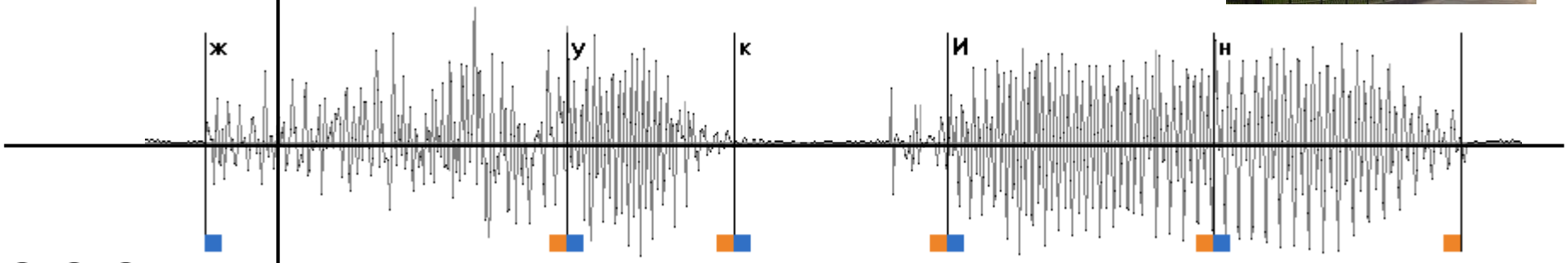
Микола Сажок, Валентина Робейко



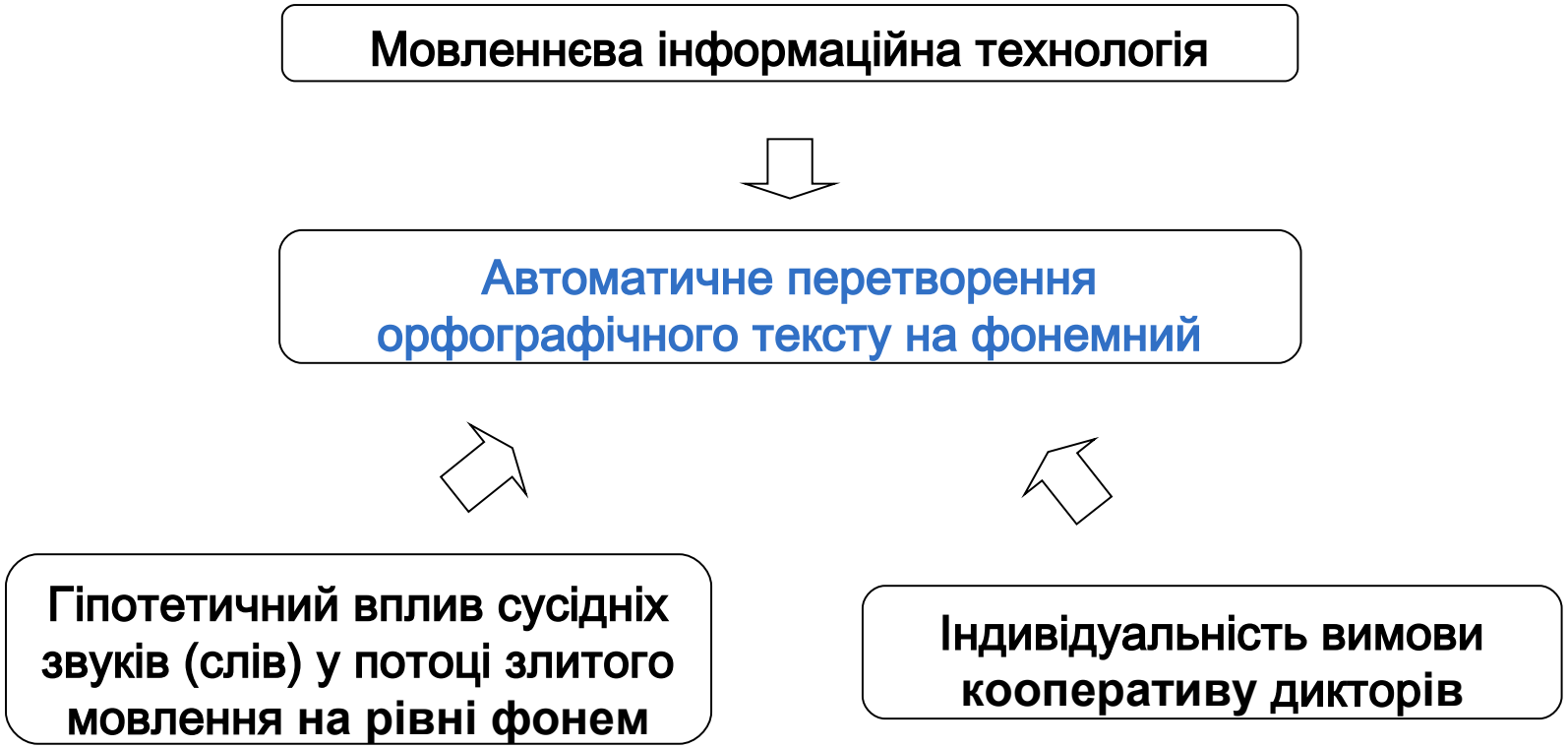
Проблеми перетворення орфографічного тексту на фонетичний для розроблення корпусів

Жукин’2010

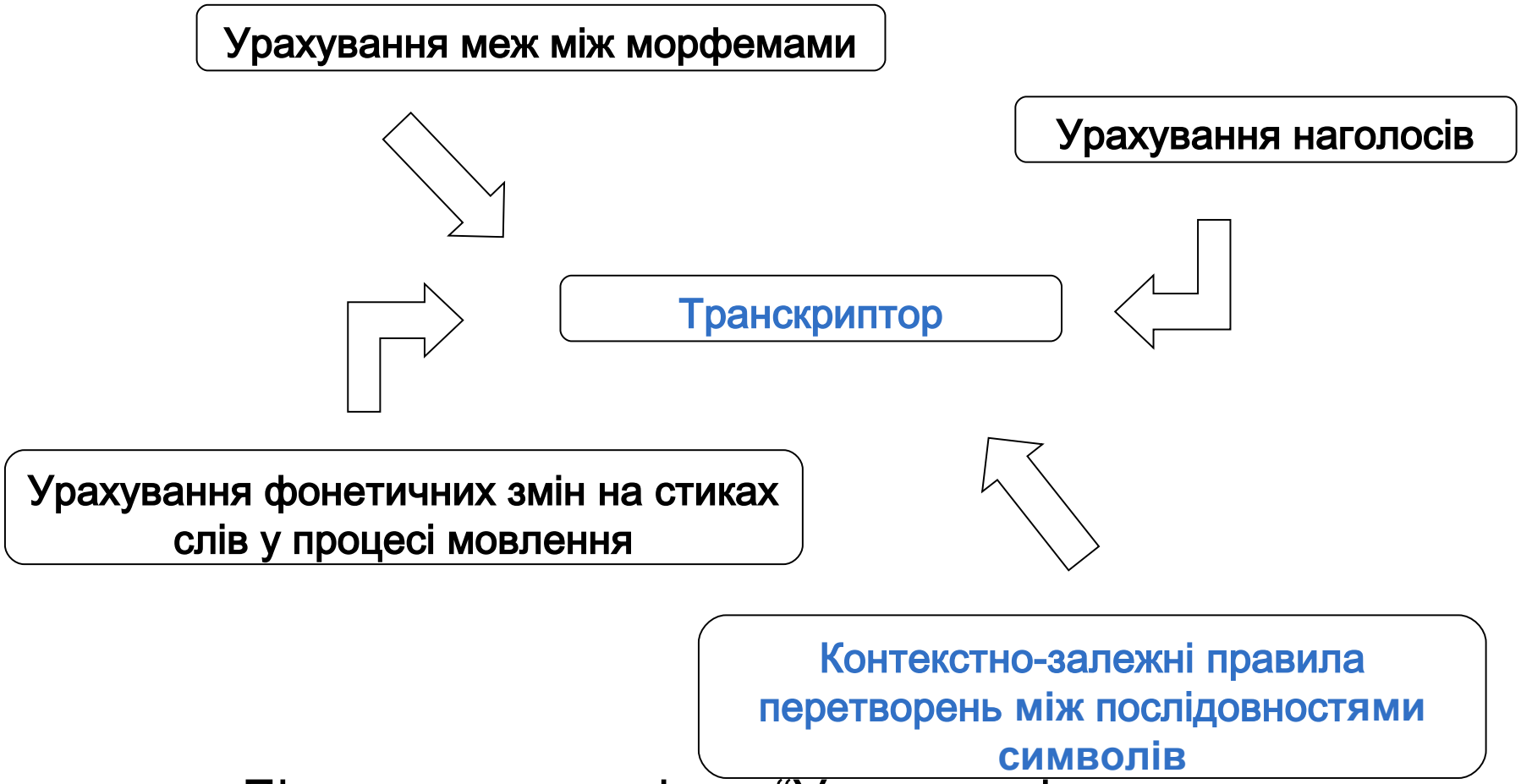
Зміст

- Загальна структура перетворювача графем на фонеми
- Перетворювач
- Розбиття на синтагми
  - Неоднозначність функції токена
- Від графеми до фонеми
  - Гомографія
- Експериментальні результати
- Висновки

Загальна структура перетворювача графем на фонеми



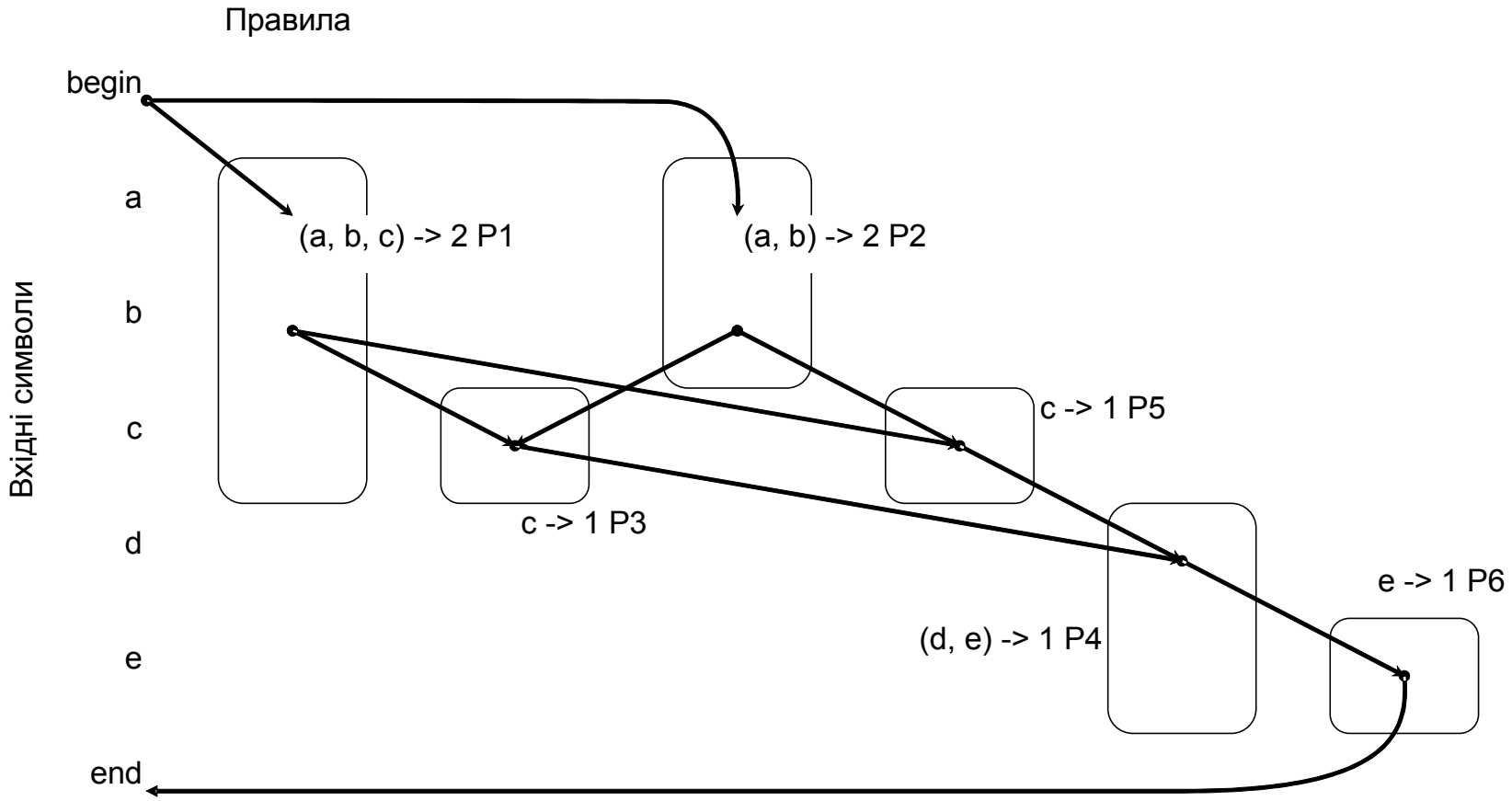
Загальна структура перетворювача графем на фонеми



Стандартні підходи

1. Таблиця замін
   - величезна кількість елементів
   - проблемно реалізувати багатозначність
1. Програмні модулі
   - потрібна перекомпіляція для внесення змін
   - нерегуляризованість

Мультиоператор



Мультиоператор

Параметри:

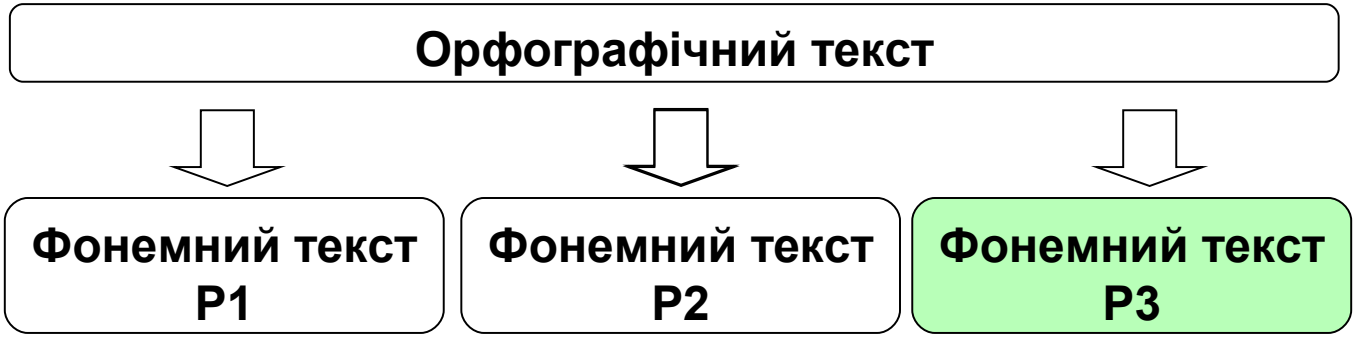
1. Найдовша довжина вхідної підпослідовності
2. Імовірності (взаємного) застосування кожного з елементарних операторів

Мультиоператор

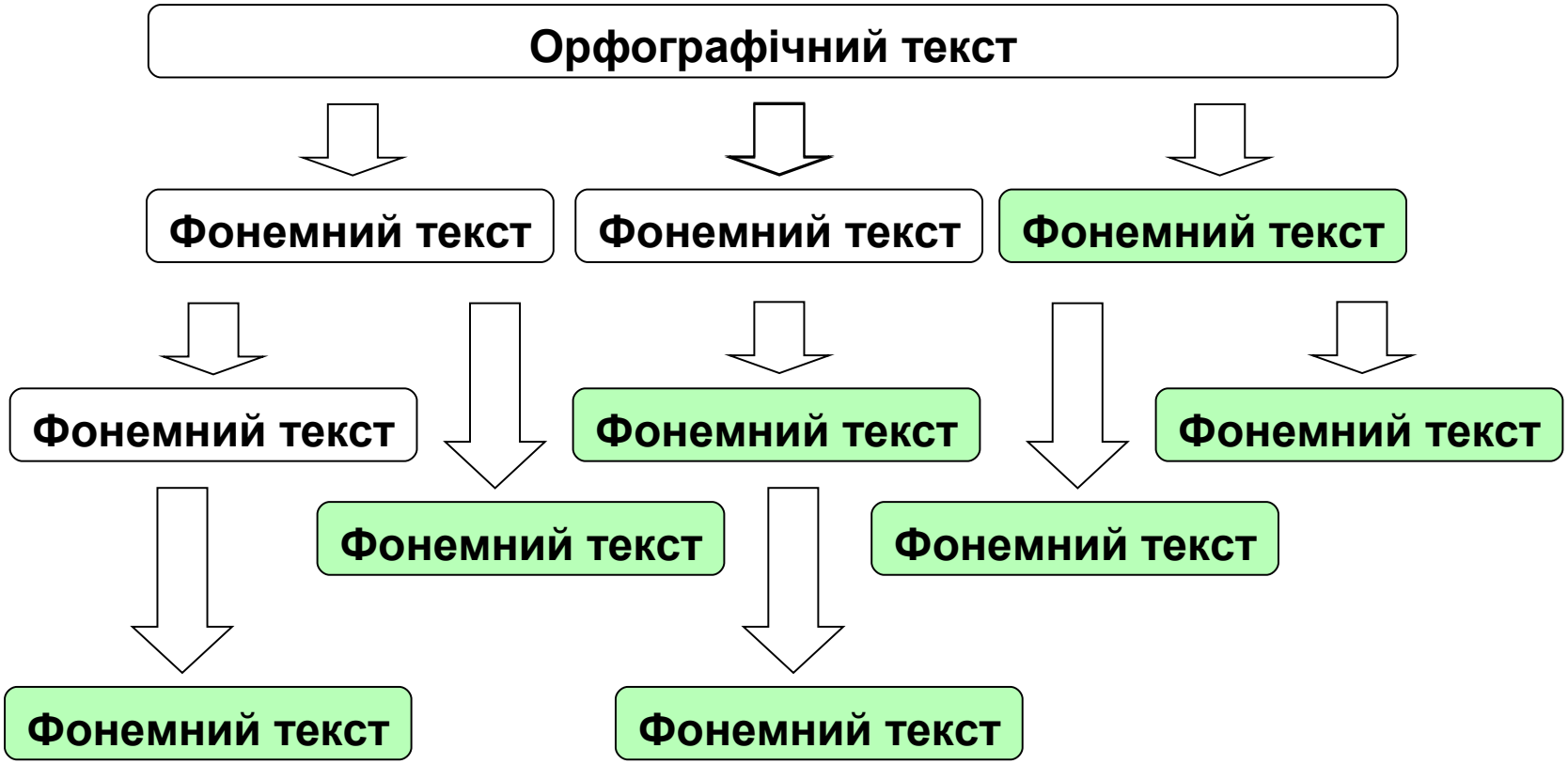
Особливості реалізації:

1. Відома вхідна послідовність повністю
2. Послідовність надходить по відліку
3. Затримка на найдовшу довжину вхідної підпослідовності

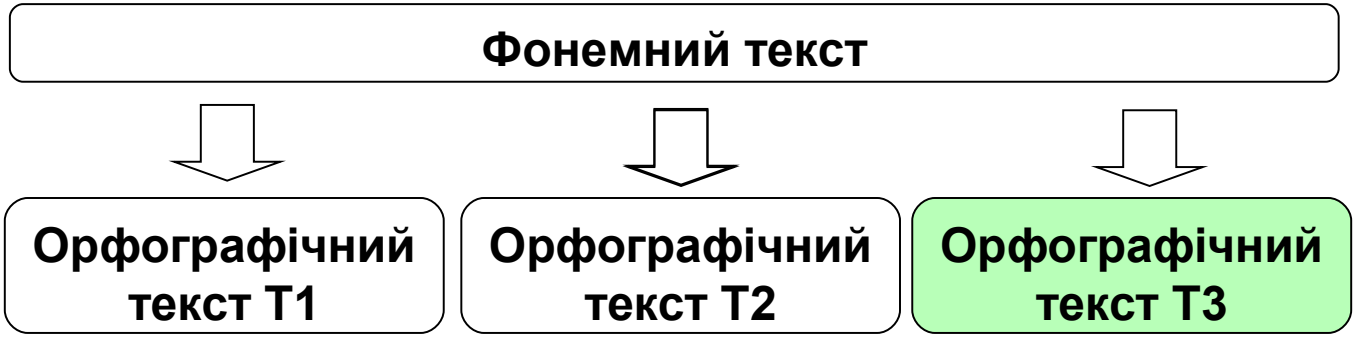
Варіанти транскрибування



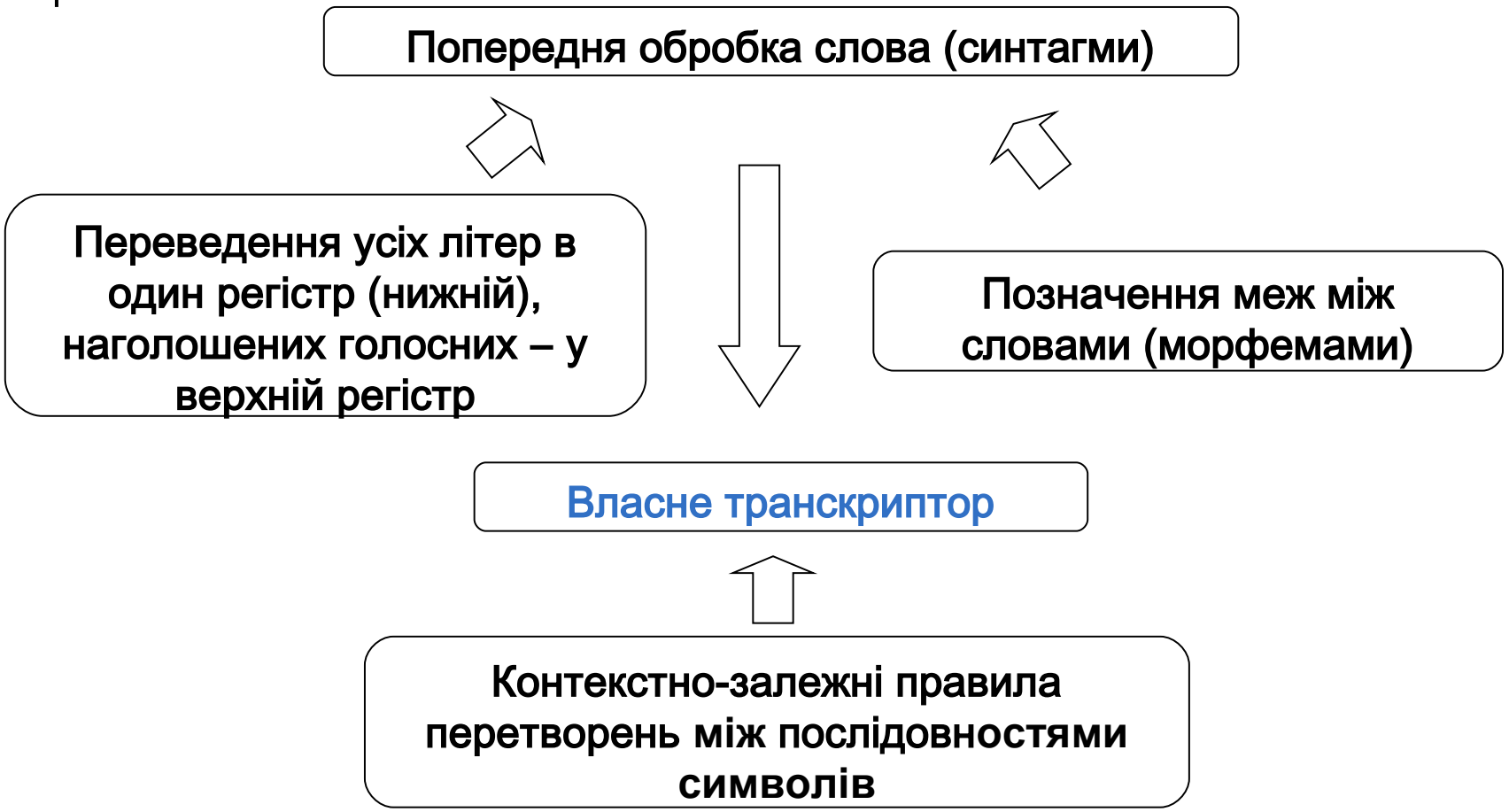
Варіанти транскрибування – каскад



Варіанти транскрибування: зворотнє



Транскриптор. Принцип роботи

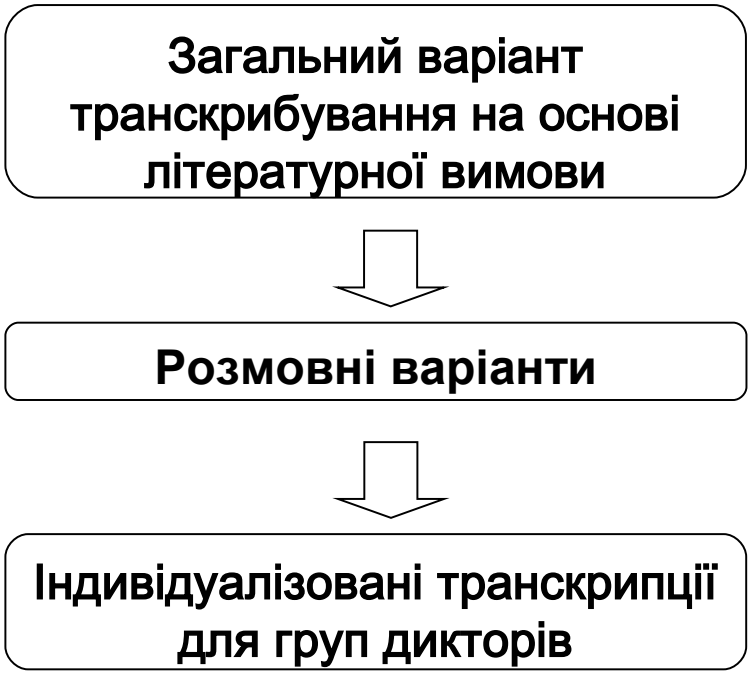


Приклади правил побудови транскриптора

Вхідна послідовність **графем**	Вихідна послідовність **фонем**	Ширина кроку аналізу	Пояснення
[зсц] [жшч]	[жшч]	1	**з**, **с**, **ц** перед **ж**, **ш**, **ч** переходять відповідно у **ж**, **ш**, **ч**
т[дтзснц] [іІєюяЄЮЯь]	т'	1	**т** перед м'якими **д**, **т**, **з**, **с**, **н**, **ц** пом'якшується
с т [лн]	с	2	**т** між **с** та **л** або **н** випадає

Близько 30 подібних правил

Варіанти транскрибування



Позиційні та комбінаторні зміни звуків у потоці мовлення

Позиційні зміни звуків у потоці мовлення

Комбінаторні зміни звуків у потоці мовлення

боротьба → б а р а д' б А
або б р а д' б А
безпекою → б е с п Е к о у
доброю → д О б р о й

книжка → к н И ш к а
квітка → к' в' І т к а
аеропорт → а р о п О р т
чесний → ч Е с т н и й

Позиційні зміни звуків у потоці мовлення

- окрім редукції ненаголошених **е**, **и** та **о** до **е**^{и}, **и**^{е}, **о**^{у}, також ослаблена вимова **о** як **а** в ненаголошеній позиції, рідше трапляється редукція ненаголошених голосних до повного зникнення (**тепер → т и п Е р**, **зозуля → з у з У л' а**, **боротьба → б а р а д' б А** або **б р а д' б А**);

- оглушення дзвінких приголосних перед паузою (**брід → б р' І т**, **зараз → з А р а с**);

- редукція у термінальних частинах слів у процесі мовлення – зникнення приголосного звука в закінченнях **-ого**, **-их**, **-ий**, **-іх**, **-ій**, **-ії**, **-ої**, **-еї**, **ою**, **-єю**, **-ити** та подібних (**коротший → к о р О ч ш и**, **синіх → с И н' і**, **безпекою → б е с п Е к о у**); зникнення кінцевого голосного звука в закінченнях **-ою**, **-ею**, **-єю** та подібних (**доброю → д О б р о й**, **землею → з е м л Е й**) та ін.

Комбінаторні зміни звуків у потоці мовлення

- повна регресивна асиміляція за глухістю у сполуці „дзвінкий+глухий” на межі будь-яких морфем у слові та на межі слів (**без причини → б е с п р и ч И н и**, **розсунути → р о с с У н у т и**, **книжка → к н И ш к а**, **сядьте → с' А т' т е**);

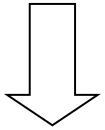
- асиміляція за м’якістю свистячих та шиплячих приголосних, губних та задньоязикових приголосних (**злі → з' л' І**, **шлях → ш' л' А х**, **квітка → к' в' І т к а**);

- вимова подовжених приголосних звуків як звичайного неподовженого звука, вимова двох голосних як одного звука (**віддати → в' і д А т и**, **знання → з н а н' А**, **зоопарк → з о п А р к**, **аеропорт → а р о п О р т**);

- неповне спрощення в групах приголосних, його відсутність (**чесний → ч Е с т н и й**) та ін.

Приклади індивідуалізованих словників

Різновиди транскрипцій словника



Інд.словник Літ. транскрипція	devocal	end_cons	a	devocal _ a
р о б И в	р о б И ф	р о б И	р а б И в	р а б И ф
в О р о г	в О р о х	в О р о	в О р а г	в О р а х

Інші застосування алгоритму

Перетворення чисел, символів і скорочень на орфографічний текст (приклад правил з фару)

Основна ідея лінгвістичної моделі (нагадування)

Імовірність речення із l слів $W_{0,l} = (w_1, \quad w_2, \ldots w_l)$ розкладається на добуток умовних імовірностей:

$$P(w_1, w_2, \ldots, w_l) = \prod_{i=1}^{l} P(w_i / w_1, w_2, \ldots, w_{i-1}) = \prod_{i=1}^{l} P(w_i / W_{0,i-1}),$$

Основна ідея лінгвістичної моделі

Обмежуємо контекст:

$$P(w_1, w_2, \ldots, w_l) \cong \prod_{i=1}^{l} P(w_i / w_{i-n+1}, \ldots, w_{i-1}) = \prod_{i=1}^{l} P(w_i / W_{i-n,i-1}),$$

де n – порядок моделі.

При $n = 1$ модель вироджується до відсутності елементарних контекстів, тобто розглядаються **уні-грами**.

При $n = 2$ отримуємо **бі-грамну** модель, коли в елементарний контекст входить граматична функція одного попереднього слова.

Збільшуючи далі параметр n, ми розширюємо елементарний контекст, підвищуючи точність моделі.

Основна ідея лінгвістичної моделі

Враховуючи:

$$\hat{P}\left(W_{0,l}\right) = \prod_{i=1}^{l} P\left(w_i / W_{i-n,i-1}\right),$$

Оцінюємо ймовірність послідовності слів при надходженні наступного слова w_{i+1}

$$\hat{P}\left(W_{0,i+1}\right) = \hat{P}\left(W_{0,i}\right) P\left(w_{i+1} / W_{i-n,i}\right)$$

Основна ідея лінгвістичної моделі

Оцінка параметрів лінгвістичної моделі

$$P\left(w_i / W_{i-n,i-1}\right) = \begin{cases} \alpha\left(W_{i-n,i-1}\right) & : \quad c\left(W_{i-n,i}\right) = 0 \\ d_{c\left(W_{i-n,i}\right)} \dfrac{c\left(W_{i-n,i}\right)}{c\left(W_{i-n,i-1}\right)} & : \quad 1 \le c\left(W_{i-n,i}\right) \le k \\ \dfrac{c\left(W_{i-n,i}\right)}{c\left(W_{i-n,i-1}\right)} & : \quad c\left(W_{i-n,i}\right) > k \end{cases}$$

Основна ідея лінгвістичної моделі

Розбиття слів на класи $g = G(w)$

Узгодження розшифрованих елементів тексту

Класи, яким належить слово w:

$$G(w) = \{g^1(w), g^2(w), \ldots, g_n(w)\}$$

Задаються на основі граматичних типів (частина мови, число, відмінок, рід, час, особа …)

Узгодження розшифрованих елементів тексту

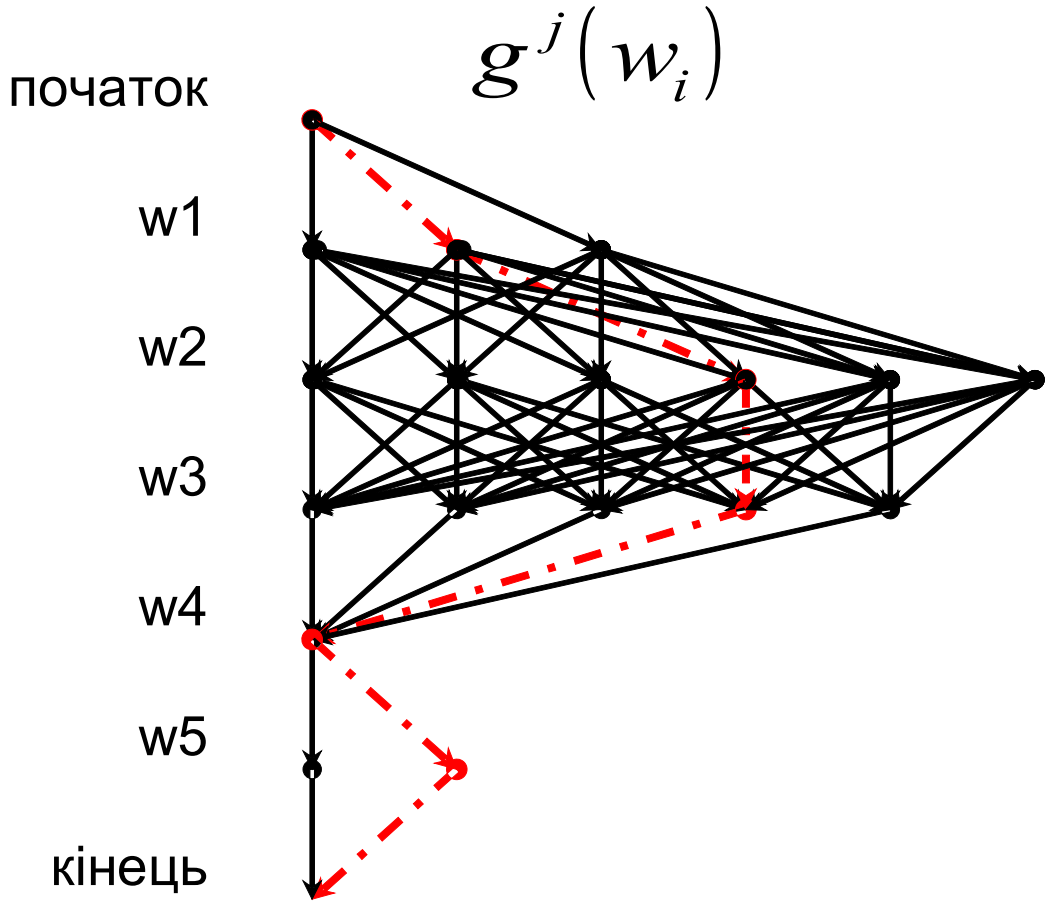
Зафіксуємо по класу з множини класів $G(w)$ для кожного слова w:

$$G_{0,l}^{j} = \left(g^{j}(w_1), g^{j}(w_2), ..., g^{j}(w_l)\right).$$

Оцінюємо ймовірність послідовності класів слів при надходженні наступного слова w_{i+1} за відомої $\hat{P}\left(G_{0,i}^{k}\right)$ для j-го класу слова w_{i+1}:

$$\hat{P}\left(G_{0,i}^{j}\right) = \hat{P}\left(G_{0,i}^{k}\right) P\left(g^{j}(w_{i+1}) / G_{i-n,i}^{k}\right)$$

Узгодження розшифрованих елементів тексту



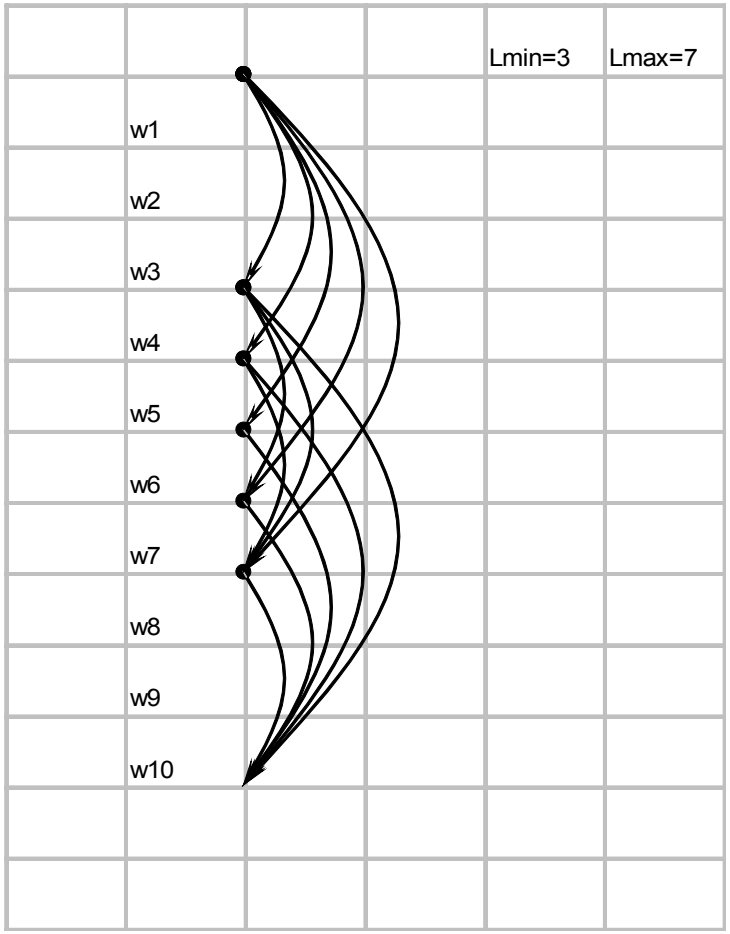
Поділ на синтагми (інтонаційні групи)

IG(w(i)) повертає номер синтагми слова w(i) у реченні W

$$P(IG(w(i)) = j \,/\, IG(w(i-1)) = j-1)$$

Lmin, Lmax – найкоротша і найдовша тривалість синтагми

Поділ на синтагми (інтонаційні групи)



Поділ на синтагми

g(i) – граматична форма слова w(i)

Тоді

P(IG(w(i)) = j / IG(w(i-1)) = j-1)

заміняємо на

P(IG(g(i)) != IG(g(i-1)))

Спасибі за увагу!

robeiko@uasoiro.org.ua

mykola@uasoiro.org.ua