НАЦІОНАЛЬНИЙ КОРПУС КРИМСЬКОТАТАРСЬКОЇ МОВИ

Звіт №2 за 03.10.2022 - 15.03.2023

Мета та основні напрямки роботи

- Провідною метою робіт за звітний період був збір і первинна підготовка текстів кримськотатарською мовою для їх подальшого імпорту на платформу Sketch Engine.
- Основні напрямки роботи для реалізації цієї мети:

Збір друкованих та електронних текстів кримськотатарською мовою

Попередній аналіз і обробка наявних матеріалів, включення у каталог Корпусу

Аналіз цінності джерел для Корпусу філологами

Форматування матеріалів з оцінкою 75% і вище на попередньому етапі

Пошук та опрацювання текстів

- Матеріали обираються до каталогу з усього масиву файлів, що надсилаються на google-форму Корпусу для збору, особисто координаторам проєкту чи виявляються учасниками команди у процесі пошуку.
- Отримані файли аналізуються на предмет наявності у Корпусі (аби запобігти дублюванню), якості сканування (для друкованих видань).
- Далі файлам присвоюється стандартне ім’я, вносяться відповідні реквізити в облікову таблицю.
- Кожен матеріал, внесений в облікову таблицю, проходить аналіз цінності для Корпусу представником Київського національного університету імені Тараса Шевченка Абібуллою Сеїт-Джелілєм.
- Після цього формується відповідна тека у каталозі, де розміщуються початковий файл, файл після оптичного розпізнавання символів (для сканованих матеріалів), відформатований файл (співставлений з оригіналом), файл після контрольної перевірки лінгвістами на предмет граматичних помилок (які не можна віднести до авторської мови), іншомовних вставок.
- Під час збору текстів враховуються основні принципи наукового підходу до укладання мовних корпусів, що передбачає створення сучасного продукту кримськотатарської тематики.

Результати збору літератури

- За звітний період було попередньо оброблено і внесено в каталог Корпусу для подальшої роботи 742 матеріали.
- Серед зібраних матеріалів: книги, періодичні видання (журнали, газети), окремі твори кримськотатарських авторів, підручники, навчально-методичні посібники, документи міжнародних організацій, сценарії озвучки фільмів тощо.
- У каталозі представлені матеріали всіма графічними системами, що використовувалися у кримськотатарській мові: арабською графікою, довоєнною латиницею, кирилицею, сучасною латиницею.
- Формально найстарішими виданнями у каталозі наразі є три номери газети “Terciman” (“Перекладач”) Ісмаїла Гаспринського 1883 року, але фактично є твори та набагато старші (13 і 17 століття), проте сучасного друку.

Результати збору літератури


ترجمان


ПЕРЕВОДЧИКЪ

ЕЖЕНЕДѢЛЬНАЯ ГАЗЕТА

ЛИТЕРАТУРЫ, ОТЕЧЕСТВЕННОЙ И ИНОСТРАННОЙ ЖИЗНИ.

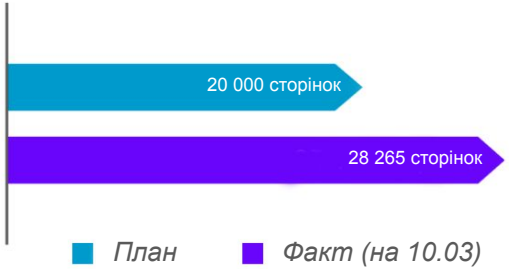
Одне з найстаріших видань у Корпусі — [газета “Терджіман”, 1883 рік, №15](#).

Головний редактор і власник — Ісмаїл Гаспринський.

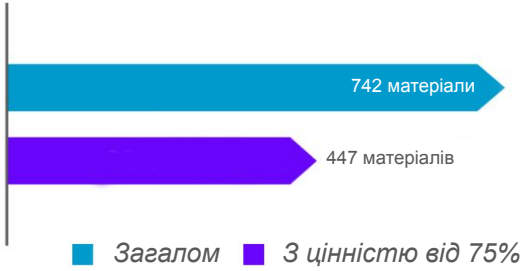
Газета друкувалася арабською графікою, мала переклад російською.

Аналіз збору літератури

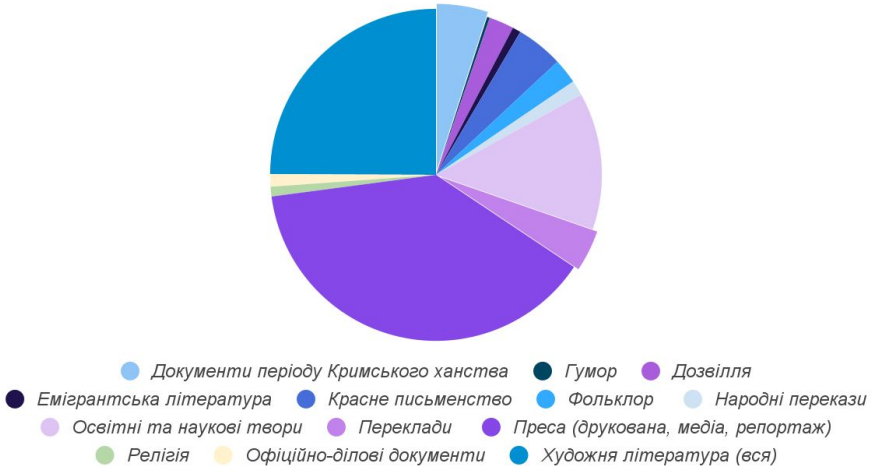
Обсяг зібраних матеріалів



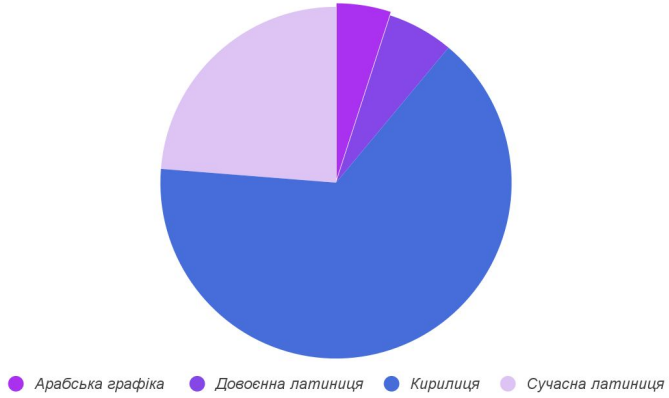
Результати аналізу цінності



Розподіл матеріалів за жанрами

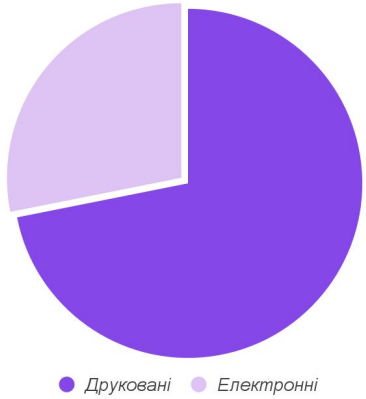


Матеріали, викладені різними графічними системами

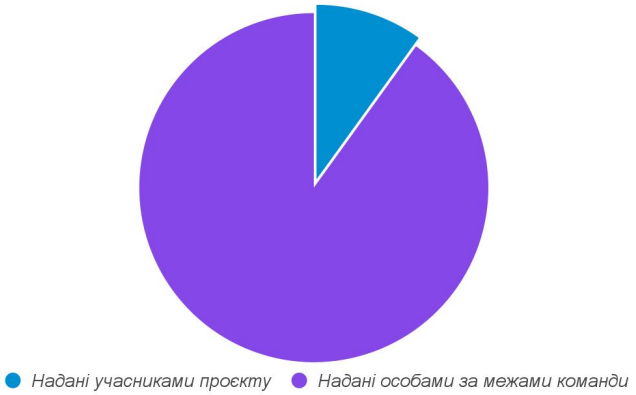


Хронологічний розподіл матеріалів

Формат матеріалів за джерелом надходження



Формат матеріалів за суб’єктом надходження



Джерела походження матеріалів

Матеріали надходять у Корпус кількома шляхами:

- пошук учасниками команди,
- отримання на google-форму за результатами поширення інформації про збір у соціальних мережах,
- залучення осіб, які мають власні архіви текстів,
- залучення кримських періодичних видань і медіаплатформ (що не співпрацюють з окупантами).

Залучення видавництв, авторів матеріалів

Для оптимізації процесу збору текстів командою проєкту було налагоджено комунікацію з рядом видавництв, медіаплатформ, ЗМІ.

У результаті вдалося отримати багато матеріалів кримськотатарською мовою. Зокрема, були залучені

Видавництво “Букрек”

Видавництво “Кримнавчпеддержаввидав”

Радіо Свобода / Крим.Реалії

Ток-шоу “ІНІЦІАТИВА”

Розробка Технічних вимог до створення додаткового функціонала до Sketch Engine для роботи з кримськотатарською мовою

Створення зазначеного функціонала є необхідним, оскільки аналогів такого програмного забезпечення для роботи з корпусами кримськотатарської мови не існує на сьогодні.

За звітний період:

- проведено консультації з технічними спеціалістами,
- виділено в окреме завдання створення словника лематизації.

Фіналізовано:

- технічні вимоги до створення словника лематизації
- технічні вимоги щодо лематизації та анотування
- технічні вимоги щодо імпорту матеріалів на платформу Sketch Engine
- технічні вимоги до транслітерації текстів кириличною графікою*

** Транслітерація ґрунтується на стимулюванні переходу кримськотатарської мови у різних сферах функціонування на алфавіт на основі латинської графіки, що закріплений Постановою Кабінету міністрів України від 22 вересня 2021 р. № 993.*

Словник лематизації

- Під терміном “словник лематизації” слід розуміти реєстр (базу даних) словоформ з позначенням відповідної початкової форми слова, а також основних мовних характеристик (частина мови, рід, число тощо).
- Словник лематизації є необхідним інструментом для запуску та оптимального функціонування модуля лематизації, що впроваджується на платформі Sketch Engine.
- За допомогою модулю лематизації для користувачів Корпусу стане можливим пошук слів за їх початковими формами або мовними ознаками.

Розробка словника лематизації

За звітний період було розроблено словник лематизації, відповідно до узгоджених Технічних вимог.

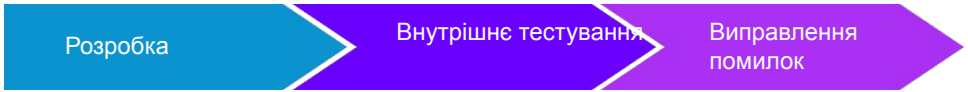
На сьогодні словник лематизації пройшов первинну і проходить фінальну перевірку лінгвістами з боку Набувача проєкту.

Додатково було розроблено інтерфейс для перевірки словника.

Розробка модуля лематизації

- Програмне забезпечення для лематизації та анотування Національного корпусу кримськотатарської мови було розроблено.
- У ході внутрішнього тестування на платформі Sketch Engine виявлено перелік випадків, для яких програма автоматично обирає подвійну характеристику (омоніми тощо), звіт по яких було розглянуто та обрано єдиний варіант для кожної з 2000+ таких помилок.
- Наразі триває виправлення помилок, після чого буде розпочато фінальне тестування створеного функціонала.

Прогрес розробки програмного забезпечення



Розробка модулів імпорту і транслітерації

На сьогодні триває розробка програмного забезпечення для імпорту і транслітерації текстів кримськотатарською мовою на платформі Sketch Engine.

Орієнтовні терміни завершення робіт по блоках:

- словник лематизації - 31 березня 2023 р.;
- транслітерація - 31 березня 2023 р.;
- імпорт - 07 квітня 2023 р.

У процесі фіналізації Технічне завдання до додаткових модулів програмного забезпечення. Орієнтовний термін завершення і фінального узгодження - 14 квітня 2023 р.

Команда проєкту

З 03 жовтня до команди проєкту додатково долучилися 31 учасник (-ця):

SMM - менеджер

створення і розміщення контенту для соціальних мереж, таргетована реклама - 1 особа

Лінгвісти

перевірка відформатованих текстів, експертні консультації, транслітерація текстів з арабської графіки - 4 осіб

Технічні спеціалісти

експертні консультації, розробка словника лематизації - 2 особи

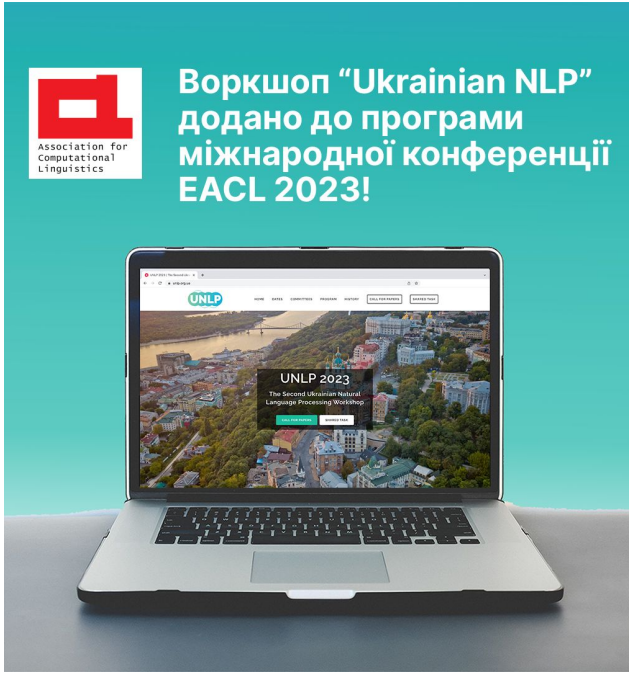
Контент — менеджери

розпізнавання просканованих матеріалів, форматування, зіставлення тексту з оригіналом, пошук матеріалів онлайн - 24 особи

Розвиток команди

Учасників проєкту запрошено до участі у воркшопі Ukrainian Natural Language Processing на 17-ту конференцію Європейського відділення Асоціації комп'ютерної лінгвістики (EACL), що відбудеться у травні 2023 року. У рамках цього воркшопу організатори планують розкрити проблему кримськотатарської мови як мови одного з корінних народів України, а також дізнатися про специфіку роботи над корпусом кримськотатарської мови.

Воркшоп пропонується до програми конференції компанією Grammarly та Українським Католицьким Університетом, який брав участь у розробці Генерального регіонально анотованого корпусу української мови.



Розвиток команди

Команда проєкту також долучилася до Української спільноти синтезу мовлення, в рамках якої запропоновано і розпочато створення систем Кримськотатарського синтезу та розпізнавання мовлення.

Синтез мовлення є практичним інструментом, що використовується системами голосових помічників, онлайн-перекладачів, а також у сфері освіти для забезпечення інклюзивності. Проєкт синтезу та розпізнавання мови є унікальним для кримськотатарської мови та дозволить популяризувати її серед широкого загалу.

За час спільної роботи було записано 4+ годин аудіозаписів — зразків правильної вимови кримськотатарською.

Прослухати аудіо

Прослухати аудіо