УДК 658.012.011.56

# ДОСЛІДЖЕННЯ ВІДПОВІДНОСТІ МІЖ ФОНЕТИЧНИМИ СИСТЕМАМИ ТА ГРАФІЧНИМ ПРЕДСТАВЛЕННЯМ ЛЕКСИКИ СХІДНОСЛОВ'ЯНСЬКИХ МОВ

**М. Ю. Кригін**

Український мовно-інформаційний фонд НАН України, м. Київ, Україна
maxus@zeos.net

Здійснено порівняння лексичного складу природної мови з його фонетичним відповідником, на основі чого зроблено висновок про ступінь відповідності між орфографічною та фонетичною системами мови. Наведено порівняльні характеристики ступеня фонетичності на рівні окремих літер, слів, а також за частинами мови. Дослідження проведено на прикладі української мови. Сформульовано певні припущення та пропозиції щодо проведення відповідних досліджень для інших східнослов'янських мов (білоруської та російської).

ЛЕКСИЧНИЙ СКЛАД, ФОНЕТИЧНА ТРАНСКРИПЦІЯ, СТУПІНЬ ФОНЕТИЧНОСТІ, ФОНЕТИЧНИЙ ВІДПОВІДНИК

## Вступ

Останніми роками в Україні досить жваво обговорюються правописні питання та принципи, на яких мусить ґрунтуватися український правопис. Зокрема, постає питання, наскільки чинний правопис відповідає фонетичному принципу, тобто наскільки фонетика української мови близька до її орфографії. Актуальним також є завдання побудови систем автоматичного озвучування українських текстів, представлених в електронній формі. Цьому присвячено чимало наукових праць [1–4]; створено й реально діючі системи озвучування [5]. Тим часом проблема якості систем озвучування з огляду на адекватне передання ними української літературної вимови та наголосу досі залишається нерозв'язаною. На нашу думку, для вирішення цієї проблеми необхідно проведення більш глибоких досліджень відповідності між знаковою системою українського письма та фонетичними процесами, притаманними українській мові. Дослідження такого роду можуть і мають проводитися на різних рівнях мови, насамперед, на лексичному, який репрезентує основну, центральну підсистему мовної системи. З огляду на значний обсяг української лексики (внаслідок розвиненості словозмінної системи української мови на реєстрі приблизно в 200 тис. одиниць він сягає 4 млн одиниць) такі дослідження неминуче набуватимуть статистичного характеру. Оскільки всі східнослов'янські мови засновані на кирилиці, мають схожі мовні системи і правила транскрибування, ми поширили дослідження на три мови: білоруську, російську та українську. Саме статистичному дослідженню відповідності між орфографією та фонетичним принципом письма для східнослов'янських мов і присвячена ця робота, проте найбільш детально ми зупинимося на аналізі української мови.

Метою роботи є дослідити на максимально формалізованому рівні, наскільки орфографічне представлення за правилами українського правопису слова $x$ відповідає його фонетичній транскрипції $x^t$ за правилами транскрибування [6] для української мови, коли $x$ пробігає деяку досить велику репрезентативну множину українських слів.

## 1. Принципи транскрибування текстів

Українська писемність, взагалі, заснована на засаді відповідності «звук – літера», але цей принцип витримується не абсолютно. Якщо знак фонетичної транскрипції відповідає написанню слова за правописними правилами, можемо робити висновок, що написання слова є фонетично адекватним. Якщо має місце лише часткова відповідність, то можна оцінити *ступінь фонетичності* написання, алгоритм розрахунку якого буде викладено далі.

Вважатимемо, що знаки письма відповідають знакам транскрипції, якщо вони позначаються на письмі однаково. З 33 букв українського алфавіту 27 мають відповідники у фонетичній транскрипції, і лише 6 не мають (**є**, **ї**, **ю**, **я**, **ь** та **щ**). До цих останніх слід додати і апостроф, який також не має свого відповідника у фонетичній транскрипції.

У багатьох випадках написання алфавітних літер збігається з транскрипційними знаками: **а, б, в, г, ґ, д, ж, з, к, л, м, н, п, р, с, т, у, ф, х, ц, ч, ш**. Якщо приголосні є пом'якшеними, за транскрипційним знаком, яким позначається літера, ставиться знак, яким позначається відповідний ступінь пом'якшення, тобто пом'якшений приголосний в транскрипції позначається принаймні двома знаками – знак, що відповідає звукові та символ пом'якшення, що йде за ним. Подовжені звуки позначаються двокрапками. Літера **в** у кінці складів при транскрибуванні передається як **ў**; літери **е, и, о** мають по два транскрипційних відповідники: **е** та $\mathbf{е}^{u}$, **и** та $\mathbf{и}^{e}$, **о** та $\mathbf{о}^{y}$ відповідно; знаком **й** у транскрипції позначається як літера **й**, так і перша частина йотованих голосних; літера **щ** транскрипційно позначається двома літерами **шч**; буквосполучення **дз** та **дж** мають власні транскрипційні знаки; м'який знак позначається за до-

помогою вертикальної рисочки: ′. У транскрипції також позначаються основні та другорядні наголоси. З правилами транскрибування слів української мови, які було взято за основу роботи програми транскрикування, можна ознайомитися в Орфоепічному словнику української мови [6].

### 2. Апарат порівняння орфографії та транскрипції

Введемо позначення. Символом $F(x)$ позначатимемо ступінь фонетичності слова $x$; через $L^C(x)$ позначимо кількість символів у слові $x$, написання яких в орфографічному представленні збігається з транскрипційним представленням; через $L^W(x)$ позначимо кількість символів у слові $x$ в орфографічному представленні; через $L^T(x)$ позначимо кількість символів у слові $x$ в транскрипційному представленні. Тоді ступінь фонетичності $F(x)$ слова $x$ природно визначити такою формулою:

$$F(x) = \frac{2L^C(x)}{L^W(x) + L^T(x)} \cdot 100\% . \qquad (1)$$

Ступінь фонетичності $F$ для всього масиву тексту вираховується як середнє $F(x)$:

$$F = \frac{\sum_{x \in G} F(x)}{N[G]} , \qquad (2)$$

де $N[G]$ – кількість слів у досліджуваному масиві $G$.

Дослідження може бути проведене на масиві текстів (наприклад, на певному лінгвістичному корпусі) або на достатньо широкому списку словоформ. У першому випадку є можливість врахування частотності вживання тих чи інших слів і відповідно комбінацій тих чи інших літер, але при цьому деякі графемні комбінації можуть взагалі не потрапити до вибірки. У другому випадку частоти вживання слів не враховуються, але подається повна картина словоформ та морфемних комбінацій у реальних та граматично коректних словоформах. З огляду на викладене саме другий шлях було обрано для нашого дослідження. Предметом вивчення став масив українських словоформ, який для змінюваних лексем виводиться за допомогою програми парадигматизації (побудови повної словозмінної парадигми) на основі словозмінної класифікації, розробленої в Українському мовно-інформаційному фонді НАН України [7].

Обраний у якості дослідного масиву перелік словоформ літературної української мови було сформовано автоматично програмою граматичного словника, створеного в Українському мовно-інформаційному фонді НАН України, лексеми в якому проіндексовано за словозмінними класами української мови згідно із словозмінною класифікацією [7]. При цьому враховувалися зміни наголосів у похідних словоформах відповідно до акцентуаційної класифікації української мови [8]. Наголоси в початкових словоформах було розставлено відповідно до Орфографічного словника української мови [9]. Загалом досліджуваний масив налічує понад 186 тис. лексем та близько 3,7 млн словоформ. Кожну з словоформ було автоматично протранскрибовано за допомогою спеціальної програми, розробленої відповідно до правил української фонетичної транскрипції, поданої в Орфоепічному словнику; зазначена програма використовується також в українському електронному словникові [10]. Таким чином, було отримано пари «словоформа – її фонетична транскрипція» у такому вигляді (див. табл. 1).

**Таблиця 1**

Транскрибування словоформ

| | |
|---|---|
| мова | [мува] |
| мови | [муви$^{e}$] |
| мові | [мув‘і] |
| мову | [муву] |
| мовою | [мувойу] |

Тотожність знаків письма і транскрипції визначалася за повною відповідністю або за відповідністю знаків з наголосами звичайним ненаголошеним буквам українського письма. Так, у першому рядку знак **ó** з правої колонки вважається тотожним з буквою **о** лівої колонки. Таке спрощення прийнято через відсутність знаків наголосу в українських текстах (крім деяких спеціальних: книжки для молодших школярів, навчальна література для іноземців тощо).

### 3. Результати дослідження

Для кожної пари «слово – транскрипція» отримано числову характеристику, яка дорівнює кількості співпадаючих символів. Це проілюстровано в табл. 2.

**Таблиця 2**

Приклад аналізу збігу символів у словоформі

| Слово $x$ | Транскрипція | Кількість символів у слові, $L^W(x)$ | Кількість символів у транскрипції, $L^T(x)$ | Кількість співпадаючих символів, $L^C(x)$ |
|---|---|---|---|---|
| тигр | [тигр] | 4 | 4 | 4 |
| хліб | [хл’іб] | 4 | 5 | 4 |
| будь-хто | [буд’хто́] | 8 | 7 | 6 |
| починається | [почи$^{e}$на́йе$^{i}$ц‘:а] | 11 | 12 | 5 |
| щит | [шчит] | 3 | 4 | 2 |

Таким чином, ступінь фонетичності словоформи становить для першого рядка табл. 2

$$F(x) = \frac{2*4}{4+4} * 100 = 100\% ,$$

для другого $- F(x) = \frac{2*4}{4+5} * 100 = 88{,}89\% ,$

для третього $- F(x) = \frac{2*6}{8+7} * 100 = 80\% .$

Наголошене **é** в транскрипції ототожнюється зі звичайним **е**.

Результати порівняння для табл. 1 подано в табл. 3.

**Таблиця 3**

Приклади обчислення ступенів фонетичності

| № | Словоформа $x$ | Транскрипція | $L^C(x)$ | $L^W(x)$ | $L^T(x)$ | $F(x)$, (%) |
|---|---|---|---|---|---|---|
| 1 | мова | мо́ва | 4 | 4 | 4 | 100 |
| 2 | мови | мо́ви$^{e}$ | 3 | 4 | 4 | 75 |
| 3 | мові | мо́в‘і | 4 | 4 | 5 | 88,88889 |
| 4 | мову | мо́ву | 5 | 5 | 5 | 100 |
| 5 | мовою | мо́войу | 4 | 5 | 6 | 72,72727 |

Ступінь фонетичності окремого символу алфавіту підраховано як відношення кількості вживань певної літери, які у транскрипції передаються її орфографічним відповідником до кількості всіх вживань цієї літери. Ступені фонетичності українських літер представлено у табл. 4.

**Таблиця 4**

Ступені фонетичності символів у масиві українських словоформ

| Символ | Кількість у масиві | Кількість відповідностей у масиві транскрипцій | Ступінь фонетичності |
|---|---|---|---|
| ' | 48015 | 0 | 0 |
| - | 68634 | 0 | 0 |
| а | 2760049 | 2757090 | 99,89 |
| б | 514549 | 508305 | 98,79 |
| в | 1913692 | 1460382 | 76,31 |
| г | 567803 | 563357 | 99,22 |
| ґ | 7469 | 7412 | 99,24 |
| д | 856574 | 734626 | 85,76 |
| е | 1744011 | 356509 | 20,44 |
| є | 199814 | 0 | 0 |
| ж | 217677 | 164120 | 75,4 |
| з | 712748 | 663714 | 93,12 |
| и | 2245339 | 557877 | 24,85 |
| і | 1737610 | 1735282 | 99,87 |
| ї | 105560 | 0 | 0 |
| й | 458373 | 458196 | 99,96 |
| к | 1239151 | 1230113 | 99,27 |
| л | 1382647 | 1374111 | 99,38 |
| м | 1584210 | 1577046 | 99,55 |
| н | 2401237 | 2254700 | 93,9 |
| о | 3500959 | 3257258 | 93,04 |
| п | 1181773 | 1174116 | 99,35 |
| р | 1875913 | 1871017 | 99,74 |
| с | 1507714 | 1428001 | 94,71 |
| т | 1710612 | 1629247 | 95,24 |
| у | 1441999 | 1426272 | 98,91 |
| ф | 151271 | 148976 | 98,48 |
| х | 388302 | 387357 | 99,76 |
| ц | 227980 | 226601 | 99,4 |
| ч | 459448 | 452804 | 98,55 |
| ш | 385905 | 369918 | 95,86 |
| щ | 81401 | 0 | 0 |
| ь | 686208 | 0 | 0 |
| ю | 426560 | 0 | 0 |
| я | 628415 | 0 | 0 |

В результаті проведення підрахунків ступеня фонетичності за формулами (1) та (2) на всьому сформованому масиві словоформ було встановлено, що українські тексти – на рівні слів – на 76,7% збігаються з власною транскрипцією.

Нерівномірність розподілу літер за їх відповідністю у транскрипції пояснюється особливістю української правописної системи. Шість вищевказаних літер (**є, ї, щ, ь, ю, я**), а також апостроф і дефіс не мають графічних відповідників у фонетичній транскрипції, хоча знак м’якшення фактично однозначно відповідає знаку вертикальної риски у транскрипції. Голосні **є, ю, я** транскрибуються у залежності від місця в слові та передаються, наприклад, через **й** та відповідний нейотований голосний або через відповідний нейотований приголосний з пом’якшенням попереднього голосного. Зрозуміло, що всі ці символи мають нулі у відповідних колонках табл. 4.

Цікавішим є факт досить низьких рівнів фонетичності для літер **в** (76,31%), **д** (85,74%) , **ж** (75,4%) і особливо **е** (20,44%) та **и** (24,85%). Це пояснюється регулярним вживанням цих літер на позначення різних звуків. Зокрема, буква **в** у нашій мові вживається для вираження щілинного приголосного **в** та напівголосного **ў**. Велика частотність цього останнього звуку, зокрема в дієслівних формах чоловічого роду минулого часу, зумовлює високий відсоток невідповідності букви **в** її фонетичному репрезентантові у масиві українських словоформ. Низький ступінь фонетичності літер **д** та **ж** зумовлений вживанням сполучення цих букв для позначення африкати **дж**. Нарешті, напрочуд низький ступінь фонетичності для високочастотних голосних **е** та **и** пояснюється регулярним вживанням транскрипційних символів **е**$^{и}$ та **и**$^{е}$, відповідно, на позначення специфічних голосних звуків у ненаголошеній позиції.

Було також досліджено залежність фонетичності графічного репрезентанта словоформи від її частиномовної приналежності. Результати подано в табл. 5.

**Таблиця 5**

Ступінь фонетичності словоформ різних частин мови

| Частина мови | Загальна кількість словоформ | Збіг з транскрипцією $F_x$ |
|---|---|---|
| Іменники | 922707 | 79,323350060 |
| Прикметники | 959898 | 80,516120718 |
| Дієслова | 1015115 | 73,614291915 |
| Прислівники | 8712 | 81,321647214 |
| Дієприслівники | 186 | 72,911394735 |
| Дієприкметники | 330338 | 81,576514615 |
| Числівники | 1738 | 72,09770427 |
| Прийменники | 149 | 81,988834356 |
| Частки | 124 | 84,685310025 |
| Сполучники | 118 | 82,307622942 |
| Займенники | 2945 | 74,090968710 |
| Вигуки | 492 | 78,744486483 |
| Присудкове слово | 173 | 82,837217673 |
| Вставне слово | 31 | 81,956417207 |

Високий відсоток невідповідності з транскрипцією для дієслів та числівників неважко пояснити регулярною появою у словозмінній парадигмі форм на -ться (для дієслів) та -дцять (для числівників), для яких написання лише віддалено відповідає вимові і відповідно графічній транскрипції: [ц:а] та [ц':ат'] відповідно.

Аналогічно можна провести аналіз відповідності транскрипції початкових форм слів та найбільш частотних форм, які використовуються в мові, а також відібраних за іншими критеріями груп словоформ, наприклад, термінів – в залежності від кінцевої мети дослідження.

Наявність переліку найуживаніших слів (зокрема, зафіксованих у найбільш авторитетних словниках), а також наявність системи парадигматизації української лексики, яка дозволяє отримати в явному вигляді всі непочаткові форми слів, робить цілком можливим кількісний аналіз відповідності написання українських слів їх звуковому вираженню.

**Висновки**

З огляду на викладене можна сказати, що ступінь фонетичності української мови досить високий, і отже, цілком реальна можливість побудови високоякісних систем для озвучування українських текстів, заснованих на простих правилах. Для цього необхідно провести статистичні дослідження розподілу транскрипційних символів (та їхніх комбінацій) в українських текстах і за результатами проведеного дослідження сформувати експериментальні масиви акустичних відповідників із урахуванням інтонаційних та синтагматичних характеристик.

Серед східнослов'янських мов найбільш фонетичною, на нашу думку, є білоруська, оскільки в ній є окремий символ в алфавіті для позначення нескладового **у – ў**, немає ефекту переходу **о – а**, **е – и** при вимові та передачі дзвінких приголосних глухими в усному мовленні. В російській мові, як ми прогнозуємо, спостерігається найменший серед східнослов'янських мов ступінь фонетичності за рахунок того, що в словах типу ***солнце, радостный*** відбувається випадіння приголосного, в словах типу ***дуб, пруд*** кінцевий дзвінкий вимовляється глухо, голосні **о** переходять на вимові в **а** у словах типу ***пошел, родник;*** спостерігаються також інші системні фонетично-правописно розходження. Більш докладному аналізу цих явищ буде присвячено окрему роботу.

Автор висловлює подяку В. А. Широкову та І. В. Шевченку, у тісній співпраці з якими була написана ця стаття.

**Список літератури: 1.** *Сажок М. М.* Автоматизовані засоби дослідження синтезу українського мовлення на основі фонемно-трифонної моделі // Автоматизовані системи управління та прогресивні інформаційні технології. Вип. І.– Київ, 2003.–С. 101–113. **2.** *Сажок М.М.* Усномовний паспорт диктора для мовленєвих діалогових систем // Автоматизовані системи управління та прогресивні інформаційні технології. Вип. ІІ.–Київ, 2004.–С. 101–111. **3.** *Сажок М.* Генерування правил розставляння наголосів у багатомовному аспекті // Праці 5-ї Всеукр. міжнар. конф. «Оброблення сигналів і зображень та розпізнавання образів» УкрОбраз'2000.– Київ, 2000.–С. 111–112. **4.** *Т. В. Людовик, Н. Н. Сажок.* Использование речевых баз данных большого объема при синтезе речи в системах искусственного интеллекта // Проблемы управления и информатики.– Київ.– 2003.– №6.– С. 82–87. **5.** *Вінцюк Т., Людовик Т., Сажок М., Селюх Р.* Автоматичний озвучувач українських текстів на основі фонемно-трифонної моделі з використанням природного мовного сигналу // Праці 6-ї Всеукр. міжнар. конф. «Оброблення сигналів і зображень та розпізнавання образів» УкрОбраз'2002.– Київ, 2002.–С. 79–84. **6.** Орфоепічний словник української мови: в 2 т. / За ред. М. М.Пещак, В. М. Русанівського.–К.: Довіра, 2001. **7.** *Шевченко І. В.* Алгоритмічна словозмінна класифікація української лексики // Мовознавство.–1996.–№ 4–5.–С. 40–44. **8.** *Шевченко І. В.* Автоматизована дистрибуція наголосів у словозмінній парадигмі українського дієслова. // Мовознавство.–2001.–№5.–С. 26-30. **9.** Український орфографічний словник. Вид. 4-те / За ред. В. М. Русанівського.– К.: Довіра, 2005. **10.** *Широков В. А., Рабулець О. Г., Шевченко І. В., Костишин О. М., Якименко К.М.* Інтегрована лексикографічна система «Словники України».– К., 2004.

***Поступила до редколегії 16.02.2009***

УДК 658.012.011.56

**Исследование соответствия между фонетическими системами и графическим представлением лексики восточнославянских языков** / М. Ю. Крыгин // Бионика интеллекта: науч.-техн. журнал. –2009.–№1 (70).–С. 60–63.

Статья посвящена исследованию того, насколько близки фонетика и письменность восточнославянских языков. Построена модель степени соответствия между орфографическим и фонетическим представлением слов. Вычислительный эксперимент проводился на выборке, состоящей из словоформ, зафиксированных в грамматическом словаре украинского языка (около 3,7 млн. слов). Сделан вывод, что степень фонетичности украинской орфографии составляет более 76%. Обсуждаются аналогичные вопросы для белорусского и русского языков.

Табл.: 5. Библиогр.: 7 назв.

UDK 658.012.011.56

**Research of the correspondence between phonetic systems and graphic representation of vocabulary for East Slavonic languages** / M.Ju. Krygin // Bionics of Intelligence: Sci. Mag. –2009. – №1(70).– P. 60–63.

The article is devoted research of the similarity of phonetics and written language for the East Slavonic languages. The model of correspondence degree between orthographic and phonetic representations of the words is built.

The computing experiment has been conducted on a sample that contains word forms fixed in the Ukrainian Grammatical Dictionary (about 3.7 million word forms). A conclusion is made that the degree of the correspondence of Ukrainian orthography and phonetics is more than 76%. Analogical problems for Byelorussian and Russian are discussed.

Tabl.: 5. Ref.: 7 items.