Створення та анотування акустичних (мовленнєвих) корпусів

Валентина Робейко

Мовленнєві бази даних

Акустичні корпуси (мовленнєві корпуси, мовленнєві бази даних) – важливий тип мовленнєвих ресурсів.

Сьогодні створення акустичних корпусів стає самостійним і популярним напрямом мовленнєвих технологій.

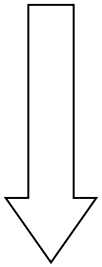
Акустичний корпус як різновид мовленнєвих ресурсів

Акустичний корпус – структурована множина мовленнєвих фрагментів.

Мовленнєвий фрагмент як базова одиниця корпусу – це оцифрований фрагмент мовленнєвого сигналу, який супроводжується асоційованою інформацією певного типу (типів). Така інформація називається також анотацією до мовленнєвого фрагменту.

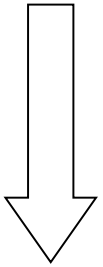
Актуальність акустичних корпусів

Синтез мовлення за текстом



Великі мовленнєві корпуси для конкатенації

Розпізнавання мовлення



Навчання на великих масивах записів анотованого мовлення

Із історії створення акустичних корпусів

Перші акустичні корпуси були створені в першій половині 80-х років минулого століття та використовувалися для тестування й оцінки роботи систем розпізнавання мовлення на однаковому стандартному мовленнєвому матеріалі.

У другій половині 80-х років минулого століття були доведені переваги систем розпізнавання мовлення на основі великих навчальних мовленнєвих корпусів – початок формування напряму мовленнєвих технологій, пов'язаного зі створенням акустичних корпусів.

Із історії створення акустичних корпусів. Коротка характеристика мовленнєвих корпусів 80-х рр. XX ст.

Назва	Використання	Мова	Рік	Характеристика
TI&DIGITS	Розпізнавання цифр і їх послідовностей	Амер. англ.	1984	?
ROAD RALLY	Розпізнавання ключових слів у мовленнєвому потоці	Амер. англ.	1-а пол. 80-х	?
KING CORPUS	Ідентифікація мовця	Амер. англ.		?
RM	Розпізнавання запитів у сфері військових служб	Амер. англ.	2-а пол. 80-х	160 дикторів; словник – 1000 слів; 21000 фраз
WSJ	Дикторонезалежне розпізнавання злитого мовлення	Амер. англ.		Словник – 20000 слів; тексти новин із сфери бізнесу
ATIS	Розпізнавання запитів у сфері обслуговування (авіація)	Амер. англ.		Спонтанні діалоги зі сфери обслуговування (авіація)
TIMIT	Для широкого використання; дикторонезалежне розпізнавання злитого мовлення; наукові дослідження	Амер. англ.	80-90-і рр.	630 дикторів; 2432 фрази; словник – 6229 слів

Із історії створення акустичних корпусів. Координаційні центри

Створення акустичних корпусів – доволі складна технологічна задача, яка потребує значних фінансових і кадрових ресурсів. Для її вирішення у 90-і рр. були створені координаційні центри для збору, зберігання, розповсюдження та створення загальнодоступних і стандартизованих мовних ресурсів, у тому числі мовленнєвих. Серед них:

- **LDC — Linguistic Data Consorcium, http://www.ldc.upenn.edu;**
- **CSLU — Center for Spoken Language Understanding, Oregon Graduate Institute, http://www.CSLU.ogi.edu**
- **ELRA — European Language Resources Association, http ://www.icp.grenet.fr/elra**

Сучасний етап розробки акустичних корпусів

LDC — Linguistic Data Consorcium (США):

- **близько 50 мовленнєвих корпусів;**
- **сотні годин записів;**
- **сучасний інструментарій для обробки мовлення та створення мовленнєвих баз даних.**

Проблема стандартизації:

- **методів;**
- **упорядкування даних;**
- **анотацій;**
- **інструментаріїв корпусних ресурсів.**

Класифікація акустичних корпусів

За метою використання: спеціалізовані, технологічні, загальні, навчально-ілюстраційні.
За типом мовленнєвого матеріалу: дискретне мовлення, неперервне читання, спонтанне мовлення, спеціальні та природні діалоги.
За типом текстового матеріалу: списки складів/слів, фрази, зв'язні тексти; монотематичні та політематичні.
За типом мовленнєвого сигналу: лабораторне, офісне, публічне, телефонне, теле-, радіомовлення, мовлення з акцентом тощо.
За типом анотацій: орфографічний запис, фонемна/фонетична, просодична транскрипція, примітки про індивідуальні особливості мовлення тощо.
Інші класифікації.

Головні проблеми створення акустичних корпусів

1) Фінансове забезпечення;
2) Кадрові ресурси;
3) Стандартизація анотацій та структури АК;
4) Забезпечення загальнодоступності;
5) Розробка інструментарію для роботи з АК;
6) Необхідність створення багаторівневих АК.

Акустичні корпуси для російської мови

- ISABASE
- RuSpeech
- Национальный корпус русского языка

Створення акустичного корпусу українського ефірного мовлення

Характеристика акустичного корпусу:

- **за метою використання: загальний;**
- **за типом мовленнєвого матеріалу: читане мовлення, підготоване мовлення, спонтанне мовлення;**
- **за типом текстового матеріалу: зв'язні політематичні тексти;**
- **за типом мовленнєвого сигналу: публічне мовлення, теле-, радіомовлення, мовлення у природних обставинах.**
- **за типом анотацій: сегментне анотування;**
- **за кількістю мов: двомовний (українська та російська мови).**

Створення акустичного корпусу українського ефірного мовлення

На даний момент анотовано близько 200 звукових файлів.

Це становить приблизно 120 годин звукових записів (разом для української та російської мови).

Корпус містить майже 30 000 слів української мови та майже 35 000 слів російської мови.

Проаналізоване мовлення кількох сотень дикторів.

Створено словник суржику (понад 1000 слів), словник територіальних та соціальних діалектів (понад 450 слів).

Анотування акустичного корпусу українського ефірного мовлення

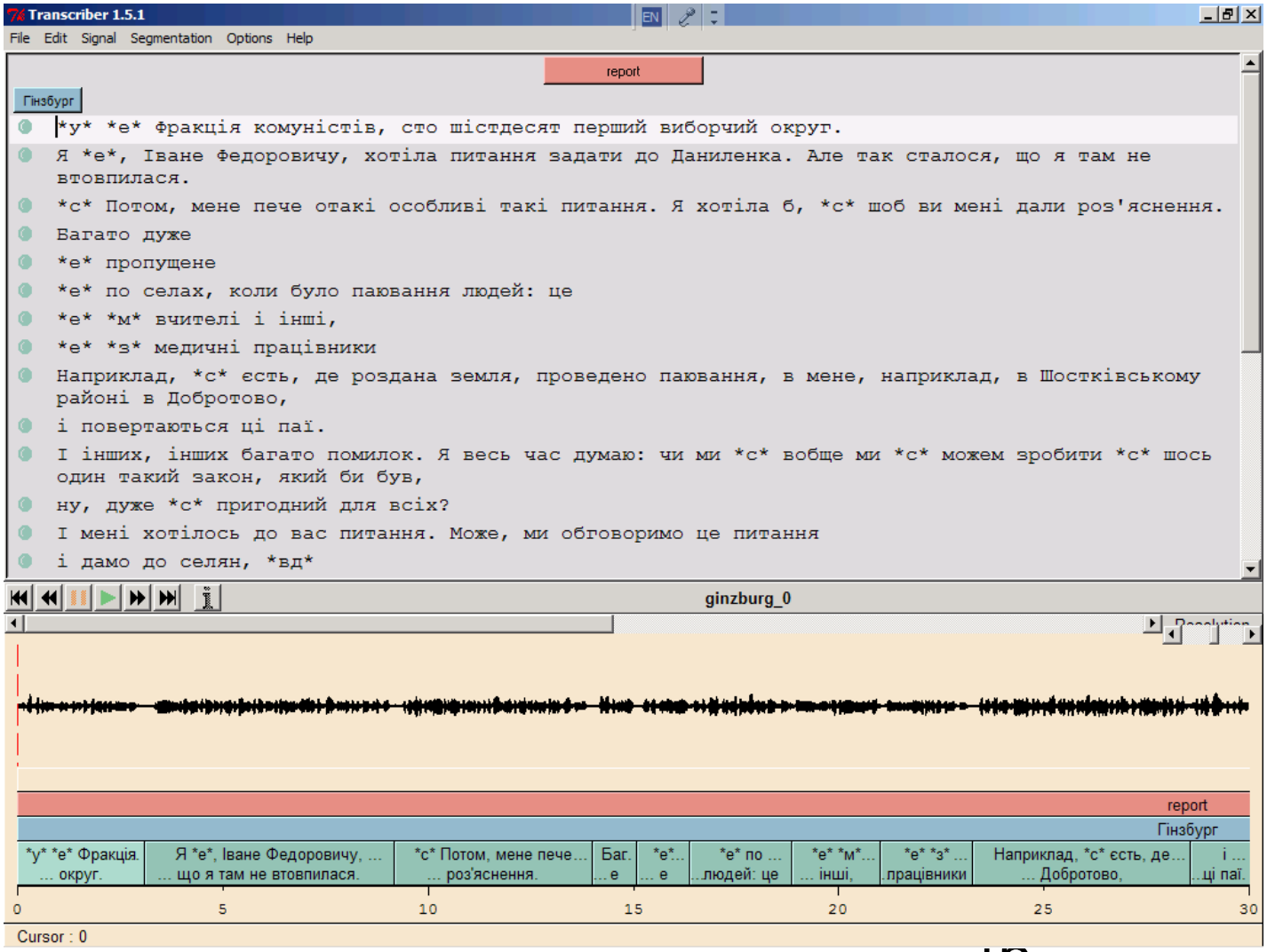
Анотування акустичного корпусу відбувається за допомогою програми Transcriber (http://www.etca.fr/CTA/gip/Projets/Transcriber/)

Особливості анотування акустичного корпусу

Система спеціальних позначень для анотування АК:

- **позначення мови;**
- **позначення нелітературних слів;**
- **позначення способу вимови слів;**
- **позначення фону;**
- **позначення неінформаційних слів та звуків, які вимовляє диктор;**
- **позначення діалогів та хорів;**
- **позначення шуму.**

Зразок розміченого звукового запису



Спасибі за увагу!

robeiko@uasoiro.org.ua