ моделювання проведено у праці [7] для імовірно рівнозатребуваних слів трьох словників російської мови і адаптованих україномовних версій цих словників, сформованих за допомогою російсько-української конвертації (перекладу). Зокрема, досліджені Словари русского языка. Словарь А.А. Зализняка [8] (СЗр — російський, СЗу — українська версія), Лопатин Владимир — Русский орфографический словарь [9] (СЛр — російський, СЛу — українська версія), Словари русского язика. Словарь русской литературы [10] (СРЛр — російський, СРЛЗу — українська версія). Для орієнтовних оцінок імовірності P_k помилок класів E_k взято значення із праці [11].

Результати моделювання наведено у табл. 1

Таблиця 1. Результати повновибіркового моделювання

k	P_k	СЗр $N = 92555$ $\overline{n} = 9{,}61$	СЛр $N = 150213$ $\overline{n} = 10{,}06$	СРЛр $N = 161730$ $\overline{n} = 8{,}44$	СЗу $N = 84575$ $\overline{n} = 9{,}49$	СЛу $N = 135401$ $\overline{n} = 9{,}93$	СРЛу $N = 129244$ $\overline{n} = 8{,}31$
1	0,56	0,39	0,41	1,2	0,28	0,28	1,0
2	0,16	0,06	0,07	0,27	0,04	0,04	0,15
3	0,12	2,14	2,16	8,8	1,39	1,40	5,2
4	0,06	0,95	1,55	1,2	0,91	1,22	1,1
$Q_{\text{нв}}$	0,9	0,54	0,6	1,84	0,38	0,41	0,77

Як видно з даних, наведених у табл. 1, реальні значення $Q_{\text{нв}}$ значно (на порядки) перевищують ідеалізовані значення $Q_{\text{нв}}^{(0)}$. Це є наслідком того, що ланцюжок взаємних спотворень слів типу $\langle кол \rangle \Rightarrow \langle пол, мол, гол, фол, вол, тол, дол \rangle$ дає значно більшу кількість збігів зі словником, ніж, наприклад, випадковий малоймовірний гіпотетичний перехід $\langle кол \rangle \Rightarrow \langle крах \rangle$, у результаті контролювальна здатність словників як російської, так і української мови виявляється значно нижчою, ніж можна було б припустити, виходячи з виразу (1). При цьому різні словники мають контролювальні властивості, що помітно розрізняються. Так, із 1000 випадкових помилкових слів словників, спотворених помилками 1, 2, 3, 4 (у вказаній пропорції), у середньому не виявляються 5,4 помилок для Словника Залізняка і 18,4 помилки для Словника російської літератури. Виходячи з отриманих даних, можна припустити, що діапазон значень $Q_{\text{нв}}$ для досліджених словників визначається як чисто лінгвістичними чинниками (мовою, структурою), так і різницею в обсягах. Зменшення обсягу словника за інших однакових умов прогнозовано повинно зменшувати $Q_{\text{нв}}$ за рахунок явного збільшення від-

носної надмірності подання слів і відповідного зменшення можливостей випадкових збігів помилкових слів з допустимими. З іншого боку, вилучення зі словника слів з ненульовою затребуваністю збільшує ймовірність хибних повідомлень про помилки. Тому комплексна оцінка контролювальних властивостей РОС, що оцінює його якість, визначається двома чинниками [12]:

– здатністю виявляти якнайбільше найімовірніших (типових) помилок;

– здатністю виявляти якнайменше хибних повідомлень про помилки.

Перший чинник оцінимо показником дисфункції РОС, який визначимо через значення ρ відносної кількості слів, що спотворені певними помилками і збіглися з іншими, реально допустимими словами; відповідні помилки системою не виявляються.

Другий чинник оцінимо ймовірністю φ відсутності запитаного слова у РОС.

Для уточнення і розвитку поняття дисфункції і питань, пов'язаних з її кількісною оцінкою, розглянемо наступну модель спотворення слів і виявлення помилок (чекінгу) (рис. 1).

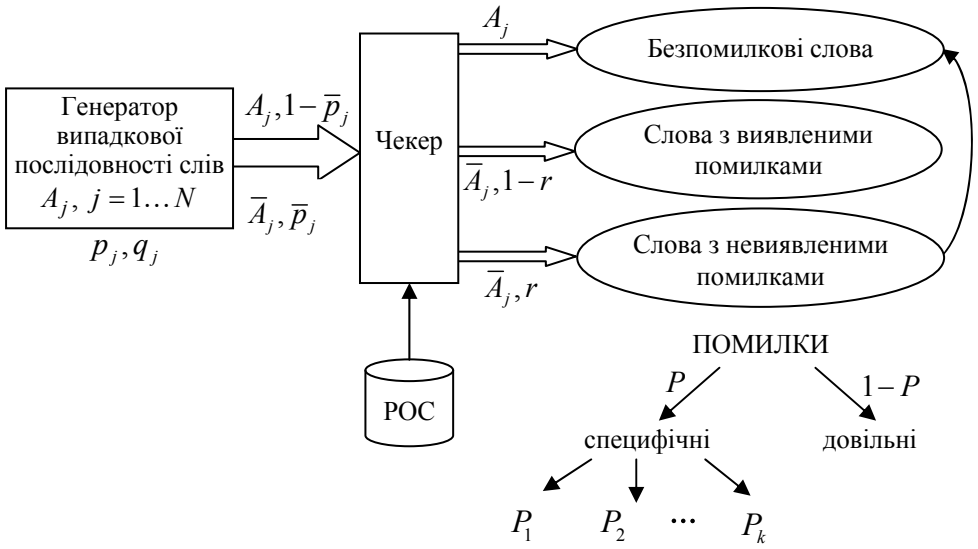

Рис. 1. Схема ймовірнісної моделі спотворення слів

Для пояснення позначень розглянемо інтерпретацію функцій, поданих на рис. 1 об'єктів.

Уявимо генератор слів A_j у вигляді деякої урни, у якій містяться деякі фішки зі словами A_j; значення $p_{j,}$ визначає відносну кількість фішок з конкретним словом A_j. Під час замісного витягання фішок з урни (генерації послідовності слів) фішки можуть пошкодитися, причому конкретна фішка A_j пошкоджується $(A_j \to \overline{A}_j)$ з імовірністю q_j.

У вихідному потоці слів генератора відносна кількість $\overline{p}_j$ пошкоджених фішок $\overline{A}_j$ визначається добутком $\overline{p}_j = p_j q_j$ (з точністю до умов нормування). Чекер порівнює фішки A_j, $\overline{A}_j$ з еталонними і виокремлює пошко-

джені, причому деяка частина пошкоджень r не розпізнається, оскільки пошкоджена фішка $\overline{A}_l$ збігається з деякою еталонною A_s. Показник дисфункції r визначає відносну кількість невиявлених пошкоджень (помилок) до повної кількості пошкоджених фішок (помилкових слів).

Визначимо величину r через властивості слів A_j, $\overline{A}_j$ словника РОС. Позначимо через r_j відносну кількість невиявлених помилок у слові $\overline{A}_j$ до всіх можливих помилок у цьому слові. Тоді

$$r = \sum_{j=1}^{N} r_j \cdot \overline{p}_j . \tag{2}$$

За визначенням, обмежуючись помилками ансамблю $\boldsymbol{K}$, для виразу (2) беремо

$$r_j \to \rho_j = \sum_k \frac{\rho_{kj} P_k}{P}$$

та

$$r \to \rho = \sum_j \sum_k \overline{p}_j \cdot \frac{\rho_{kj} P_k}{P}, \tag{3}$$

де ρ_{kj} — відносна кількість невиявлених помилок підкласу k до різноманітних помилок цього підкласу у слові A_j.

З очевидних міркувань

$$\rho_{kj} = \frac{v_{kj}}{V_{kj}}, \tag{4}$$

де v_{kj} — кількість збігів слова $\overline{A}_j$, спотвореного помилками підкласу k, з іншими словами РОС; V_{kj} — повна кількість різноманітних спотворень слова A_j у підкласі помилок k (кількість варіацій $\widetilde{A}_{kj}$) [7].

Таким чином, вирази (3), (4) визначають зміст та значення показника дисфункції довільного РОС та довільної функції q_j.

Для процесу послідовного введення (передавання) символів слів, що перевіряються, можна скористатися відомим допущенням про пряму залежність імовірності спотворення слова від його довжини:

$$q_j \approx \pi_c n_j \mu ,$$

де π_c — статистична ймовірність спотворення довільного символу у процесі послідовного введення; n_j — кількість символів у слові A_j; μ — нормувальний множник.

Тоді з урахуванням нормувальної умови $\sum_j q_j = 1$

$$q_j = \frac{n_j}{\sum_{j=1}^{N} n_j}.$$

В окремому випадку для $n_j = \text{const} = n$ отримаємо $q_j = \frac{1}{N}$.

Якщо ще і $p_j = \frac{1}{N}$, то нормоване значення $\overline{p}_j$ також дорівнює $\frac{1}{N}$.

Задача узгодження критеріїв якості РОС та підхід до її розв'язання

Нині для одних і тих самих предметних галузей (зокрема, понять і слів української, російської та інших природних мов) існують різні готові орфографічні словники, що розрізняються широтою охоплення тезауруса (обсягами) і контролювальними властивостями. Зменшення обсягу словника за інших однакових умов протилежним чином впливає на показники якості — значення ρ і φ. З одного боку, за рахунок збільшення відносної надмірності подання слів і відповідного зменшення випадкових збігів помилкових слів з допустимими зменшується показник дисфункції ρ, з другого боку, вилучення з РОС слів з ненульовою затребуваністю (імовірністю звернення) збільшує значення φ.

Уявимо тезаурус T деякої предметної галузі, складеним з двох частин — дійсної (видимої) і уявної (прихованої). Дійсна частина є конкретним реальним РОС, а уявна (УРОС) — частина слів T, не поданих у РОС, але потенційно затребуваних. Задача узгодження критеріїв якості РОС формулюється як завдання вилучення і переміщення в УРОС слів, які більше за інших зменшують ρ і менше ніж інші збільшують φ. Інакше кажучи, ідеться про побудову парето-оптимальної траєкторії значень ρ, φ у міру переміщення вибраних слів в УРОС з метою можливого вибору прийнятного поєднання ρ, φ. Формування точкового критерію, що оцінює конкретний внесок слів, які потенційно вилучаються, у значення показників якості, ґрунтується на таких положеннях:

1. Вилучення нейтрального слова, помилки у якому не спричиняють збігів з реальними словами, не зменшують значення ρ, але збільшують значення φ. Отже, вилучення таких слів не входить у парето-оптимальні розв'язки.

2. Для кожного слова A_l, пряма помилка $A_l \to A_s$ у якому викликає збіг зі словами A_s (наприклад, *влечь* $\Rightarrow$ {*слечь*, *увлечь*, *лечь*}), існують обернені помилки, що не виявляються, $A_s \to A_l$ ({*слечь*, *увлечь*, *лечь*} $\Rightarrow$ *влечь*).

3. Із вилученням слова A_l (зокрема, *влечь*) обернені помилки будуть виявлятися (оскільки слова *влечь* не буде у РОС), а прямі помилки $A_l \to A_s$ — ні (оскільки затребуваність слова *влечь* не зникає, а пов'язані слова A_s залишаються у РОС).

Наведені якісні міркування узагальнює критерій відповідності (~), який може бути покладений в основу покрокового алгоритму розв'язання задачі вибору для вилучення слова A_l:

$$A_l \sim \min_l \alpha_l = \frac{p_l}{\Delta\rho_l},\ \Delta\rho_l = \sum_k p_k \sum_s \overline{p}_s \, {}^{v_{ks}}\!\Big/\!_{V_{ks}}, \qquad (5)$$

де V_{ks} — повна кількість різноманітних спотворень слів A_s у класах помилок E_k.

У наведеній постановці задачу можна розглядати як деяке узагальнення задачі рюкзака (Knapsack Problem [13]), а покроковий алгоритм її розв'язання на основі критерію (5) — як різновид жадібного алгоритму GA (Greedy algorithm [14]), у якому в рюкзак поміщають предмети з максимальним співвідношенням ціни (у розгляданому випадку $\Delta\rho$) до маси (p_l). Ця задача відрізняється від класичної Knapsack Problem тим, що там ціна і маса предметів залишаються постійними в процесі укладання рюкзака, а в цьому випадку ціна предметів, що залишилися після часткового завантаження рюкзака, може змінюватися залежно від того, що було завантажено перед цим. Останнє зумовлено тим, що вилучення слова A_l змінює розподіл наслідків можливих помилок у частині РОС, що залишилася.

Результати моделювання алгоритму GA з критерієм (5) наведено у табл. 2, 3.

Таблиця 2. Словник Лопатіна

$\frac{N-Y}{N}$	$\lambda = 8/N$				$\lambda = 24/N$			
	Випадкове зменшення		Розрахункове зменшення		Випадкове зменшення		Розрахункове зменшення	
	$\rho^{(Y)}/\rho^{(N)}$	$\varphi\cdot10^2$	$\rho^{(Y)}/\rho^{(N)}$	$\varphi\cdot10^4$	$\rho^{(Y)}/\rho^{(N)}$	$\varphi\cdot10^2$	$\rho^{(Y)}/\rho^{(N)}$	$\varphi\cdot10^4$
1	1,0	0	1,0	0,0	1,0	0,0	1,0	0,0
0,94	0,942	5,660	0,812	4,00	0,937	6,229	0,822	0,000014
0,88	0,888	11,909	0,563	28,38	0,877	12,525	0,554	0,003772
0,82	0,838	17,901	0,224	299,36	0,822	17,957	0,197	9,72321
0,81	0,827	19,064	0,126	632,099	0,814	19,009	0,041	229,8810

Таблиця 3. Українська версія словника Лопатіна

$\frac{N-Y}{N}$	$\lambda = 8/N$				$\lambda = 24/N$			
	Випадкове зменшення		Розрахункове зменшення		Випадкове зменшення		Розрахункове зменшення	
	$\rho^{(Y)}/\rho^{(N)}$	$\varphi\cdot10^2$	$\rho^{(Y)}/\rho^{(N)}$	$\varphi\cdot10^2$	$\rho^{(Y)}/\rho^{(N)}$	$\varphi\cdot10^2$	$\rho^{(Y)}/\rho^{(N)}$	$\varphi\cdot10^2$
1	1,0	0,0	1,0	0,0	1,0	0,0	1,0	0,0
0,94	0,939	6,077	0,710	5,881	0,949	5,487	0,701	0,000065
0,88	0,879	12,150	0,327	108,187	0,897	11,314	0,322	0,663742
0,86	0,864	14,001	0,126	449,207	0,874	13,122	0,042	1749,524

Дані табл. 2 мають такий зміст. Параметр λ визначає крутизну експоненціальної кривої, що апроксимує ступеневий розподіл щільності ймовірності звернень до слів. Для $\lambda = 8/N$ розподіл характеризується відношенням 20/80 (80% звернень охоплює всього 20% слів), а для $\lambda = 24/N$ — співвідношенням 10/90 (з аналогічним змістом) [12].

У разі випадкового зменшення Y слів переносилися в УРОС випадковим чином, за розрахункового — відповідно до результатів роботи описаного вище алгоритму (точковим підбиранням).

Моделювання проводилося для словника Лопатіна (табл. 2) і його україномовної версії (табл. 3). Дані таблиць ілюструють відносну результативність роботи алгоритму. Зокрема, наприклад, для $\lambda = \frac{8}{N}$ вибіркове вилучення 6% слів призводить до зниження значення показника дисфункції на 18% (російський словник Лопатіна) і 29% (україномовна версія словника), а випадкове — усього на 5,8% і 6,1 %. Відповідне значення φ становить $5.7 \cdot 10^{-4}$ і $5.9 \cdot 10^{-4}$ для вибіркового вилучення і $5.7 \cdot 10^{-2}$ і $6.1 \cdot 10^{-2}$ для випадкового.

КОРИГУВАЛЬНІ ВЛАСТИВОСТІ РОС

Загальні положення. Логіко-імовірнісна модель коригування

Уведемо такі поняття та позначення:

$d(A_j^i, \overline{A}_j)$ — функція відстані, що визначає в деякій метриці орфографічну близькість слів $\overline{A}_j$ та слів РОС $(i = 1 \ldots N)$;

$F_1(A_j^i, \overline{A}_j)$ — функція попереднього вибору, що визначає множину слів РОС, для яких $d(A_j^i, \overline{A}) < d_{\max}$;

$\hat{A}_j^l$ — слова РОС, для яких $d(\hat{A}_j^l, \overline{A}_j) = \min_i d(A_j^i, \overline{A}_j); \quad l = 1 \ldots z;$ $z = 0,1,\ldots$ Для $z = 0$ таких слів не знайдено;

$F_2(\hat{A}_j^l \to A_j)$ функція преференцій, що визначає вибір із z слів конкретного слова $\hat{A}_j^l$ для коригування (заміни) помилкового слова $\overline{A}_j$.

Унаслідок коригування помилкового слова $\overline{A}_j$ можливі такі результати:

- помилка $A_j \to \overline{A}_j$ не виявлена (фінальна подія $S_{j\text{нв}}$, імовірність результату $Q_{j\text{нв}}$);
- помилка $A_j \to \overline{A}_j$ виявлена (подія S_0), знайдено одне або більше слів-кандидатів $\hat{A}_j^{il}\,(m \ge 1)$, функція $F_2(\hat{A}_j^l, A_j)$ визначила правильний вибір, коригування виконано правильно (фінальна подія $S_{j\text{пк}}$, імовірність результату $Q_{j\text{пк}}$);
- помилка $A_j \to \overline{A}_j$ виявлена, $z \ge 1$, функція $F_2(\hat{A}_j^{il} \to A_j)$ визначила помилковий розв'язок, коригування виконано помилково (фінальна подія $S_{j\text{хк}}$, імовірність результату $Q_{j\text{хк}}$);

• помилка $A_j \to \overline{A}_j$ виявлена, не знайдено жодного $(z = 0)$ слова-кандидата, для якого $d(A_j^i, \overline{A}_j) < d_{\max}$; коригування не виконується (фінальна подія $S_{j\text{нк}}$, імовірність результату $Q_{j\text{нк}}$).

Мета побудови й аналізу конкретної логіко-імовірнісної моделі полягає у визначенні для конкретного РОС значень імовірності відповідних результатів, що визначають коригувальні властивості РОС для окремих слів і словника в цілому.

Унаслідок реалізації процесу коригування можливі різні рішення щодо вибору функцій відстаней і переваг. Для оцінювання властивостей РОС конкретизуємо узагальнену модель (рис. 2) для таких умов.

1. Визначаючи функцію попереднього вибору, обмежимося базовими помилками ансамблю $\boldsymbol{K}$, що за визначенням звужують зону пошуку варіантів коригування помилкового слова.

2. Покладемо

$$d_0(A_j^i, \overline{A}_j) = \begin{cases} 0 \text{ для } A_j^i \equiv \widetilde{\overline{A}}_j^i, \text{ де } \widetilde{\overline{A}}_j^i - \text{ вторинне;} \\ \quad \text{сппотвореня (вваріація слова } \overline{A}_j; \\ d_{\max} \text{ в іншому випадку.} \end{cases} \qquad (6)$$

Рівність $d_0(A_j^i, \overline{A}_j) = 0$ означає, що варіація $\widetilde{\overline{A}}_j^i$ збігається зі словом A_j^i. У межах уведених умов відстань Дамерау–Левенштейна мінімальна для слів A_j^i, з якими збігається варіація $\widetilde{\overline{A}}_j^i$ у класах $E_1 - E_4$ ансамблю коригованих помилок.

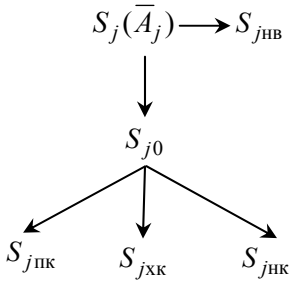

Рис. 2. Узагальнена модель подій

3. Для функції переваги визначимо найгірше рішення — рівноймовірний вибір із z збігів. Оскільки генеруються усі варіації помилкового слова, принаймні один збіг тут забезпечений, тобто $z \geq 1$.

Логіко-імовірнісну модель, конкретизовану для прийнятих умов, наведено на рис. 3.

Додаткові позначення для окремих подій:

S_{j0} і $S_{j\text{нв}}$ — помилка виявлена/не виявлена;

S_{j1} і $S_{j2} = \overline{S}_{j1}$ — помилка належить/не належить до ансамблю K відповідно;

S_{j1}^k — помилка належить до класу E_k;

S_{j11} і $S_{j12} = \overline{S}_{j11}$ — помилка однозначна $(z = 1)$ / неоднозначна $(z > 1)$;

$S_{j\text{пкl}}^k$ — помилка класу E_k коригується однозначно правильно;

S_{j121}^k — фактично багатозначній помилці класу E_k відповідає перший $(l = 1)$ зі збігів;

$S^k_{j122} = \overline{S}^k_{j11}$ — фактично багатозначній помилці класу k відповідають збіги з $l = 2 \ldots z$.

Таким чином,

$$S^k_{j\text{пк}1} = S_{j0} \wedge S^k_{j1} \wedge S^k_{j11};$$

$$S^k_{j\text{пк}} = (S_{j0} \wedge S^k_{j1} \wedge S^k_{j11}) \vee (S^k_{j0} \wedge S^k_{j1} \wedge S^k_{j12} \wedge S^k_{j121})) =$$

$$= (S_{j0} \wedge S^k_{j1}) \wedge (S^k_{j11} \vee (S^k_{j12} \wedge S^k_{j121}));$$

$$S^k_{j\text{лк}1} = S^k_{j0} \wedge S^k_{j1} \wedge S^k_{j12} \wedge S^k_{j122},$$

$$S_{\text{пк, лк}} = \bigvee_j \bigvee_k S^k_{j\text{пк, }j\text{лк}}.$$

Натурно-імітаційне моделювання

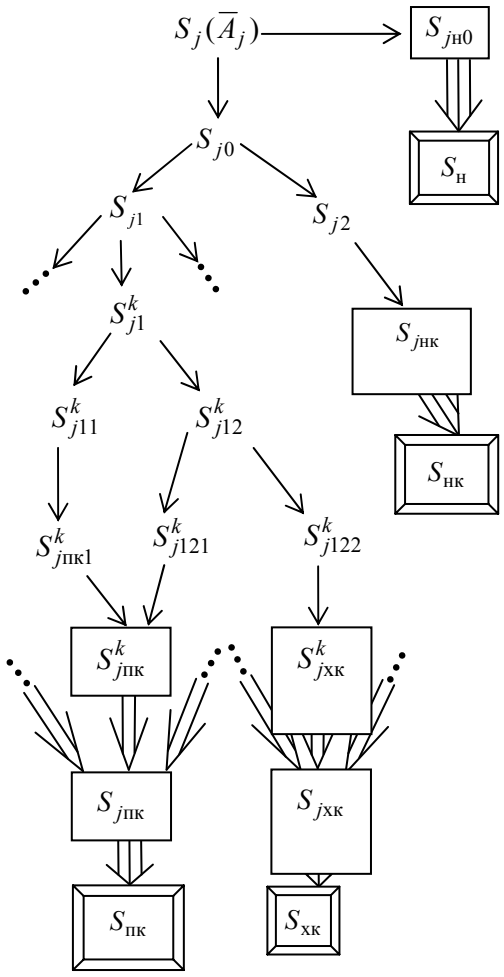

Рис. 3. Логіко-імовірнісна модель визначення коригувальних властивостей РОС моделювання

Натурно-імітаційне моделювання процесу спотворення та коригування слів РОС грунтується на генерації для кожного слова A_j можливих коригованих помилок ансамблю *K*, перевірці можливості виявлення помилки, генерації для кожної помилки можливих варіантів коригування (зворотних спотворень) і пошуку збігів у словнику. При цьому ймовірність проміжних і фінальних подій визначається де-факто для конкретного словника через відповідні кількості збігів.

Приклади можливих окремих випадків для конкретного слова $A_j := арак$:

$\overline{A}_{jks} = мрак$, помилка не виявляється;

$\overline{A}_{jks} := аеак$, $\tilde{\overline{A}}_{jks} = \{арак\}$, $z = 1$, помилка коригується однозначно;

$\overline{A}_{jks} := прак$, $\tilde{\overline{A}}_{jks} = \{арак, брак, мрак, трак, рак, парк\}$, $z = 6$, за випадкового вибору помилка коригується правильно з імовірністю 1/6 і неправильно з імовірністю 5/6;

$\overline{A}_{jks} := \textit{аарк}$, $\tilde{\overline{A}}_{jks} = \{\textit{барк}, \textit{карк}, \textit{марк}, \textit{парк}, \textit{тарк}, \textit{арак}\}$, $z = 6$, за випадкового вибору помилка коригується правильно з імовірністю 1/6 і неправильно з імовірністю 5/6;

Оскільки під час моделювання генеруються усі можливі кориговані помилки та всі варіанти їх виправлення, результат моделювання (зокрема, значення ймовірності $Q_{\text{пк}}$ і $Q_{\text{хк}}$) повністю характеризує коригувальні властивості конкретного РОС.

Моделювання виконано для наборів словників і значень P_k. У зв'язку з відносно високою обчислювальною трудомісткістю процесу генерації помилок і варіантів їх виправлення оброблялися випадковим чином сформовані вибірки обсягом 20000 слів (з оцінкою відповідної довірчої ймовірності). Результати моделювання наведено у табл. 4.

Таблиця 4. Результати моделювання для натурно-імітаційного коригування

Словник	$Q_{0\text{пк}}$	$Q_{0\text{лк}}$	$Q_{\text{пк}}$	$Q_{\text{хк}}$	$Q_{\text{нв}}$	$Q_{\text{нк}}$
Словник російської літератури, $N = 161730$	0,7549	0,1443	0,7410	0,1416	0,0184	~01
Словник Лопатіна, $N = 150213$	0,8282	0,0709	0,8233	0,0706	0,0060	~01
Словник Залізняка, $N = 92555$	0,8281	0,0710	0,8236	0,0706	0,0054	~01
Словник Лопатіна скорочений, $N = 84575$	0,8518	0,0474	0,8483	0,0472	0,0038	~01
Українська версія Словника Лопатіна $N = 84575$	0,8610	0,0382	0,8585	0,0381	0,0028	~01

Довірчі інтервали для отримання середніх загальних значень $Q_{0\text{пк}}$, $Q_{0\text{хк}}$, обчислені на основі припущення про близький до нормального закон розподілу окремих значень $Q_{j0\text{пк}}$, $Q_{j0\text{хк}}$, з імовірністю 0,99 становлять $\pm 0{,}5\%$ для словника російської літератури, $\pm 0{,}3\%$ для словників Лопатіна та Залізняка і $\pm 0{,}2\%$ для скорочених словників.

Із даних табл. 4 видно, що коригувальні властивості так само, як і контролювальні помітно розрізняються для різних словників. Так, для словника російської літератури з 1000 довільних помилок не виявляється 18,4 помилки, правильно коригується 741 помилка і неправильно — 141 помилка. Відповідні значення для скороченого словника Лопатіна складають 2,5; 850 і 47.

Розкид значень $Q_{\text{пк}}$, $Q_{\text{хк}}$ для різних словників пояснюється двома чинниками. З одного боку, словник меншого обсягу за інших однакових умов повинен мати більш високі значення $Q_{\text{пк}}$ і менші $Q_{\text{хк}}$ за рахунок більшого значення відносної надмірності подання слів і відповідного зменшення можливостей збігу згенерованих варіантів виправлення помилок з реальними словами словника. Так, для словника Лопатіна обсягом 92555 слів значення $Q_{\text{пк}} = 0{,}8233$, а для скороченого (випадковим чином) цього ж словника обсягом 84575 $Q_{\text{пк}} = 0{,}8483$. Із другого боку, відіграють роль і чисто лінгвістичні чинники (мова, тезаурус). Так, для української версії скороченого словника Лопатіна, що має такий самий обсяг і такий самий набір

слів, що і російськомовна версія $Q_{\text{пк}} = 0{,}8594$. У цілому, як видно з даних табл. 4, існує явно високий ступінь кореляції між значеннями $Q_{\text{нв}}$ і $Q_{\text{хк}}$. Цей чинник у поєднанні із впливом відносної надмірності словника дає підстави для таких попередніх висновків:

– словник, оптимізований (за Парето) щодо контролювальних властивостей, має і кращі коригувальні властивості;

– показник відносної надмірності словника може бути використаний як основа для оцінювання його коригувальних властивостей.

Натурно-аналітична модель коригувальних властивостей

Зупинимося детальніше на значенні згаданого поняття «відносна надмірність словника» і його кількісного зв'язку з контролювальними і коригувальними властивостями. Розглянемо ідеалізований гіпотетичний словник обсягом *N* слів однакової довжини *n* символів в алфавіті *q*.

На рис. 4 показано лінійну модель такого словника, у якій q^n активних комірок позначають різноманітні значення комбінацій *n* символів, а виділені комірки A_j позначають комбінації, що відповідають реально існуючим словам ($j = 1 \ldots N$).

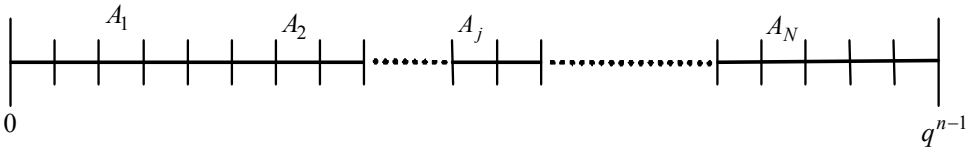

Рис. 4. Лінійна інтерпретація словника

Акт перевірки правильності слова A_j, спотвореного довільною помилкою $A_j \to \overline{A}_j$, можна розглядати тут як акт очищення комірки A_j і «кидання» комбінації $\overline{A}_j$ на регістр комірок. За припущення випадкового характеру розподілу активних комірок в інтервалі $0..q^n - 1$ імовірність потрапити комбінацією $\overline{A}_j$ в зайняту комірку $r = N \big/ q^n$, а відносну надмірність словника *C* можна оцінити як $C = 1 - r = 1 - \dfrac{N}{q^n}$.

Чим більше N за інших однакових умов, тим більше r, тим гірші і контролювальні властивості (вища ймовірність випадкового збігу помилкового слова з реально існуючим) та коригувальні властивості (більша кількість рівноправних варіантів коригування помилкового слова, зокрема варіантів повного збігу).

Для функції (6) і гіпотетичного ідеалізованого словника можлива ймовірнісна оцінка кількості випадкових збігів довільного помилкового слова (генерованої варіації) зі словником на основі моделі незалежних випробувань Бернуллі і відповідної формули біноміального розподілу

$$P(g, r, V) = C_V^g r^g (1 - r)^{V - g}, \qquad (7)$$

де $P(g,r,V)$ — імовірність отримання точно g випадкових збігів у результаті V дослідів, у кожному з яких імовірність сприятливого наслідку дорівнює r; C_V^g — кількість сполучень з V за g.

Проте для реального словника така оцінка значень $Q_{\text{пк}}$, $Q_{\text{хк}}$ є надто грубою, оскільки випробування не є однорідними: згенеровані варіації так само, як і слова словника, мають різну довжину і різну лексикографічну вразливість у сенсі можливостей взаємних збігів. Для підвищення адекватності моделі (7) регістр (рис. 4) слід розглядати у двох вимірах (номер комірки і довжина комірки), а значення V і r — індивідуально для кожного слова словника і варіації помилкового слова.

Припустімо заданими ймовірність β_{1j} збігу зі словником помилкового слова та ймовірність β_{2j} збігу варіації помилкового слова $\tilde{\tilde{A}}_j$. Тоді у відповідності з логічними виразами для подій (див. рис. 2) і моделі випробувань (7) можемо записати такі вирази для ймовірності окремих подій:

$$Q_{j\text{нв}} = \beta_{1j} P;$$

$$Q_{j\text{пк}} = (1 - Q_{j\text{нв}}) P \left[(1-\beta_{2j})^{V_j - 1} + \sum_{g=1}^{V_j - 1} \frac{1}{g+1} P(g, \beta_{2j}, V_j - 1) \right];$$

$$Q_{j\text{пк}} = (1 - Q_{j\text{нв}}) P \sum_{g=1}^{V_j - 1} \frac{g}{g+1} P(g, \beta_{2j}, V_j - 1);$$

$$Q_{j\text{нк}} = (1 - Q_{j\text{нв}})(1 - P).$$

Під час виведення виразів ураховано, що із z можливих збігів слів, що перевіряються, одне правильне, таке, що відповідає спотвореному слову $\overline{A}_j$, і g випадкових збігів — неправильні. Правильним є коригування в разі, якщо $g = 0$ (імовірність події $P(0, \beta_{2j}, V_j - 1) = (1-\beta_{2j})^{V_j - 1}$, або якщо з $g+1$ варіантів зроблено правильний вибір (імовірність $\frac{1}{g+1}$).

Для визначення величин β_{1j} і β_{2j} розглянемо таку інтерпретацію залежності значень імовірності $\beta_{xj}(x)$ збігу зі словником x разів спотвореного типовою помилкою слова A_j (рис. 5).

Якщо $x = 0$, то $\beta_{0j} = 1$, оскільки неспотворене слово безперечно збігається зі словником.

Якщо $x = 1$, величина β_{1j} дорівнює відносній кількості збігів слів $\overline{A}_j$, спотворених типовими помилками. Ця величина визначається безпосередньо за допомогою імітаційної моделі (табл. 4).

Якщо $x = m_j >> 1$, величина β_{mj} асимптотично прагне до значення

$$r_j = \frac{\hat{N}(n_j)}{q^{nj}},$$

де $\hat{N}(n_j)$ — кількість слів словника довжиною $n_j \pm 1$.

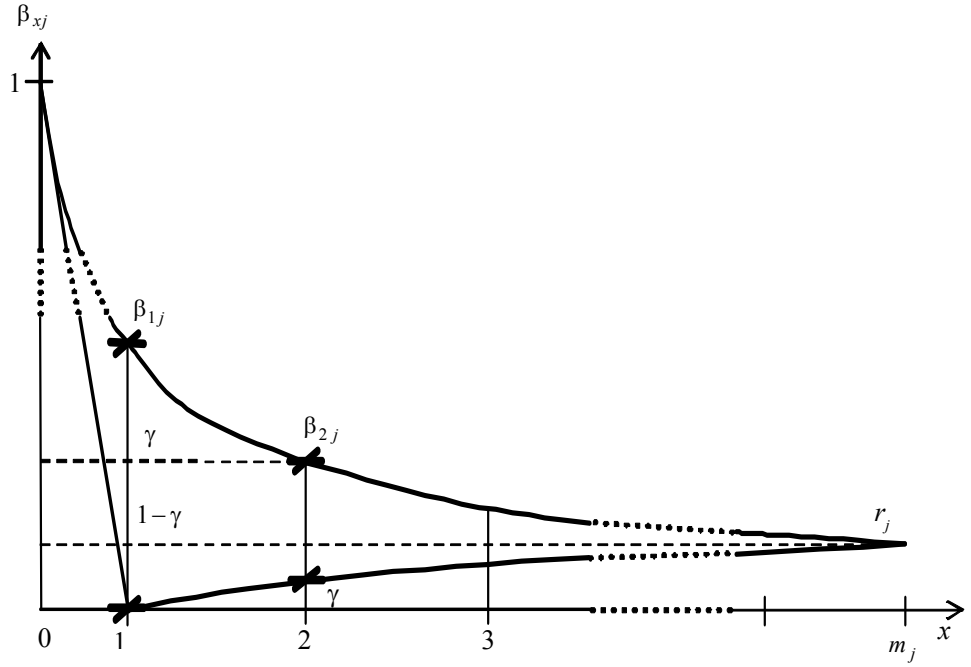

Рис. 5. Графічна інтерпретація залежності значення ймовірності збігу зі словником

На підставі попередніх міркувань для $x = 2$ покладемо

$\beta_{2j} := \beta_{1j} - \left(\beta_{1j} - \dfrac{\hat{N}(n_j)}{q^{nj}}\right)\gamma$, де коефіцієнт γ визначає крутизну спадання кривої $\beta_{xj}(x)$.

Для розрахунків за моделлю беремо

$$\beta_{1j} = \frac{\sum_k v_{kj}}{V(n_j)},$$

де v_{kj} — кількість збігів зі словником слова $\overline{A}_j$, спотвореного типовою помилкою E_k; $V(n_j) = V_j$ — сумарна кількість різноманітних типових помилок слова довжиною n_j символів;

$$\hat{N}(n_j) = N(n_j)\frac{V_1(n_j) + V_4(n_j)}{V(n_j)} + N(n_j - 1)\frac{V_2(n_j)}{V(n_j)} + N(n_j + 1)\frac{V_3(n_j)}{V(n_j)},$$

де $N(n_j)$, $N(n_j - 1)$, $N(n_j + 1)$ — фактична кількість слів довжиною n_j, $n_j - 1$, $n_j + 1$; $\gamma = 0{,}87$.

У виразі для $\hat{N}(n_j)$ взято до уваги зміну довжини слова $\overline{A}_j$, спотвореного пропусками та вставками символів, а значення коефіцієнта γ підбиралось у процесі моделювання (за траєкторією $0{,}9 \to 0{,}88 \to 0{,}86 \to 0{,}87$).

Результати моделювання наведено в табл. 5.

Таблиця 5. Результати моделювання для натурно-аналітичного коригування

Словник	$Q_{0\text{пк}}$	$Q_{\text{пк}}$	$Q_{\text{хк}}$	$Q_{\text{нв}}$
Словник російської літератури, $N = 161730$	0,7490	0,7355	0,1470	0,0184
Словник Лопатіна, $N = 150213$	0,8373	0,8323	0,0628	0,0600
Словник Залізняка, $N = 92555$	0,8383	0,8338	0,0614	0,0054
Словник Лопатіна скорочений, $N = 84575$	0,8608	0,8576	0,0394	0,0038
Українська версія скороченого Словника Лопатіна, $N = 84575$	0,8698	0,8674	0,0300	0,0028

Як видно з даних табл. 2, результати розрахунків за аналітичною моделлю близькі до результатів табл. 4. Так, відхилення значень основного показника коригувальних властивостей $Q_{\text{пк}}$ становить 0,75% для словника російської літератури і не перевищує 1,2% для інших словників. При цьому оброблення словника потребує на порядки менше часу — для використовуваного малопотужного комп'ютера і послідовної схеми моделювання час оброблення одного слова за імітаційною моделлю становило 6 с, а за аналітичною моделлю — 0,04 с. Крім того, відхилення могло б бути ще меншим (до 1%) за ретельнішого підбору значень γ. Із порівняльних даних випробувань і їх інтерполяційних оцінок випливає, що оптимальне значення, яке відповідає мінімальному сумарному відхиленню, $\gamma = 0{,}865 - 0{,}867$. Істотно, що відхилення результатів мало залежить від лінгвістичної структури і змісту словників, їх обсягів і мов. Подібна «універсальність» дає підстави для підтвердження правомірності пропонованого підходу до побудови аналітичної моделі. У свою чергу, це означає, що аналітична модель може бути покладена в основу оцінки коригувальних властивостей щодо більш складних помилок, що мають відстань Дамерау–Левенштейна більшу, ніж типові помилки.

ВИСНОВКИ

1. Подані моделі можуть бути покладені в основу інструменту порівняльної оцінки потенційних контролювальних і коригувальних властивостей конкретного орфографічного словника щодо типових помилок тайпінгу.

В оцінюванні контролювальних властивостей (див. табл. 1–3) моделі та алгоритм дають можливість для конкретного словника отримати дані про значення очікуваного показника дисфункції і можливості його зменшення за рахунок прийнятного підвищення ймовірності неправдивого сигналу про помилковість слова. Такі дані можуть бути корисні для прийняття обґрунтованих рішень для вибору готового словника, що описує задану предметну галузь, оцінювання можливості та доцільності його поліпшення з урахуванням критеріїв ρ, φ та адаптивного супроводу РОС (поповнення, вилучення) на основі критерію (5).

Під час оцінювання коригувальних властивостей моделі дозволяють оцінити потенційне співвідношення ймовірності правильного і неправильного коригування. При цьому аналітична модель може слугувати для попередніх рішень, а імітаційна для уточнених оцінок, повнота яких визначається урахуванням усіх можливих типових помилок і внеску кожного слова

в підсумкове значення $Q_{\text{пк}}$. За заданих імовірностей спотворення слів A_j цей внесок може бути відповідним чином зважений.

2. Існує високий ступінь кореляції між значеннями показників контролювальних і коригувальних властивостей $Q_{\text{нв}}$ і $Q_{\text{пк}}$. З одного боку, це дає підстави вважати, що словники, поліпшені щодо контролювальних властивостей, мають і відповідно кращі коригувальні властивості, з другого боку, говорити про деякий загальний показник орфографічної вразливості словника щодо як до окремих типових помилок, так і їх кратних комбінацій. Якщо взяти за основу прийняту інтерпретацію залежності значень імовірності збігу зі словником від кратності типової помилки (рис. 5) — інтерпретацію, правомірність якої попередньо підтверджують результати моделювання, тоді як загальний показник можна брати значення зваженої ймовірності збігу довільного слова, спотвореного типовою помилкою ансамблю *K*. Кількісна оцінка можливого зв'язку цього показника з коригувальними властивостями словника за іншими функціями попереднього вибору і преференцій (наприклад, властивих застосуванню фонетичних алгоритмів) потребує окремого дослідження.

ЛІТЕРАТУРА

1. *Нечёткий* поиск в тексте и словаре [Електронний ресурс]. — Режим доступу: https://habrahabr.ru/post/114997/.
2. *Расстояние* Левенштейна в MySQL и алгоритмы нечёткого поиска средствами PHP [Електронний ресурс]. — Режим доступу: https://habrahabr.ru/post/342434/.
3. *Фонетические* алгоритмы [Електронний ресурс]. — Режим доступу: https://habrahabr.ru/post/114947/.
4. *Phonetic* Algorithms [Електронний ресурс]. — Available at: https://deparkes.co.uk/2017/12/01/phonetic-algorithms/.
5. *Hodge V.J.* A comparison of standard spell checking algorithms and a novel binary neural approach / V.J. Hodge, J. Austin // IEEE Transactions on Knowledge and Data Engineering. — 2003. — С. 1073–1081.
6. *de Amorim R.C.* Effective Spell Checking Methods Using Clustering Algorithms [Електронний ресурс] / R.C. de Amorim, M. Zampieri. — Available at: http://www.aclweb.org/anthology/R13-1023.
7. *Литвинов В.А.* Оценка контролирующих свойств базового словаря допустимых слов в системе автоматического обнаружения ошибок пользователя / В.А. Литвинов, С.Я. Майстренко, К.В. Хурцилава // Математичні машини і системи. — 2014. — № 2. — С. 65–70.
8. *Словари* русского языка [Електронний ресурс]. — Режим доступу: http://speakrus.ru/dict.
9. *Словарь* Лопатина [Електронний ресурс]. — Режим доступу: http://royallib.ru/book/ lopatin_vladimir/russkiy_orfograficheskiy_slovar.html.
10. *Словари* русского языка [Електронний ресурс]. — Режим доступу: http://speakrus.ru/dict.
11. *Литвинов В.А.* Контроль достоверности и восстановления информации в человеко-машинных системах / В.А. Литвинов, В.В. Крамаренко. — К.: Техніка, 1986. — 200 с.
12. *Литвинов В.А.* Дисфункция референтного словаря системы проверки орфографии и подход к ее снижению / В.А. Литвинов, С.Я. Майстренко, К.В. Хурцилава // Математичні машини і системи. — 2017. — № 2. — С. 39–48.
13. *Knapsack* problem [Електронний ресурс]. — Режим доступа: http://en.wikipedia.org/wiki/Knapsack_problem.
14. *Задача* о рюкзаке: жадный алгоритм [Електронний ресурс]. — Режим доступу: http://traditioru.org/wiki/Задача_о_рюкзаке: жадный_алгоритм.

Надійшла 15.01.2019