https://orcid.org/0000-0003-0817-8957
https://orcid.org/0000-0001-5568-7629

УДК 681.3

С.В. КОСТЕНКО[*], В.А. ЛИТВИНОВ[**]

МОДЕЛЬ ОЦІНКИ ВЛАСТИВОСТЕЙ АЛГОРИТМІВ ВИПРАВЛЕННЯ ОРФОГРАФІЧНИХ ПОМИЛОК ТА ЇЇ ВИКОРИСТАННЯ ДЛЯ УКРАЇНОМОВНИХ ЗАСТОСУНКІВ

[*]Національний університет харчових технологій, м. Київ, Україна
[**]Інститут проблем математичних машин і систем НАН України, м. Київ, Україна

Анотація. *Значна частина підходів та методів автоматичного виправлення помилок правопису є мовозалежною, орієнтованою на врахування граматичних правил і фонетики конкретної мови. Переважна більшість робіт у цій сфері присвячена англомовним текстам, менша частина – іншим мовам германської групи, зовсім мала – слов'янським мовам і мізерна – українській мові. Розглядаються інструментарій (імітаційна модель (ІМ)) та пробні результати оцінки коригуючих властивостей деяких алгоритмів по відношенню до виправлення типових помилок тайпінга в україномовних словах за умови попередньої індексації словника. ІМ має модульну структуру і конфігурується до конкретних словників, типів помилок, алгоритмів генерації індексів. Для заданої комбінації «словник – тип помилок – алгоритм» ІМ спотворює слова заданого словника помилкою і шукає найбільш «близькі» слова-кандидати на виправлення помилкового слова. Загальний алгоритм обробки слова, в якому виявлено помилку, включає попередній вибір (ПВ) множини слів-кандидатів за правилами алгоритму індексації та остаточний вибір (ОВ) – пріоритезація обраних слів і звуження області пошуку «правильного» слова за прийнятими критеріями близькості до слова, що виправляється, і різними критеріями відбору. Пробне моделювання проведено для фонетичних алгоритмів Soundex і Metaphone, адаптованих до української мови. Як ансамбль помилок прийнятий набір помилок тайпінга, що включає чотири різні базові одиночні помилки (заміни, вставки, пропуски і перестановки символів), а також подвійні помилки, що представляють собою зважену суміш базових помилок. Опрацьовано 59,6 млн помилкових слів, середній час обробки одного слова склав 0,07 мс. Обговорюються результати моделювання, що містять загальну кількість коректованих помилкових слів за видами помилок, кількість коректних пропозицій слів-кандидатів, кількість помилкових пропозицій, кількість відсутніх пропозицій, середню кількість кандидатів у пропозиціях на етапах ПВ і ОВ.*

Ключові слова: *перевірка і забезпечення правопису, виправлення помилок у словах української мови, імітаційна модель оцінки алгоритмів індексації словника.*

Abstract. *A significant part of the methods and approaches to automatic spelling correction is language-dependent since they are focused on grammatical rules and phonetics of a particular language. The vast majority of works in this field is devoted to texts in English, a smaller part – to other Germanic languages, a very small part – to Slavic languages and the smallest part – to the Ukrainian language. A special toolkit (a simulation model (SM)) and the evaluation results of the corrective properties of some algorithms in relation to the correction of typical typing errors in Ukrainian words, taking into account the preliminary dictionary indexation, are considered in the article. SM has a modular structure and is configurable for specific dictionaries, types of errors and index generation algorithms. For the «dictionary – type of errors – algorithm» combination SM distorts the words of the given dictionary with a mistake and searches for the «closest» words-candidates to correct the wrong word. A general algorithm for processing a word with an error includes pre-selection (PS) of a set of words-candidates according to the rules of the indexing algorithm, and final selection (FS) as a prioritization of the selected words and narrowing the search area for a «correct» word according to the accepted proximity criteria for the word to be corrected, and various selection criteria. A trial simulation was performed for Soundex and Metaphone phonetic algorithms adapted to the Ukrainian language. An ensemble of errors is a set of typing errors which includes four different basic single misspellings (replacements, insertions, omissions and rearrangement of letters), as well as double errors which are a mixture of basic errors. 59,6 million words with errors were processed; the average time for processing one word was 0,07 ms. The simulation results*

contain the total number of corrected words grouped by the type of error, the number of correct words-candidates offers, the number of incorrect offers, the number of missing offers and the average number of candidates in the offers at the PS and FS stages.
Keywords: *Spell Checking, correction of errors in the words of the Ukrainian language, simulation model for estimating dictionary indexing algorithms.*

DOI: 10.34121/1028-9763-2021-2-62-73

1. Вступ

Задача перевірки і забезпечення коректного правопису слів і речень природної мови в комп'ютерних інформаційних технологіях має більш ніж сорокарічну історію. Одна з перших узагальнюючих публікацій на цю тему належить Д. Петерсону (J.L. Peterson, 1980) [1], згодом було опубліковано величезну кількість робіт. Фундаментальний огляд проблематики і досягнень у даній області був свого часу зроблений К. Кукічем [2] (K.K. Kukich, 1992). З тих пір тематика відповідних досліджень продовжує розвиватися, з'являються нові сфери застосувань (web-сайти, месенджери, голосові дані, системи автоматичного онлайн перекладу та ін.), нові підходи (використання технологій нейронних мереж, вивчення поведінки конкретного користувача та ін.), нові методи і системи. Уявлення про сучасний стан проблематики і відповідне інструментальне забезпечення може дати огляд [3].

Значна частина досліджуваних підходів і методів є мовозалежними – орієнтованими на врахування граматичних правил і фонетики конкретної мови. Загальна задача забезпечення коректного правопису (в ширшому сенсі – це розпізнавання і коректна ідентифікація слів і словосполучень) включає в себе дві підзадачі: виявлення можливих помилок і виправлення виявлених. Виявлення базується на перевірці належності слова до конкретної мови. Зокрема, наприклад, на основі заданого референтного орфографічного словника (РОС). Тут залежність від мови проявляється в меншій мірі, тому методи і алгоритми, розроблені для одних мов, здебільшого прямо застосовні і для інших. Виправлення, необхідність в якому виникає при виявленні помилки, вимагає аналізу помилкового слова на основі граматичних та/або фонетичних правил, специфічних для конкретної мови. Тому на тлі переваги робіт і результатів по англомовних застосуваннях методи і алгоритми виправлення помилок широко розглядаються і для інших мов германської групи, зокрема, німецької, шведської, португальської, іспанської та інших (наприклад, [4–7]). Є окремі роботи, які пропонують метод виправлення фонетичних помилок (фонетичний алгоритм) і для російської мови [8].

Особливості виправлення граматичних і фонетичних помилок в українській мові практично не досліджувалися (відповідних робіт знайти не вдалося), хоча при базовій схожості з російською в українській мові є своя безсумнівна специфіка, особливо з урахуванням розвитку нових правил правопису [9]. Можна згадати лише [10], де аналізуються результати виявлення та виправлення типових помилок тайпінга в російськомовних і україномовних текстах «силовим» методом, а саме генерацією всіх можливих варіантів виправлення, які перебувають на відстані Дамерау-Левенштейна від помилкового слова, що дорівнює 1, і перевіркою по РОС коректності варіанта. Застосування цих результатів обмежується переважно порівняльною оцінкою контролювальних і коригувальних можливостей досліджених словників.

Мета статті полягає в поданні інструментальної моделі, призначеної для досліджень як результативності алгоритмів виправлення різних видів помилок в україномовних словах, так і коригувальних властивостей різних українських словників.

2. Вихідні положення

Загальні методи ідентифікації та корекції помилок правопису поділяються на дві основні групи:

- автономна корекція окремих слів за референтним орфографічним словником;
- контекстно-залежна корекція.

У зв'язку з тим, що автономна корекція є складовою частиною контекстно-залежної корекції та з огляду на недостатню загальну вивченість україномовних застосувань, завдання моделювального комплексу обмежимо дослідженням автономної корекції. Виділимо три фактори, які є важливими для цільових властивостей моделювання процесу виявлення і виправлення помилок:

- типи помилок;
- алгоритми обробки помилкового слова;
- критерії оцінки властивостей і результативності процесу.

2.1. Типи помилок

З усього різноманіття помилок в окремому слові виділимо три основних типи: механічні помилки (стосовно до введення із клавіатури – помилки тайпінга), фонетичні і когнітивні помилки.

Помилки тайпінга виникають, коли користувач (оператор) знає правильне написання слова, але допускає порушення коректної моторики (послідовності рухів) у процесі «спілкування» із клавіатурою. Наприклад, замість клавіші «і» натискає сусідню клавішу «в» і слово *піарити* перетворюється у слово *пварити*. Такі помилки, що дуже мало залежать від мови, є найбільш вивченими. Вони, як відомо, поділяються на 4 базових підтипи: заміни, вставки, видалення і перестановки символів. Імовірності «спотворення» слова такими помилками в межах кожного підтипу є максимальними для одиночних помилок та швидко зменшуються з ростом їх кратності. Вважалося [2], що при наборі тексту близько 80% помилок є поодинокими. В оцінках експериментів [11] по набору російськомовних текстів ця цифра становила близько 90%, а розподіл одиночних помилок за підтипами характеризувався такими значеннями: заміни символу – близько 56%, додавання – 15%, пропуск – 12%, перестановка – до 7%.

Фонетичні помилки пов'язані з фонетично коректним, але орфографічно неправильним написанням слова. Наприклад, помилки *<коритце – корицце>*, *<чекати – чикати>*. Часто фонетичні помилки виникають у результаті явища, що дістало назву «Метатеза» – зміна у слові менш зручних для вимови поєднань звуків більш зручними. Метатеза властива для маловідомих, етимологічно непрозорих, іншомовних слів [12].

У разі когнітивних помилок у джерела формування слова відсутнє знання про його правильне написання, у зв'язку з чим виникають помилки типу *<бароко – барокко>*, *<пріоритезація – приоретизація>*.

Прийнята класифікація дає уявлення про походження і механізми помилок, але не завжди дозволяє чітко віднести конкретну помилку до одного з зазначених типів, особливо в разі когнітивної помилки і, наприклад, помилки-метатези. Тому в оцінці коригувальних властивостей методів і словників української мови при дослідженні алгоритмів обмежимося помилками тайпінга і фонетичними помилками, для яких існують загальні правила їх визначення: помилки тайпінга – це сукупності механічних одиночних і багаторазових базових спотворень символів (заміни, вставки, видалення, перестановки), а фонетичні помилки – це спотворення слів, пов'язані з порушеннями фонетичних правил української мови.

2.2. Алгоритми обробки помилкового слова

Приймемо в подальшому такі позначення:

A_j – слово РОС $(j = 1..N)$;

$\bar{A}_j$ – слово A_j, спотворене помилкою.

Загальний алгоритм виправлення слова, в якому виявлено помилку, включає таку послідовність етапів:

1) Попередній вибір в РОС множини слів-кандидатів, які відповідають прийнятим критеріям близькості до слова $\bar{A}_k$.

2) Пріоритезація обраних слів і звуження області пошуку «правильного» слова A_k у множині слів-кандидатів.

3) Прийняття рішення про виправлення.

Третій етап базується на специфічних для конкретного додатку формальних та евристичних правилах, які враховують, зокрема, і можливість участі користувача у прийнятті рішень. Тому при оцінці властивостей конкретних алгоритмів представляється доцільним обмежитися аналізом перших двох етапів, для реалізації яких можливе використання деяких загальних формальних положень.

Найбільш простий підхід до реалізації перших двох етапів полягає в застосуванні «грубої сили» – послідовного перебору слів A_j і визначення їх близькості до слова A_k на основі деякої метрики, зокрема, наприклад, відстані Дамерау-Левенштейна (ВДЛ). У зв'язку з високою трудомісткістю процедури обчислення ВДЛ такий підхід практично можливий лише у разі малих N. Можлива й інша реалізація силового підходу – зворотний перебір всеможливих помилок $\bar{A}_k \to A_k$ (варіацій помилкового слова $\bar{A}_k$) [10], що знаходяться в межах заданого ВДЛ від A_k. Така реалізація менш трудомістка, але практично можлива лише для $ВДЛ(\bar{A}_k, A_k) = 1$.

Конструктивний підхід до пошуку слів-кандидатів у словниках великого обсягу полягає в попередній індексації РОС – розбитті слів на групи з однаковим значенням індексу (ключа) K_j, який формується за допомогою алгоритму $H(A_j)$, що реалізує, по суті, деяку функцію хешування. Для помилкового слова $\bar{A}_k$ обчислюється $K_k = H(\bar{A}_k)$, і якщо для нього існує $H(A_k) = H(\bar{A}_k)$, отже, успіх, правильне слово знаходиться у відповідній групі РОС і може бути знайденим у результаті аналізу слів даної групи. Приклад обробки помилкового слова показаний на рис. 1.

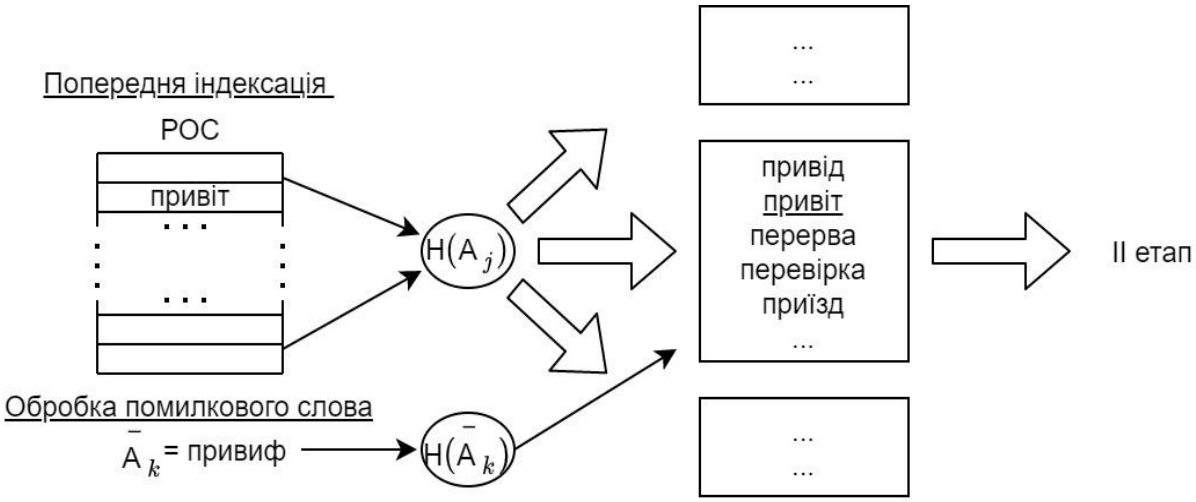

Рисунок 1 – Процес обробки помилок

2.3. Критерії оцінки результатів

У результаті коригування помилкового слова $\bar{A}_j$ можливі випадки, показані на рис. 2:

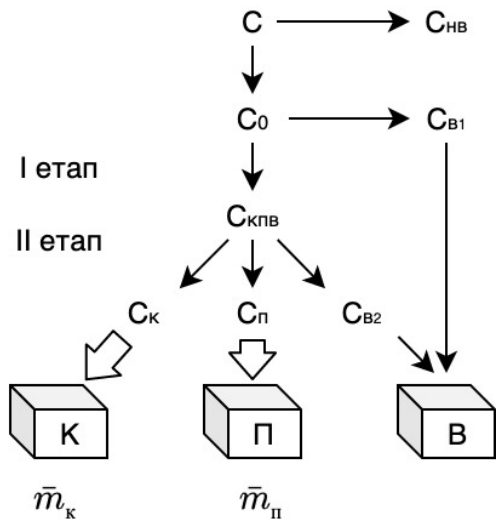

Рисунок 2 – Узагальнена модель подій

• помилку $A_к \to \bar{A}_к$ не було виявлено (фінальна подія C_{HB}, ймовірність Q_{HB});

• помилку $A_к \to \bar{A}_к$ виявлено (подія C_0), ключ відсутній у переліку ключів індексованих груп слів РОС, коригування неможливе (подія C_{B1});

• помилку $A_к \to \bar{A}_к$ виявлено, в РОС знайдена група з відповідним ключем, пропонується набір можливих кандидатів на виправлення (подія $C_{КПВ}$);

• у пріоритезованій групі запропонованих кандидатів є правильне слово (подія C_K, ймовірність Q_K);

• у пріоритезованій групі запропонованих кандидатів правильне слово відсутнє (подія $C_П$, ймовірність $Q_П$);

• пріоритезована група запропонованих кандидатів порожня (подія C_{B2}).

Результативність процесу коригування для заданого РОС, ансамблю помилок і алгоритму обробки помилкового слова визначається:

• ймовірностями (відносною кількістю) подій C_K, $C_П$ і середніми обсягами відповідних груп $\bar{m}_K, \bar{m}_П$;

• ймовірністю (сумарною відносною кількістю) подій C_{B1}, C_{B2}.

3. Структура моделі та алгоритм моделювання

3.1. Структура моделі

Комплекс моделювання має модульну структуру (рис. 3), де кожен із компонентів може бути заміненим. Це дозволяє йому бути легко конфігурованим та пристосовним до конкретних словників, типів помилок, алгоритмів індексації.

Основні компоненти комплексу:

• РОС – референтний орфографічний словник, що використовується при моделюванні.

• Постачальник РОС – модуль роботи зі словниками.

• Генератор помилок – модуль, що виконує спотворення слів РОС у відповідності з прийнятим типом помилок.

• Алгоритм кодування – модуль, що виконує обчислення ключа для попередньої індексації слів РОС і обробки помилкового слова.

• Модуль обчислення показників моделювання.

Моделювальний комплекс зв’язує дані компоненти в єдину систему.

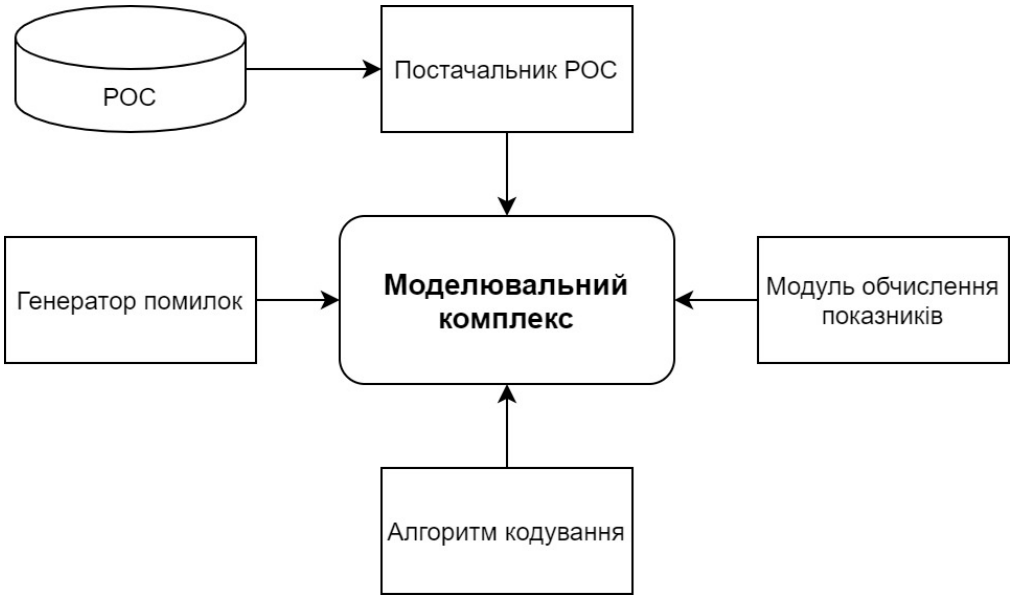

Рисунок 3 – Структура комплексу моделювання

3.2. Загальний алгоритм моделювання

Процес моделювання для заданої трійки «словник, тип помилок, алгоритм індексації» включає такі етапи.

1. Підготовка даних для моделювання.

1.1.Зчитування словника та завантаження в пам’ять.

1.2. Підготовка слів – очистка від можливих пробілів та спецсимволів.

1.3. Індексація словника за заданим алгоритмом індексації – створення структури <Ключ, Слово>.

2. Обробка помилкових слів (ПС) чергового слова РОС.

2.1. Генерація масиву ПС для чергового слова РОС.

2.2. Виключення ПС, що співпадають із деяким реальним словом РОС. Відповідні помилки не викриваються і не підлягають коригуванню.

2.3. Моделювання процесу виправлення чергового ПС.

2.3.1. Обчислення ключа ПС за заданим алгоритмом індексації.

2.3.2. Пошук і аналіз індексованої групи РОС із обчисленим ключем.

2.3.3. Фіксація поточних результатів у відповідності з подіями (рис. 2).

3. Повторення п. 2.3 для всіх ПС чергового слова і п. 2.1 для всіх слів РОС.

4. Обчислення показників K, $П$, B, $\bar{m}_K$, $\bar{m}_П$ для РОС у цілому.

4. Результати моделювання обраних алгоритмів виправлення помилок

4.1. Алгоритми, помилки та критерії відбору

Пробне моделювання проведено для таких вихідних умов.

1) Як досліджувані алгоритми індексації прийняті оригінальні англомовні алгоритми Soundex і Metaphone. Вибір обумовлений, з одного боку, їх широкою популярністю і поширеністю, а з іншого боку, наявністю в бібліотеках готових модулів програмної реалізації (зокрема, для англійської мови). Адаптацію до української мови здійснено на основі офіційної англо-української транслітерації, затвердженої Кабінетом Міністрів України [13]. Приклади транслітерації конкретних слів: *<займатися – zaimatysia>*, *<ідентифікатор – identyfikator>*, *<співочий – spivochyi>*, *<зв’язок – zviazok>*. Результати моделювання транслітерованих алгоритмів мають створити певну точку відліку для порівняльної оцінки результативності україномовних алгоритмів.

Правила обробки слів A_j, A_k і обчислення значень ключів в оригінальних англомовних алгоритмах Soundex і Metaphone наведені в багатьох джерелах, наприклад, у [14].

2) Як ансамбль помилок взято набір помилок тайпінга, що включає чотири всіляких базових одиночних помилки, а також подвійні помилки, що представляють собою зважену суміш базових помилок. Прийнято такий алгоритм формування подвійних помилок.

1. Фіксується n – довжина слова.

2. Випадковим чином вибирається тип першої помилки (з імовірністю $\pi_i^{(1)}$).

3. Слово A_j спотворюється першою помилкою $(A_j \to \bar{A}_j)$.

4. Випадковим чином вибирається тип другої помилки (з імовірністю $\pi_i^{(2)}$).

5. Слово A_j спотворюється другою помилкою $(\bar{A}_j \to \bar{\bar{A}}_j)$.

Визначається $ВДЛ(\bar{\bar{A}}_j, A_j)$. Якщо $ВДЛ(\bar{\bar{A}}_j, A_j) = 2$, слово з помилкою додається в результуючий набір помилкових слів, інакше – друга помилка генерується заново (перехід до п. 4).

Операції пп. 2–5 повторюються n раз. Вектор π_1, π_2, π_3, π_4 (заміна, пропуск, додавання, перестановка символу) сформований з урахуванням відповідних даних [11], використаних в [10].

Вибір типу помилок для пробного моделювання обумовлений визначеністю і простотою їх генерації. Хоча алгоритми Soundex і Metaphone орієнтовані в першу чергу на виправлення фонетичних помилок, їх використання до помилок тайпінга дає певну інформацію як для аналізу властивостей фонетичних алгоритмів, так і можливого вдосконалення алгоритмів виправлення інших типів помилок.

3) Для генерації конкретних помилок використано готовий словник українських слів обсягом 84575 слів, запозичений з [10].

4) Розглядаються результати моделювання та оцінки двох за припущенням конкурентоздатних критеріїв вибору слів-кандидатів на другому етапі відбору:

1. $ВДЛ = ВДЛ_{\min}$.

2. $ВДЛ \le ВДЛ_{\max}$, де $ВДЛ_{\max}$ є конвенційним максимальним ВДЛ для помилкових слів (зокрема, для досліджуваного ансамблю помилок $ВДЛ_{\max} = 2$).

4.2. Результати моделювання оригінальних алгоритмів

У табл. 1–6, що відображають результати моделювання, прийняті такі позначення:

C – кількість згенерованих помилкових слів;

C_0 – кількість слів для корекції (абсолютна і відносна по відношенню до C);

K – кількість коректних пропозицій (абсолютна і відносна по відношенню до C_0);

X – кількість хибних пропозицій (абсолютна і відносна по відношенню до C_0);

B – кількість відсутніх пропозицій (абсолютна і відносна по відношенню до C_0);

$\overline{K}$ – середній обсяг групи коректної пропозиції;

$\overline{X}$ – середній обсяг групи хибної пропозиції.

Таблиця 1 – Soundex – оригінальний транслітерований попередній відбір (ПВ)

Тип помилки	C	C_0	K	X	B	$\overline{K}$	$\overline{X}$
Заміна символу	26856042	26828123	14525821	11846740	455562	76,33	47,02
		0,9990	0,5409	0,4411	0,0170		
Додавання символу	30485400	30480461	20869394	9202342	408725	74,34	48,1
		0,9998	0,6846	0,3019	0,0134		
Пропуск символу	839225	834286	514139	311140	9007	74,1	41,82
		0,9941	0,6126	0,3707	0,0107		
Перестановка символів	754650	745422	569414	168686	7322	74,15	39,14
		0,9878	0,7545	0,2235	0,0097		
Подвійні помилки	839225	834535	281168	530358	23009	78,87	44,86
		0,9944	0,3350	0,6320	0,0274		

Таблиця 2 – Soundex – оригінальний транслітерований остаточний відбір (ОВ), критерій 1

Тип помилки	C	C_0	K	X	B	$\overline{K}$	$\overline{X}$
Заміна символу	26856042	26828123	14525821	11846740	455562	1,02	2,08
		0,9990	0,5409	0,4411	0,0170		
Додавання символу	30485400	30480461	20869394	9202342	408725	1	2,08
		0,9998	0,6846	0,3019	0,0134		
Пропуск символу	839225	834286	514139	311140	9007	1,07	2,02
		0,9941	0,6126	0,3707	0,0107		
Перестановка символів	754650	745422	569414	168686	7322	1,01	2,03
		0,9878	0,7545	0,2235	0,0097		
Подвійні помилки	839225	834455	279861	531636	22958	1,05	2,15
		0,9943	0,3335	0,6335	0,0274		

Таблиця 3 – Soundex – оригінальний транслітерований остаточний відбір (ОВ), критерій 2

Тип помилки	C	C_0	K	X	B	$\overline{K}$	$\overline{X}$
Заміна символу	26856042	26828123	14525821	1138251	11164051	1,24	1,93
		0,9990	0,5409	0,0424	0,4157		
Додавання символу	30485400	30480461	20869394	314632	9296435	1,15	1,52
		0,9998	0,6846	0,0103	0,3049		
Пропуск символу	839225	834286	514139	76295	243852	1,46	2,54
		0,9941	0,6126	0,0909	0,2906		
Перестановка символів	754650	754650	745422	569414	13446	1,23	1,79
		1,0000	0,9878	0,7545	0,0178		
Подвійні помилки	839225	834535	281168	35248	518119	1,12	2,26
		0,9944	0,3350	0,0420	0,6174		

Таблиця 4 – Metaphone – оригінальний транслітерований попередній відбір (ПВ)

Тип помилки	C	C_0	K	X	B	$\overline{K}$	$\overline{X}$
Заміна символу	26856042	26828123	13283355	10324630	3220138	28,92	15,03
		0,9990	0,4946	0,3844	0,1199		
Додавання символу	30485400	30480461	19397238	8071770	3011453	27,86	14,75
		0,9998	0,6363	0,2648	0,0988		
Пропуск символу	839225	834286	500592	269586	64108	27,89	14,87
		0,9941	0,5965	0,3212	0,0764		

Продовж. табл. 4

Перестановка символів	754650	745422	542194	142838	60390	27,89	14,87
		0,9878	0,7185	0,1893	0,0800		
Подвійні помилки	839225	834558	241688	435172	157698	30,37	14,41
		0,9944	0,2880	0,5185	0,1879		

Таблиця 5 – Metaphone – оригінальний транслітерований остаточний відбір (ОВ), критерій 1

Тип помилки	C	C_0	K	X	B	$\overline{K}$	$\overline{X}$
Заміна символу	26856042	26828123	13283355	10324630	3220138	1,02	1,65
		0,9990	0,4946	0,3844	0,1199		
Додавання символу	30485400	30480461	19397238	8071770	3011453	1	1,65
		0,9998	0,6363	0,2648	0,0988		
Пропуск символу	839225	834286	500592	269586	64108	1,06	1,65
		0,9941	0,5965	0,3212	0,0764		
Перестановка символів	754650	745422	542194	142838	60390	1,01	1,62
		0,9878	0,7185	0,1893	0,0800		
Подвійні помилки	839225	834558	241039	435821	157698	1,04	1,68
		0,9944	0,2872	0,5193	0,1879		

Таблиця 6 – Metaphone – оригінальний транслітерований остаточний відбір (ОВ), критерій 2

Тип помилки	C	C_0	K	X	B	$\overline{K}$	$\overline{X}$
Заміна символу	26856042	26828123	13283355	833386	12711382	1,2	1,7
		0,9990	0,4946	0,0310	0,4733		
Додавання символу	30485400	30480461	19397238	223240	10859983	1,13	1,39
		0,9998	0,6363	0,0073	0,3562		
Пропуск символу	839225	834286	500592	61245	272449	1,36	2,18
		0,9941	0,5965	0,0730	0,3246		
Перестановка символів	754650	745422	542194	10223	193005	1,18	1,59
		0,9878	0,7185	0,0135	0,2558		
Подвійні помилки	839225	834611	241390	25555	567666	1,09	1,95
		0,9945	0,2876	0,0305	0,6764		

Наведені дані свідчать про таке.

1. Результативність досліджених алгоритмів по відношенню до помилок тайпінга порівняно невисока – помітно нижче, ніж для перебору варіантів корекції [10]. Там значення ймовірності правильної корекції одиночних помилок, зважене за прийнятими ймовірностями їх появи, склало величину 0,858. Як ми бачимо, для обох алгоритмів (особливо для Metaphone) для всіх типів помилок відносна кількість коректних пропозицій помітно нижча, а аналогічне зважене середнє становить 0,562 для Soundex і 0,520 для Metaphone. Оскільки досліджені фонетичні алгоритми від самого початку орієнтовані більшою мірою на корекцію багаторазових помилок фонетичного походження, цей порівняльний результат в якісному сенсі є очікуваним і дає лише відповідні орієнтовні кількісні оцінки.

2. В основу оцінки результативності критеріїв 1 та 2 покладене міркування, що якщо в відповідній групі пропозиції ПВ правильне слово для коректування відсутнє, то для цієї пропозиції краще статус «пропозиція відсутня», ніж «помилкова (хибна) пропозиція».

Отже, оцінка результативності обраних критеріїв здійснюється за співставленням кількості коректних, хибних та відсутніх пропозицій з урахуванням середніх обсягів відповідних груп. У табл. 7 наведено дані, які відображають результативність критеріїв на прикладі моделювання оригінального ФА Soundex з англо-українською транслітерацією для 5 типів помилок, відносно яких прийняті такі позначення: 1 – заміна символу, 2 – вставки, 3 – пропуски, 4 – перестановки, 5 – подвійні помилки.

Таблиця 7 – Результати моделювання оригінального ФА Soundex

Тип	К-ть $*10^6$	Критерій 1						Критерій 2					
		Розподіл пропозицій $*10^6$			Обсяг пропозицій		S	Розподіл пропозицій $*10^6$			Обсяг пропозицій		S
		K	X	B	$\overline{K}$	$\overline{X}$		K	X	B	$\overline{K}$	$\overline{X}$	
1	26,8	14,5	11,8	0,45	1,02	2,08	39,33	14,5	1,13	11,2	1,24	1,93	20,16
2	30,5	20,8	9,2	0,41	1	2,08	39,94	20,8	0,31	9,3	1,15	1,52	24,39
3	0,83	0,51	0,31	0,01	1,07	2,02	1,17	0,51	0,01	0,24	1,46	2,54	0,77
4	0,74	0,57	0,17	0,01	1,01	2,03	0,92	0,57	0,01	0,16	1,23	1,79	0,72
5	0,83	0,28	0,53	0.02	1,05	2,15	1,43	0,28	0,03	0,52	1,12	2,26	0,38

Як бачимо з даних таблиці, застосування критерію 2 пов’язане з істотно меншим сумарним обсягом (S) обробки коректних і помилкових пропозицій на заключному етапі прийняття конкретного рішення щодо корекції помилкового слова. Ця властивість пов’язана з «перетіканням» на етапі ОВ деякої частини пропозицій X до групи В, яка не потребує аналізу та прийняття рішення. Загальні «баланси» обсягів пропозицій K, X та В на етапах ПВ та ОВ у цілому характеризуються такими співвідношеннями.

Для критерія $ВДЛ = ВДЛ_{\min}$:

$$K_{ОВ} \le K_{ПВ}, X_{ОВ} \ge X_{ПВ}, B_{ОВ} = B_{ПВ}, K_{ОВ} + X_{ОВ} + B_{ОВ} = B_{ПВ} + K_{ПВ} + B_{ПВ}.$$

Для критерія $ВДЛ \le 2$:

$$K_{ОВ} = K_{ПВ}, X_{ОВ} \le X_{ПВ}, B_{ОВ} \ge B_{ПВ}, K_{ОВ} + X_{ОВ} + B_{ОВ} = B_{ПВ} + K_{ПВ} + B_{ПВ}.$$

Слово	Soundex	Metaphone
ліфуваьник	л150	ЛФФНК
шліфувалнк	ш414	ШЛФФЛНК
шлфувальнк	ш414	ШЛФФЛНК
шліфувашьник	ш412	ШЛФФШНК
шліувальнки	ш414	ШЛФЛНК
шліуфвалник	ш414	ШЛФЛНК
шліфувальцик	ш414	ШЛФФЛЦК
шліуфувуальник	ш414	ШЛФФЛНК
ліфувальинк	л145	ЛФФЛНК
лшіфвуальник	л214	ЛШФЛНК
шліфвулаьник	ш414	ШЛФЛНК
шліфуваьинк	ш414	ШЛФФНК

6 різних значень з 12

9 різних значень з 12

Рисунок 4 – Подвійні помилки для слова *шліфувальник*, $n = 12$

Наступний приклад ілюструє згадану властивість «перетікання» для помилкового слова *дуфер* (помилка *буфер – дуфер*), ФА – транслітерований Metaphone, ключ – TFR.

Група $X_{ПВ}$={*двір* (ВДЛ=3), *двері* (ВДЛ=3), *довіру* (ВДЛ=4), *тавро* (ВДЛ=5), *твір* (ВДЛ=4), *тварь* (ВДЛ=5), *тефра* (ВДЛ=4)}.

Таким чином, для критерія 1) $X_{ОВ}$={*двір, двері*}, $B_{ОВ}$={0}. У результаті маємо помилкову пропозицію.

Для критерія 2) $X_{ОВ}$={0}, $B_{ОВ}$={*двір*, *двері*, *довіру*, *тавро*, *твір*, *тварь*, *тефра*}. В результаті пропозиція відсутня.

Отже, розглянуті особливості застосування критеріїв ОВ можуть бути використ-

тані для зниження загальної трудомісткості обробки помилкового слова при застосуванні алгоритмів коректування з попередньою індексацією словника.

3. Із загальних властивостей процесу індексації РОС випливає, що результативність залежить від двох взаємопов'язаних факторів – від здатності алгоритму групувати близькі слова в однойменні групи і від обсягів груп, що утворюються. Якщо порівнювати перший етап відбору з пострілом із рушниці по певній мішені, то якість алгоритму можна зіставити з якістю прицілу, а обсяг групи з розміром мішені. Чим вище, тим більша ймовірність потрапити в мішень. При цьому обидва чинники пов'язані з довжиною формованого алгоритмом ключа – чим довше ключ, тим більше підстав очікувати менших обсягів груп. Це положення ілюструє фрагмент результатів обчислення ключів для помилкових слів, наведений на рис. 4. Тому важливим завданням при підборі функції перетворення «слово – ключ» є забезпечення балансу, що визначає прийнятне поєднання результативності та продуктивності.

4. У цілому загальні шляхи підвищення результативності розглянутих алгоритмів індексації полягають, з одного боку, у використанні безпосередньо української літерації, а з іншого, в реалізації алгоритмів, більш пристосованих до ансамблю помилок, ніж ФА до помилок тайпінга. Попередні результати по першому шляху для однієї із пробних україномовних адаптацій ФА Soundex наведені в табл. 8, 9.

Таблиця 8 – Soundex – україномовний попередній відбір (ПВ)

Тип помилки	C	C_0	K	X	B	$\overline{K}$	$\overline{X}$
Заміна символу	26856042	26828123	16813603	9613902	400618	342,29	186,3
		0,9990	0,6261	0,3580	0,0149		
Додавання символу	30485400	30480461	21809905	8283848	386708	336,79	182,23
		0,9998	0,7154	0,2717	0,0127		
Пропуск символу	839225	834286	591213	238809	4264	336,79	182,23
		0,9941	0,7045	0,2846	0,0051		
Перестановка символів	754650	745422	612300	128452	4670	338,19	146,59
		0,9878	0,8114	0,1702	0,0062		
Подвійні помилки	839225	834567	350610	461832	22125	338,19	146,59
		0,9944	0,4178	0,5503	0,0264		

Таблиця 9 – Soundex – україномовний, остаточний відбір (ОВ), критерій 2

Тип помилки	C	C_0	K	X	B	$\overline{K}$	$\overline{X}$
Заміна символу	26856042	26828123	16813603	979279	9035241	1,31	2,09
		0,9990	0,6261	0,0365	0,3364		
Додавання символу	30485400	30480461	21809905	324152	8346404	1,18	1,62
		0,9998	0,7154	0,0106	0,2738		
Пропуск символу	839225	834286	591213	61550	181523	1,63	2,77
		0,9941	0,7045	0,0733	0,2163		
Перестановка символів	754650	745422	612300	11316	121806	1,28	1,85
		0,9878	0,8114	0,0150	0,1614		
Подвійні помилки	839225	834567	350610	32156	451801	1,14	2,26
		0,9944	0,4178	0,0383	0,5384		

Як видно, відносна кількість коректних пропозицій для всіх типів помилок помітно краща, ніж для оригінального транслітерованого алгоритму.

5. Висновки

Таким чином, є підстави вважати, що наведений комплекс може слугувати інструментальною основою моделювання системи автоматичного виправлення орфографічних помилок у задачах нечіткого інформаційного пошуку слів у тексті та словнику. Імітаційна модель орієнтована на методи, пов'язані з попередньою індексацією словника великого обсягу, і дає можливість оцінити очікувані коригувальні властивості основних компонентів системи – алгоритмів індексації і словників – по відношенню, зокрема, до україномовних застосувань. Попередні експерименти продемонстрували можливості моделі і намітили шляхи її подальшого адекватного застосування.

Задачі подальших досліджень в обраному напрямі полягають в аналізі наявних словників сучасної української мови, формуванні тестових ансамблів типових помилок для конкретних умов їх формування (зокрема, і фонетичних помилок), розробці і оцінці порівняльної ефективності алгоритмів стосовно україномовних застосувань.

СПИСОК ДЖЕРЕЛ

1. Peterson J.L. Computer programs for detecting and correcting spelling errors. *Communications of the A.C.M.* 1980. Vol. 23 (12). P. 676–687.
2. Kukich K.K. Technique for automatically correcting words in text. *ACM Computing Surveys.* 1992. Vol. 24 (4). P. 377–439.
3. The Top 31 Spellcheck Open Source Projects. URL: https://awesomeopensource.com/projects/ spell-check.
4. Rimrott A., Heift T. Evaluating automatic detection of misspellings in German. *Language Learning & Technology*. 2008. Vol. 12, N 3. P. 73–92.
5. Sarr M. Improving Precision and Recall Using a Spellchecker in a Search Engine. Department of Numerical Analysis and Computer Science. Stockholm Royal Institute of Technology, Stockholms Universitet, 2003. 39 p.
6. Andrade G., Teixeira F., Xavier C.R., Oliveira R.S., Rocha L.C., Evsukoff A.G. HASCH: High Performance Automatic Spell Checker for Portuguese Texts from the Web. *Procedia Computer Science.* 2012. Vol. 9. P. 403–411.
7. Aqeel S.U., Beitzel S., Jensen E., Grossman D., Frieder O. On the Development of Name Search Techniques for Arabic. *Journal of the American Society for Information Science and Technology*. 2006. Vol. 57, Issue 6. P. 728–739.
8. Каньковский П. Русский метафон. URL: http://forum.aeroion.ru/topic461.html.
9. Український правопис. 14 нових правил. URL: https://glavcom.ua/country/science/ukrajinskiy-pravopis-14-novih-pravil-yaki-varto-zapamyatati--596631.html.
10. Литвинов В.А., Майстренко С.Я., Хурцилава К.В., Костенко С.В. Критерии и модели оценки корректирующих свойств референтного орфографического словаря при автоматическом исправлении типовых ошибок пользователя. *Математичні машини і системи*. 2018. № 2. С. 72–81.
11. Литвинов В.А., Крамаренко В.В. Контроль достоверности и восстановление информации в человеко-машинных системах. Київ: Техніка, 1988. 200 с.
12. Фонетика. Орфоепія. Теоретичний матеріал. URL: https://eschool.dn.ua/mod/book/tool/print/index.php?id=203161.
13. Про впорядкування транслітерації українського алфавіту латиницею. URL: https://zakon.rada.gov.ua/laws/show/55-2010-%D0%BF.
14. Phonetic Matching Algorithms. URL: https://medium.com/@ievgenii.shulitskyi/phonetic-matching-algorithms-50165e684526.

Стаття надійшла до редакції 10.02.2021