UKRAINIAN URL: ЗВОРОТНА ТРАНСЛІТЕРАЦІЯ У БУДОВІ ОНЛАЙН-СЛОВНИКІВ ТА ВЕБ-САЙТІВ

***Макуха** Андрій Вікторович*

Український мовно-інформаційний фонд Національної академії наук України
andriy.makukha@gmail.com

Анотація

У роботі представлено алгоритм Ukrainian URL (UURL) — нову цілком зворотну схему транслітерації алфавітних символів східнослов'янських мов за допомогою обмеженого набору символів ASCII не зарезервованих синтаксисом URI. Проведено порівняння пропонованої схеми транслітерації з більшістю існуючих стандартів транслітерації літер української мови; вказано ряд її переваг. Зокрема показано як схема може бути застосована для підвищення практичності, індексованості та рейтингу україномовних веб-сайтів та онлайн-словників зокрема.

ВСТУП: ТРАНСЛІТЕРАЦІЯ У МЕРЕЖІ

Хороша архітектура веб-сайтів визначається низкою параметрів, серед яких *зручність використання* (usability), *індексовність* (indexability) та задоволення умов для високого *рейтингу* сайту в пошукових системах.

Індексовність означає досяжність кожної сторінки сайту для ботів пошукових систем. Для онлайн-словників, здебільшого, це передбачає досяжність для ботів кожної словникової статті (які, зазвичай, мають окремі веб-сторінки). Відтак, з міркувань індексовності, кожна сторінка сайту мусить мати свій унікальний ідентифікатор (URI). Існує декілька підходів для побудови таких ідентифікаторів.

Один з них – це використання унікального чисельного (чи символьно-чисельного) коду кожної сторінки. Інший – використання для ідентифікації включеної в URL назви статті. При цьому написання назви може бути оригінальним (в нашому випадку – кирилицею), або транслітерованим за певними правилами. Третій варіант – змішана система, у якій URL-адреса містить як унікальний ідентифікатор, так і назву статті, яка грає допоміжну роль та може бути видаленою без втрати дієвості посилання.

Кожен з цих підходів має свої переваги та недоліки, що їх перелічено у Таблиці 1. Читабельність, одночасна висока сумісність кодувань і позитивний вплив на

пошуковий рейтинг сайту від «інформативних» URL-адрес – саме ті причини, з яких транслітерація використовується для адресації кириличних веб-сайтів. У Таблиці 2

Тип адресування		Переваги	Недоліки
За кодом		• висока сумісність • лаконічність посилання	• неінформативність посилання
За назвою	кирилиця	• інформативність посилання • відсутність кодового ідентифікатора у адресі • висока читабельність	• низька сумісність • вірогідна нечитабельність та довжина («процентне кодування») • може потребувати уваги до унікальності назв
	трансліт	• висока сумісність • інформативність посилання • відсутність кодового ідентифікатора у адресі • задовільна читабельність	• може потребувати уваги до унікальності назв • можуть виникати складнощі через неоднозначність транслітерації
Змішаний	кирилиця	• стійкість до несумісності • інформативність посилання • стійкість до помилок у адресі • висока читабельність	• вірогідна нечитабельність та довжина («процентне кодування»)
	трансліт	• висока сумісність • інформативність посилання • стійкість до помилок у адресі • задовільна читабельність	

Таблиця 1. Порівняння схем адресування сторінок кириличних сайтів.

наведено приклади кожного типу URL-

адресування та статистику їх використання сьогодні серед 50 найпопулярніших українських сайтів новин. Як бачимо, найчастіше використовуються схеми з високим рівнем сумісності

та дієвості: а саме з унікальним кодом та/чи з транслітерацією. Схеми з вмістом назви

Тип адресування		Сайт	Приклад «відносної» частини URL	К-ть
За кодом		dt.ua	articles/86389	27
За назвою	кирилиця	uk.wikipedia.org	wiki/Ураган_Айрін	0
	трансліт	tsn.ua	nauka_it/rosiya-yevropeyskiy-lider-za-obsyagom-spamu.html	7
Змішаний	кирилиця	mukachevo.net	ua/News/view/45263-У-школі-вивчатимуть-психологію	0
	трансліт	gazeta.ua	articles/anna-usik/_yak-vijti-zamizh/396366	16

Таблиця 2. Приклади та поширеність п’яти типів адресування серед 50 найпопулярніших українських сайтів новин. Остання колонка – кількість сайтів, що використовують схему.

(транслітерованої чи ні), попри деякі складнощі реалізації, використовуються на рівні з адресацією кодом (23 проти 27 відповідно). Розробники найпопулярніших онлайн-словників, натомість, майже одностайно надали перевагу схемам із вмістом назви (Таблиця 3), хоча жоден з них не використовує схему з транслітерацією. Замість неї використовується нечитабельне для людини процентне кодування кирилиці та/чи власне

кирилиця (адже кириличні посилання в «довільний час» можуть бути для сумісності закодовані в процентне кодування однією з програм, до якої потрапило посилання). Виникнення процентного кодування у посиланні знижує читабельність, лаконічність та практичність посилання. При використанні некодованої кириці у адресі веб-сторінки – страждає сумісність, надто – для української мови. (Під сумісністю розуміємо міру сприйняття різноманітним програмним забезпеченням посилання як нерозривної сутності. Щодо української мови, на жаль, багато сайтів та програм не розпізнають українські літери як частину URL.) За таких умов доцільно використовувати транслітерацію кириличних назв в URL-адресах. Натомість, однією з причин невикористання переваг транслітерації у онлайн-словниках слугувала відсутність відомої цілком зворотної схеми транслітерації. Таку схему запропоновано в цій статті.

Сайт	Вигляд «відносної» частини URL (*path* + *query*)	Вміст слова	Google PageRank
dictionary.reference.com	browse/enthusiast	+	8
thefreedictionary.com	Enthusiast	+	8
dictionary.cambridge.org	dictionary/british/enthusiast?q=enthusiast	+	8
oxforddictionaries.com	definition/enthusiast	+	6
slovnyk.net	?swrd=%E5%ED%F2%F3%E7%B3%E0%F1%F2	+	6
slovnyk.org	fcgi-bin/dic.fcgi?hn=sel&iw=enthusiast&il=en-us&ol=uk-ua&ul=uk-ua	+	6
lcorp.ulif.org.ua/dictua	—	–	5
rozum.org.ua	index.php?a=term&d=18&t=14020	–	4
slovopedia.org.ua	36/53397/239635.html	–	4
slovari.yandex.ru	энтузиаст/правописание	+	7
multitran.ru	c/m.exe?l1=1&l2=2&s=%FD%ED%F2%F3%E7%E8%E0%F1%F2	+	6
gramota.ru/slovari	dic/?bts=x&word=%FD%ED%F2%F3%E7%E8%E0%F1%F2	+	6
lingvo.abbyyonline.com	ru/en-ru/энтузиаст	+	5
slovarik.kiev.ua	ojegov/ye/112839.html	–	2

Таблиця 3. Приклади будови URL-адреси сторінок в популярних онлайнових словниках – англійських, українських та російських. Друга колонка містить вигляд «відносної» частини URL для словникових статей «enthusiast», «ентузіаст» та «энтузиаст» для англійської, української та російської мов відповідно. Третя колонка — «Вміст слова» — показує чи розпізнається (або чи *може* бути розпізнана) назва словникової статті як частина URL. Четверта колонка – рейтинг сайту в пошуковій системі Google.

ЗВОРОТНА СХЕМА ТРАНСЛІТЕРАЦІЇ БЕЗ ДІАКТРИЧНИХ ЗНАКІВ

Існує близько десяти поширених та стандартизованих схем транслітерації українського тексту. Кожна з них відрізняється у таких аспектах як використання чи невикористання діактричних (наприклад, š, č, ê, ġ, i͡u) та неалфавітних знаків (апостроф, кавички, середня точка, гравіс тощо), призначення (транскрипція оригінального звучання, можливість відтворення оригінального тексту), природності (максимальної самозрозумілості), тяжіння до англійської (й = y, ц = ts) чи центральноєвропейської

фонології (й = j, ц = c), можливості транслітерації кількох мов одночасно та ін. Порівняння схем траслітерації української мови наведено у Таблиці 5.

На меті автора було створення цілком зворотної схеми транслітерації з максимальною самозрозумілістю для української мови, використанням лише символів не зарезервованих синтаксисом URI (26 символів латиниці, точка, тильда, дефіс та знак підкреслення [1]) та можливості одночасного кодування усіх символів східнослов'янських мов. «Повна зворотність» означає можливість відтворення будь-якого кириличного тексту, включаючи випадкового. Необхідність одночасного кодування усіх символів східнослов'янських мов походить із сучасної *de facto* двомовності України та історичної близькості з Росією та Білоруссю, адже багато класичної та сучасної української літератури містить вкраплення російського та білоруського тексту.

Символ	Позначення UURL		Інші стандарти, що використовують таке позначення
	технічний	загальний	
Апостроф	точка (.)	одинарні машинні лапки (')	Замість машинних лапок прийнято використовувати типографічний апостроф. У деяких схемах – подвійні лапки чи подібні до них символи.
Стоп-символ	тильда (~)	середня точка (·)	Середня точка – BGN/PCGN (1965)
Пробіл	знак підкреслення	*без змін* (пробіл)	Без змін – усі* стандарти
А, а	A, a		Усі*
Б, б	B, b		
В, в	V, v		
Г, г	Gh, gh		ТКПН; в деяких випадках – УКППТ, телеграмний, паспортний
Ґ, ґ	G, g		Усі*, крім ISO 9:1995 та ГОСТ 7.79-2000
Д, д	D, d		Усі*
Е, е	E, e		
Є, є	Je, je		ТКПН, Прусські інструкції, ISO/R 9:1968, телеграмний
Ж, ж	Zh, zh		Усі*, крім Прусських інструкцій, ISO/R 9:1968 та ISO 9:1995
З, з	Z, z		Усі*
И, и	Y, y		Усі*, крім ISO 9:1995 та ГОСТ 7.79-2000
І, і	I, i		Усі*, крім ISO 9:1995
Ї, ї	Ji, ji		ТКПН, Прусські інструкції, телеграмний
Й, й	J, j		Усі з європейською традицією**
К, к	K, k		Усі*
Л, л	L, l		
М, м	M, m		
Н, н	N, n		
О, о	O, o		
П, п	P, p		
Р, р	R, r		
С, с	S, s		
Т, т	T, t		

У, у	U, u		
Ф, ф	F, f		
Х, х	Kh, kh		ТКПН та усі з англійською традицією***
Ц, ц	C, c		Усі з європейською традицією** (ГОСТ 7.79-2000 Б – частково)
Ч, ч	Ch, ch		Усі*, крім Прусських інструкцій, ISO/R 9:1968 та ISO 9:1995
Ш, ш	Sh, sh		
Щ, щ	Shh, shh		ТКПН та ГОСТ 7.79-2000 Б
Ь, ь	J, j (після приголосних)		ТКПН (не беручи до уваги стоп-символ)
Ю, ю	Ju, ju		ТКПН, Прусські інструкції, ISO/R 9:1968, телеграмний
Я, я	Ja, ja		
Ё, ё	Joh, joh		—
Ў, ў	W, w		ТКПН (білоруський варіант)
Ъ, ъ	.H, .h	'H, 'h	—
Ы, ы	Yh, yh		—
Э, э	Eh, eh		ТКПН (російський варіант)

Таблиця 4. Схема транслітерації UURL (технічний та загальний варіанти).
* Під «усіма» стандартами мається увазі десять схем транслітерації із Таблиці 5.
** «Усі з європейською традицією» – шість схем транслітерації з Таблиці 5, які тяжіють до центральноєвропейської фонологічної традиції позначення символів.
*** «Усі з англійською традицією» – чотири схеми транслітерації з Таблиці 5, які тяжіють до англійської фонологічної традиції позначення символів.

Із існуючих стандартів транслітерації української мови близькими до описаних вимог є схеми BGN/PCGN (1965) [8], транслітераційна схема розроблена Термінологічною комісією з природничих наук (ТКПН) при Київському національному університеті ім. Тараса Шевченка [2] та російський стандарт транслітерації ГОСТ 7.79-2000 Б [4]. Жодна з цих схем, однак, не гарантує повну зворотність та не передбачає одночасну транслітерацію українських та російських букв.

Пропонована схема транслітерації *Ukrainian URL* (*UURL*), що її представлено у Таблиці 4, базується на схемі ТКПН. На відміну від транслітерації ТКПН та подібно до стандарту BGN/PCGN (1965), схема UURL розрізняє символи апострофу та службовий «стоп-символ» для уникнення неоднозначностей. Окрім того схема передбачає кодування пробілу. Апостроф, стоп-символ та пробіл, з міркувань сумісності з URL, запропоновано кодувати як точку, тильду та знак підкреслення; проте ці символи можуть бути змінені на інші в залежності від конкретних потреб. Відтак схему UURL можна використовувати у різних варіантах. Таблиця 4 наводить два з них – *технічний* та *загальний* (зручний для прочитання).

Схема UURL, так само як схема ТКПН, використовує літеру “h” суто для модифікації попередньої літери. За цим принципом призначені позначення для неукраїнських букв. Інша особливість успадкована від транслітерації ТКПН – позначення м’якого знаку (який в українській мові використовується лише після приголосних) символом “j”, який також позначає літеру «й». Такий вибір цілком обґрунтовується спільною природою літер «й» та «ь», адже палаталізацію (м’який знак)

можна розглядати як другорядну артикуляцію звуку [й], що відображено у Міжнародному фонетичному алфавіті (IPA): «й» = [j], «ь» = [ʲ]. Відтак “j” після приголосної позначає м’який знак, інакше – йот. Для вирішення неоднозначності, використовується стоп-символ: якщо він передує літері “j” – її функції міняються до навпаки.

Загалом, стоп-символ у схемі UURL використовується у всіх випадках «двозначності», наприклад:

- найактивніший → naj·aktyvnishyj
- серйозний → ser·joznyj
- лин(ь) → lyn(·j)

Символи, що не відносяться до транслітерованих ігноруються та залишаються без змін. Зокрема при транслітерації без змін залишається дефіс.

Додатковою перевагою UURL, успадкованою від схеми ТКПН, є можливість повного відтворення регістру літер. Із перелічених у Таблиці 5 десяти стандартів транслітерації, таку властивість має лише транслітерація ТКПН.

ПОРІВНЯННЯ З ІНШИМИ СХЕМАМИ ТРАНСЛІТЕРАЦІЇ

Як можна бачити із Таблиці 5, схема транслітерації UURL поєднує у собі переваги максимальної сумісності та зворотності (як у ГОСТ 7.79-2000 Б), одночасної підтримки кількох мов (як у ISO 9:1995) та простоти прочитання (як у BGN/PCGN (1965)). При цьому, як показано у Таблиці 6, UURL – єдина схема транслітерації,

Стандарт трансліту	Нелатиничні* символи	Фонологічна традиція	Використання / призначення	Підтримка мов
ALA-LC	діактрика, єднальний знак	англійська	бібліографічне	усі кириличні та ін.; окремі правила
Прусські інструкції	діактрика, апостроф	європейська	наукове / запис вимови	слов’янські кириличні; окремо
BGN/PCGN (1965)	середня точка, апостроф	англійська	універсальне / інтуїтив. вимова	усі кириличні; окремі правила
ISO/R 9:1968	діактрика	європейська	наукове / запис вимови	слов’янські кириличні; окремо
ISO 9:1995	діактрика	європейська	однозначний запис кирилиці	усі кириличні; єдині правила
ГОСТ 7.79-2000 Б	гравіс, апостроф	європейська	однозначний запис кирилиці	слов’янські кириличні; окремо
ТКПН	апостроф	європейська	універсальне	східнослов’янські; окремі правила
УКППТ 1996 (спрощений)	*жодних*	англійська	власні назви / інтуїтив. вимова	українська
Телеграмний (2005) [7]	апостроф	європейська	міжнародні телеграми	українська, російська; окремо
Паспортний	*жодних*	англійська	транслітерація	українська

(2010)			імен	
UURL (технічний варіант)	точка, тильда, знак підкреслення	європейська	технічне / універсальне	східнослов'янські; єдині правила

Таблиця 5. Порівняння стандартів транслітерації української мови.
* «Нелатиничні символи» – усі символи поза 26-буквеним латинським алфавітом.

Схема транслітерації	Роман «Місто»		Випадковий текст	
	Зворотність	К-ть символів	Зворотність	К-ть символів
UURL	100,000%	109,2%	100,00%	128,3%
ТКПН	99,985%	109,2%	73,25%	126,0%
BGN/PCGN (1965)	99,972%	109,1%	76,21%	128,1%
ГОСТ 9.97-2000 Б	100,000%	112,3%	99,46%	127,3%
ISO 9:1995	100,000%	100,0%	100,00%	102,3%
ALA-LC	99,427%	112,4%	79,54%	134,7%

Таблиця 6. Порівняння рівня зворотності та відносної кількості символів у транслітерованій формі (до кириличного оригіналу) для різних схем транслітерації. Міра «зворотності» – частка кириличних слів, яка точно відтворилася при автоматичному переведенні тексту українського роману та неосмисленого тексту в трансліт та назад у кирилицю. Випадковий неосмислений текст з понад 500 тис. символів, що був використаний для випробувань містив, символи української абетки, апострофи, коми та дефіси. Для конвертації використано сайт *translit.kh.ua.*

окрім діактричної ISO 9:1995, яка здатна цілком відтворити *будь-який* текст записаний літерами східнослов'янських мов. Це, зокрема, робить систему UURL стійкою до помилок у тексті. З Таблиці 6 також бачимо, що UURL є також достатньо компактним записом української мови (109,2% символів від початкової кількості при транслітеруванні роману «Місто» Валер'яна Підмогильного).

UURL не є абсолютно новою чи надто оригінальною схемою транслітерації. Таблиця 4 вказує, що більшість позначень символів у схемі використовуються кількома офіційними стандартами транслітерації. Для переважної більшості *українських* слів (99,978% серед слів роману «Місто»), форма транслітерації UURL тотожна формі транслітерації ТКПН. Це особливо корисно для використання схеми у всесвітній мережі, адже схема ТКПН є однією з офіційних транслітерацій для національного домену .UA [3]. Водночас схема UURL не надто далека від таких схем транслітерації англійської традиції як BGN/PCGN (1965), до яких тяжіють у своєму «розумінні» трансліту пошукові системи. Форма BGN/PCGN цілком співпадає з формою UURL для 58,15% слів роману «Місто».

РЕАЛІЗАЦІЇ

Схему транслітерації UURL реалізовано автором на мовах програмування Python, PHP та C#. Відповідні програмні бібліотеки поширюються безкоштовно (на умовах ліцензії *MIT Licence*) під назвою `cyr2url` на сервісі Google Code:

http://code.google.com/p/cyr2url/.

ВИСНОВОК

Запропонована схема транслітерації для східнослов'янських мов має низку переваг порівняно з існуючими офіційними стандартами транслітерації: вона є цілком зворотною, максимально сумісною (зокрема з URL-адресами), простою для прочитання та підтримує три мови водночас. Схему можна розглядати як значну модифікацію та надбудову транслітерації ТКПН. Серед відомих аналогів, схема є єдиним цілком зворотним транслітом без використання діактричних знаків. Може використовуватися принаймні у двох варіантах: технічному та загальному.

Завдяки повній зворотності та URL-сумісності, схема може бути застосована для формування URL-адрес сторінок кириличних онлайн-словників із вмістом транслітерованої форми шуканого слова. Таке рішення не потребує призначення стороннього унікального ідентифікатора та може позитивно вплинути на практичність, індексовність та рейтинг онлайн-словників у пошукових системах. Схема вже успішно застосовується у онлайн-словниках sum.in.ua та rymy.in.ua (приклад адреси: http://sum.in.ua/s/kavun)

Серед недоліків схеми UURL (як і всіх інших схем транслітерації) – неможливість цілком зворотно записувати текст, що містить як кириличні, так і латиничні символи. Однак схему можна використовувати як економне (7-бітне) кодування кириличного тексту. В подальшому схема також може бути розширенна на підтримку інших символів.

ДЖЕРЕЛА

[1] *RFC 3986 – Uniform Resource Identifier (URI): Generic Syntax*. – January 2005
http://tools.ietf.org/html/rfc3986

[2] Вакуленко М. *Derzhavna mova Ukrajiny v mizhnarodnomu spilkuvanni: ukrajinsjka ta skhidnoslov'jansjka latynyci*. – 1997.
http://translit.ndivision.net/nicy.dll?dread&article=2

[3] *Правила домену .UA* – 22.11.2004,
http://hostmaster.net.ua/policy/Policy_of_.UA.pdf

[4] *ГОСТ 7.79-2000: Правила транслитерации кирилловского письма латинским алфавитом* – Нормативна баса ГСНТИ, 2000,
http://gsnti-norms.ru/norms/common/doc.asp?0&/norms/stands/7_79.htm

[5] *Українська транслітерація – онлайн конвертор,*
http://translit.kh.ua/

[6] *Romanization of Ukrainian* – Wikipedia,
http://en.wikipedia.org/wiki/Romanization_of_Ukrainian

[7] *Постанова КМУ від 09.08.2005 № 720 «Про затвердження правил надання та отримання телекомунікаційних послуг»*
http://zakon.rada.gov.ua/cgi-bin/laws/main.cgi?page=1&nreg=720-2005-%EF

[8] *BGN/PCGN 1965 Romanization System for Ukrainian*,
http://earth-info.nga.mil/gns/html/Romanization/Romanization_Ukrainian.pdf