Збереження мовної спадщини у цифрову епоху

ЛЬВІВ
2025

УДК 81’374:004; 81’27; 811

Практичний посібник з диджиталізації мовної спадщини

(кейс кримськотатарської /
приклади з цифрового словникарства)

ЄДРПОУ 37201663 • Адреса: 03056, м. Київ, вул. Борщагівська, буд. 117, корп. 3, офіс 13
Виготовлено в Україні.

Редактор: Лілія Ганушевська
Верстка: ТОВ «Майстер Книг»
Технічна підтримка цифрових ресурсів: ГО QIRI’M Young

Підготовлено в межах проєкту з диджиталізації мовної спадщини. За підтримки: House of Europe.

Контакти видавця: zamov@masterknyg.com.ua • https://masterknyg.com.ua/

Формат: А5 • Дог. № 1/230425.

Збереження мовної спадщини у цифрову епоху

ЛЬВІВ
2025

Зміст

1. Вступ: навіщо рідкісним мовам потрібна цифрова присутність

Понад половина з понад 7 000 мов світу знаходиться під загрозою зникнення. За оцінками ЮНЕСКО, мова вмирає кожні 14 днів, і навіть мови з мільйонами носіїв можуть зникнути, якщо між поколіннями порушується передача знань. У XXI ст. небезпеку посилює домінування кількох «глобальних» мов у цифровому просторі. Європейський парламент попереджає, що більшість європейських мов ризикує «цифровим вимиранням», якщо для них не буде створено мовних технологій та ресурсів. Молодь спілкується, навчається і споживає контент переважно онлайн; якщо мова відсутня в інтернеті, вона стає «невидимою» і перестає передаватися наступному поколінню.

Цей виклик ми відчули на власному досвіді під час роботи над диджиталізацією кримськотатарської етнолінгвістичної спадщини в рамках проєкту, підтриманого програмою House of Europe та Goethe-Institut Ukraine. Ми побачили, як оцифрування понад 1000 сторінок унікальних словників із використанням адаптованої OCR-технології, розробка електронних посібників для інших рідкісних мов та створення Національного корпусу кримськотатарської мови суттєво вплинули на збереження культурної ідентичності. Завдяки інтеграції цієї спадщини в онлайн-платформи, зокрема Sketch Engine, вдалося створити доступні інструменти для дослідників та освітян. Водночас цифрові технології відкривають унікальні можливості: доступ до онлайн-освіти, машинний переклад, електронні словники, соціальні мережі можуть стати потужними інструментами відродження мов. Європейський парламент визнає, що мови можуть бути порятовані завдяки цифровим сервісам; для цього спільноти мають збирати дані, створювати корпуси й інструменти, щоб молодь могла активно користуватися рідною мовою в мережі. Важливим є також

міжнародне співробітництво, що дозволяє ділитися досвідом та технологіями, які вже підтвердили свою ефективність. Саме цю двояку природу цифровізації – як загрозу і водночас унікальну можливість – висвітлює цей посібник.

2. Оцінка життєздатності мови: класичні та цифрові критерії

2.1 Класифікація ЮНЕСКО

ЮНЕСКО розробила детальну систему з дев'яти критеріїв, які допомагають оцінити, наскільки життєздатною є конкретна мова та наскільки вона ризикує зникнути. Ці критерії охоплюють різні аспекти, які впливають на виживання мов.

Перший і найважливіший критерій – це міжпоколінна передача, тобто чи передається мова від старшого покоління до молодшого.

Другий критерій – кількість носіїв мови. Чим менше людей говорять мовою, тим більше вона вразлива.

Третій критерій стосується частки носіїв мови в загальній популяції: навіть якщо кількість носіїв може бути відносно великою, мова все одно перебуває в небезпеці, якщо вона розпорошена серед значно більшої кількості носіїв інших мов.

Наступний критерій – використання мови у різних суспільних сферах, таких як освіта, бізнес, урядові установи та побутове спілкування. Якщо мова не використовується активно в різноманітних сферах життя, це свідчить про її поступове витіснення іншими мовами.

П'ятий критерій оцінює здатність мови адаптуватися до нових сфер використання, таких як технології та медіа. Якщо мова не розвивається і не пристосовується до сучасних викликів, її витісняють інші мови, що пропонують кращі комунікаційні можливості.

Шостий критерій стосується наявності освітніх матеріалів. Це підручники, посібники, словники, онлайн-курси, що дозволяють вивчати та викладати мову. Відсутність таких матеріалів значно ускладнює збереження мови.

Сьомий критерій – мовна політика держави або регіону. Підтримка з боку державних органів або, навпаки, її відсутність відіграє ключову роль у збереженні чи зникненні мови.

Восьмий критерій – це ставлення самої мовної спільноти до своєї мови. Позитивне ставлення та бажання підтримувати і розвивати рідну мову можуть суттєво посилити її позиції, тоді як байдужість чи негативне ставлення, навпаки, прискорюють її зникнення.

Останній, дев'ятий критерій – це наявність якісної мовної документації. Це записи мови, письмові документи, архіви, які можуть стати основою для її відродження, навіть якщо носіїв залишилося зовсім небагато.

За допомогою цих критеріїв ЮНЕСКО класифікує мови за ступенем загрози їхньому існуванню на п'ять категорій: «уразлива», «безумовно загрожена», «небезпечно загрожена», «критично загрожена» та «мертва». Наприклад, кримськотатарська мова класифікується як «небезпечно загрожена». Це означає, що мова перебуває в дуже складній ситуації: кількість її носіїв зменшується, міжпоколінна передача порушена, а використання в сучасних сферах та медіа обмежене. Відповідно, ця мова потребує активних заходів підтримки та відродження.

2.2 Цифрова життєздатність

Становище мови у цифровому середовищі суттєво відрізняється від традиційних способів оцінки життєздатності мови. Це пов'язано з тим, що цифрові технології мають власні особливості й критерії, які

визначають присутність і поширення мови в онлайн-просторі. У сучасному світі мова, яка відсутня в цифрових середовищах, втрачає можливості для комунікації, навчання, бізнесу і культури, що негативно позначається на її майбутньому.

Група дослідників Digital Language Diversity Project (DLDP) разом з Національною дослідницькою радою Італії (CNR) розробили спеціальну шкалу цифрової життєздатності мов. Відповідно до цієї класифікації, мови можуть перебувати на одному з кількох рівнів цифрового розвитку:

- Перед-цифровий етап означає повну або майже повну відсутність мови в цифровому середовищі. Це ситуація, коли немає навіть базових цифрових ресурсів, а носії мови не мають належного доступу до інтернету або цифрових інструментів.

- Дрімотний рівень характеризує мову, яка має мінімальну присутність в інтернеті. Можуть існувати окремі згадки або фрагменти текстів, але немає активних ресурсів, які регулярно використовуються спільнотою.

- Зародковий етап описує початкову цифрову активність: починають створюватися окремі ресурси, такі як прості вебсайти, сторінки в соціальних мережах, невеликі словники чи перші спроби цифрового архівування текстів.

- Етап розвитку свідчить про активну роботу спільноти щодо створення цифрових матеріалів, таких як корпуси текстів, онлайн-словники, платформи для навчання та інші цифрові ресурси, які активно використовуються у повсякденному житті носіїв мови.

- Життєздатний рівень відображає впевнене становище мови у цифровому просторі: мова активно використовується в соціальних мережах, створюються блоги, відеоконтент, активно ведуться дискусії й обговорення.

- Квітучий етап – найвищий рівень цифрової життєздатності. Це означає, що мова повністю інтегрована у цифровий світ, має потужні онлайн-ресурси, поширена у великих сервісах, активно підтримується великими технологічними платформами та має високий рівень локалізації.

У рамках DLDP було створено Digital Language Survival Kit – спеціальний інструментарій, який допомагає мовним спільнотам підвищити цифрову життєздатність своєї мови. Цей набір рекомендацій складається з трьох головних напрямів:

- «Підготовка ґрунту» (Tilling & Seeding). Це перший і найважливіший етап, який передбачає забезпечення доступу до інтернету для всіх носіїв мови, розвиток базових цифрових навичок, таких як робота з текстовими редакторами, соціальними мережами, пошуковими системами, а також створення фундаментальних мовних ресурсів – словників, баз текстів, простих онлайн-платформ.

- «Полив» (Watering). На цьому етапі важливо активно заохочувати спільноту використовувати мову в цифровому просторі. Це включає ведення сторінок у соціальних мережах, регулярне створення й оновлення блогів, сайтів, створення підписів до фотографій, відео, редагування статей у Вікіпедії та інших онлайн-ресурсах. Активне використання мови у різних цифрових форматах значно посилює її присутність у цифровому світі.

- «Збирання врожаю» (Harvesting & Consuming). Цей етап орієнтований на максимально глибоку інтеграцію мови в цифровий простір, що передбачає локалізацію популярних програм та додатків, розробку й впровадження технологій машинного перекладу, створення спеціалізованих доменів та інтернет-сервісів, які забезпечують повноцінну цифрову екосистему мови.

Оцінювання цифрової життєздатності мов здійснюється за кількома ключовими показниками. Серед них – наявність стабільного інтернет-підключення, повна підтримка стандарту Unicode для адекватного відображення текстів, наявність мовних ресурсів (корпусів текстів, словників, навчальних матеріалів), кількість і якість локалізованих програм та сервісів, активність спільноти в соціальних мережах і онлайн-обговореннях. Чим вище значення цих показників, тим більше можливостей відкривається для подальшого розвитку мовних технологій, що є критично важливим для виживання мови в сучасному цифровому середовищі.

3. Кримськотатарська мова як приклад

3.1 Соціолінгвістичний контекст

Кримськотатарська мова належить до тюркської мовної сім'ї та має складну та драматичну історію, яка безпосередньо вплинула на її сучасний стан. Щоб зрозуміти, чому сьогодні ця мова опинилася на межі зникнення, важливо коротко оглянути її історичний і соціальний контексти.

Кримськотатарська мова традиційно була мовою спілкування кримських татар – корінного народу Кримського півострова. Протягом багатьох століть вона успішно функціонувала як мова освіти, культури, торгівлі та побутового спілкування. Однак у XX столітті одним із найтрагічніших моментів стала депортація кримських татар у 1944 році. Радянська влада виселила майже весь кримськотатарський народ у віддалені регіони Центральної Азії та Росії, що призвело до масових людських втрат і руйнації соціальної структури. В умовах заслання та заборони на використання кримськотатарської мови у публічній сфері вона поступово втрачала позиції.

Лише наприкінці 1980-х років, коли кримські татари почали повертатися на батьківщину, з'явилась можливість відновлювати мову та культуру. Проте цей процес відбувався повільно, а мовна ситуація залишалася складною. Окупація Криму у 2014 році ще більше загострила проблему, адже вона значно ускладнила розвиток освіти кримськотатарською мовою та її використання у публічних установах і медіа.

За результатами дослідження фонду «Східна Європа», сьогодні лише близько 20–25% кримських татар, які проживають в Україні, вільно володіють рідною мовою. Цей факт свідчить про критичний рівень загрози її існуванню. Одна з найбільших проблем – розрив у міжпоколінній передачі мови. Старші покоління, які народилися до депортації або в перші роки після повернення, продовжують активно використовувати кримськотатарську мову у щоденному спілкуванні. Їхні діти, хоча і розуміють мову, але вже значно менше її використовують, особливо у професійному чи публічному житті. Внуки ж цього покоління, які народилися вже в незалежній Україні чи після 2014 року, часто не володіють мовою навіть на базовому рівні, обмежуючись лише окремими фразами чи словами.

Причиною такого стану є, зокрема, недостатній рівень освітньої підтримки, недостатня кількість кваліфікованих викладачів, а також брак сучасних навчальних матеріалів, які б відповідали актуальним потребам молоді. Важливим є і психологічний фактор: багато молодих кримських татар не відчувають достатньої мотивації для вивчення мови через її слабке представлення у медіа, культурі, соціальних мережах та професійному житті.

Щоб змінити цю ситуацію, важливо посилити підтримку на державному рівні, забезпечити створення та поширення сучасних навчальних матеріалів, популяризувати мову у соціальних мережах, медіа та молодіжних просторах. Також необхідно заохочувати молодь до активного використання мови у повсякденному спілкуванні та створювати сприятливі умови для цього – від організації мовних курсів до підтримки цифрових ініціатив, таких як

створення онлайн-словників, мобільних додатків, платформ для навчання та спілкування.

Таким чином, відновлення та підтримка кримськотатарської мови – це не лише завдання самих кримських татар, але й спільний обов'язок українського суспільства, адже збереження мовної різноманітності є важливою частиною загальнокультурної ідентичності країни.

3.2 Національний корпус

У 2023 році відбулася важлива подія в історії кримськотатарської мови – було створено Національний корпус кримськотатарської мови. Цей масштабний проєкт вдалося реалізувати завдяки підтримці програми EGAP та Міжнародної благодійної організації «Фонд Східна Європа». Такий корпус є сьогодні найбільшою електронною базою текстів кримськотатарською мовою і виконує ключову роль у збереженні та розвитку мови в цифрову епоху.

Національний корпус – це не просто архів текстів, а складний інструмент, що дозволяє проводити глибокий і всебічний аналіз мови. Він містить широкий спектр текстів – від класичної літератури та історичних документів до сучасних публікацій у медіа та наукових робіт. Завдяки цьому користувачі можуть досліджувати, як змінюється мова, як вона адаптується до сучасних реалій, які слова поступово виходять з ужитку (архаїзми), а які навпаки – набирають популярності (неологізми).

Одна з головних переваг Національного корпусу – можливість швидкого й точного аналізу контекстів вживання слів. Наприклад, якщо дослідник чи викладач хоче дізнатися, як часто певне слово чи вираз зустрічається в різних текстах, які слова найчастіше використовуються разом з ним (колокації), корпус за кілька секунд надасть усю необхідну інформацію. Це значно ефективніше, ніж переглядати сотні сторінок вручну, що економить багато часу і робить дослідження більш точними та ґрунтовними.

Корпус активно використовується в лексикографії, тобто у створенні словників. Лексикографам він допомагає підібрати автентичні приклади для тлумачних та історичних словників, демонструючи, як слова використовуються в реальних життєвих ситуаціях. Це також важливо для освітніх потреб: викладачі можуть легко знаходити цікаві приклади для занять, а студенти – вивчати реальну, живу мову.

Важливим напрямком застосування корпусу є розвиток мовних технологій. Завдяки аналізу великих обсягів текстів можна створювати сучасні сервіси автоматичного перекладу, які значно полегшують комунікацію та роботу з кримськотатарською мовою. Також корпус використовується для розробки програм перевірки правопису та граматики. Основою для цих технологій є N-грамні моделі, що базуються на частотності вживання слів та словосполучень.

Реалізація цього складного та амбітного проєкту була нелегкою. Понад 30 фахівців та волонтерів протягом року працювали над створенням і наповненням корпусу.

Загалом було опрацьовано понад 900 різноманітних текстів. Робота вимагала вирішення низки складних завдань. Наприклад, через окупацію Криму та відсутність доступу до кримських бібліотек важко було знайти та оцифрувати багато цінних історичних і культурних матеріалів. Крім того, особливим викликом стала наявність чотирьох різних графічних систем письма, які використовувалися в різні історичні періоди кримськотатарської мови (арабиця, латиниця, кирилиця та сучасна латиниця). Команда мала розробити спеціальні технічні рішення, щоб конвертувати й уніфікувати тексти в єдиному форматі, придатному для автоматичного аналізу.

Підтримка проєкту на державному та міжнародному рівні відіграла важливу роль у його успіху. Міністерство реінтеграції тимчасово окупованих територій України надало значну підтримку та офіційне сприяння. Важливим міжнародним партнером стала Програма EGAP, яка фінансується Урядом Швейцарської Конфедерації через Швейцарську агенцію з розвитку та співробітництва (SDC). Посольство Швейцарії також брало активну участь у підтримці цього проєкту, наголошуючи на його значенні для збереження культурного різноманіття та міжнаціонального діалогу.

Таким чином, Національний корпус кримськотатарської мови став не просто черговим цифровим ресурсом, а важливим інструментом збереження і розвитку мови, який сприяє її активному використанню у сучасних умовах. Він допомагає долати наслідки історичних травм і сприяє відновленню міжпоколінної передачі мови, забезпечуючи її присутність у різних сферах суспільного життя. Реалізація такого проєкту є прикладом успішного об’єднання зусиль

громадських активістів, науковців, державних інституцій та міжнародних партнерів заради спільної мети – збереження кримськотатарської мови та культурної спадщини.

3.3 Інші цифрові ресурси

У сучасному цифровому світі для збереження будь-якої мови важливим є її присутність у популярних онлайн-сервісах та ресурсах. Останніми роками кримськотатарська мова досягла значних успіхів у цій сфері, що є результатом злагодженої роботи активістів, мовознавців, міжнародних організацій та підтримки державних структур.

Один із найяскравіших успіхів у цьому напрямку стався у червні 2024 року, коли компанія Google офіційно оголосила про включення кримськотатарської мови до переліку підтримуваних у сервісі Google Translate. Це був значний прорив, який дозволив кримським татарам та всім зацікавленим людям використовувати потужний інструмент машинного перекладу для перекладу текстів з і на кримськотатарську мову. Міністерство закордонних справ України позитивно оцінило цю подію, підкреслюючи, що інтеграція мови до такого впливового сервісу робить кримськотатарську культуру набагато доступнішою для міжнародної спільноти, допомагає поширювати інформацію про історію, культуру та сучасне життя кримських татар. Це також суттєво сприяє збереженню та розвитку мови, роблячи її активною частиною глобальної цифрової екосистеми.

Окрім перекладацьких сервісів, значним напрямком розвитку цифрових ресурсів є Вікіпедія

кримськотатарською мовою. За останні роки кримськотатарська Вікіпедія активно розвивається і регулярно поповнюється новими статтями завдяки ентузіазму волонтерів та громадських активістів. Це не лише надає носіям мови доступ до важливої інформації рідною мовою, але й дозволяє зберігати та поширювати знання про власну історію, традиції, видатних особистостей і культурні події.

Важливим компонентом сучасного цифрового середовища є медіаконтент кримськотатарською мовою, який постійно розширюється. Сьогодні існує ціла низка YouTube-каналів, де кримськотатарські активісти, блогери та освітяни регулярно створюють та публікують відеоматеріали на різні теми – від історії та культури до сучасних подій і мовних уроків. Окрім того, набирають популярності подкасти кримськотатарською мовою, що охоплюють різні сфери життя і дозволяють людям практикувати слухання та спілкування кримськотатарською мовою у повсякденних форматах. Онлайн-курси також стають дедалі доступнішими та популярнішими серед тих, хто бажає вивчати кримськотатарську мову з нуля або вдосконалювати свої навички.

Наявність великої електронної бази текстів значно спрощує інтеграцію мови в сучасні цифрові технології. Важливим кроком у цьому напрямку може стати впровадження кримськотатарської мови в операційні системи комп'ютерів та мобільних пристроїв, а також створення зручних клавіатурних розкладок, адаптованих під специфіку мови. Така інтеграція значно підвищить зручність використання мови в цифрових середовищах і сприятиме її активному використанню в повсякденному житті.

Весь цей успіх демонструє, що навіть мова, яка перебуває під критичною загрозою зникнення, може суттєво посилити свою цифрову присутність, якщо спільнота системно та цілеспрямовано працює над її збереженням та розвитком. Це є прикладом того, як сучасні технології, об'єднані з активністю громадських ініціатив та підтримкою державних структур, можуть не тільки захистити, але й надати нові можливості для розвитку мов, які перебувають у складній ситуації.

4. Створення корпусу та лексичних ресурсів

4.1 Пошук і оцифрування матеріалів

Створення якісного лексичного корпусу та інших мовних ресурсів – це процес, що складається з кількох важливих етапів, першим і одним з найважливіших з яких є пошук, збір та оцифрування текстових матеріалів. Саме від якості цього етапу залежить успішність подальшої роботи з корпусом та можливості для його застосування у різних сферах.

Спочатку необхідно зібрати документи, які планується включити до корпусу. Це можуть бути газети, журнали, книжки, архівні рукописи, листування, офіційні документи, наукові публікації та інші джерела. Для кримськотатарської мови ця задача мала додаткову складність через історичні обставини: депортації, окупацію та відсутність доступу до значної частини джерел, які зберігаються в бібліотеках на тимчасово окупованій території Криму. Саме тому дуже важливо було залучати волонтерів та партнерів з різних регіонів України та за її межами, які могли надавати доступ до приватних колекцій, сімейних архівів і маловідомих матеріалів.

Після того як матеріали були зібрані, наступним кроком стала процедура їх оцифрування, тобто сканування або фотографування документів у високій якості. Щоб забезпечити достатню якість оцифрованих матеріалів для подальшої роботи (розпізнавання тексту та аналізу), важливо дотримуватися певних технічних стандартів.

Для сканування текстових документів (наприклад, книжок чи газет) рекомендується роздільна здатність сканування в межах 300–600 dpi (точок на дюйм). Цей дозвіл дозволяє отримати чіткі зображення тексту, які потім легко піддаються оптичному розпізнаванню (OCR). Для ілюстрацій, креслень та інших графічних матеріалів, які містять багато деталей, оптимальним є значення до 1200 dpi. Такий підхід гарантує максимальну якість і збереження усіх деталей, які можуть бути важливими для дослідників або лексикографів у майбутньому.

Для збереження отриманих цифрових копій важливо вибрати відповідні формати файлів. Для архівного зберігання оптимальними є формати TIFF з безвтратним стисненням, оскільки вони дозволяють зберігати файли без втрати якості, незалежно від подальших маніпуляцій. Формат PDF/A також є відповідним варіантом для документів, які планується активно використовувати або поширювати, адже цей формат спеціально розроблений для довготривалого зберігання і забезпечує стабільність відображення на різних пристроях.

Важливою частиною процесу оцифрування є також дотримання правил безпеки та резервування отриманих цифрових матеріалів. Це означає, що необхідно регулярно створювати резервні копії всіх цифрових файлів, використовуючи для цього кілька незалежних сховищ – наприклад, поєднання хмарних сховищ та локальних серверів. Такий підхід гарантує, що у разі технічних збоїв або інших непередбачуваних ситуацій оцифровані матеріали не будуть втрачені та залишатимуться доступними для подальшого використання.

Окрім технічних вимог, в процесі оцифрування необхідно також враховувати питання авторських прав та етичні аспекти. Перед початком сканування важливо отримати дозвіл від авторів або власників прав на документи. Це особливо актуально для сучасних публікацій або матеріалів, які охороняються законодавством про авторське право. Для історичних документів, де авторські права можуть бути не такими очевидними, слід керуватися принципом дотримання поваги до авторства та збереження автентичності оригінальних текстів. Не менш важливою є вимога дотримання конфіденційності особистих даних, які можуть бути присутніми в документах, особливо якщо мова йде про приватні листування або архівні матеріали.

Не менш значущою є рекомендація не змінювати зміст історичних документів під час оцифрування. Навіть якщо в текстах зустрічаються помилки, неточності або застарілі формулювання, їх слід залишати без змін, щоб забезпечити максимально точну передачу первісного змісту та можливість подальших досліджень.

Таким чином, процес пошуку та оцифрування матеріалів є ключовим етапом у створенні цифрових лексичних ресурсів і потребує уважності, точності та дотримання чітких стандартів і рекомендацій. Від його якості значною мірою залежить ефективність і надійність всього лексичного корпусу, що робить можливим його подальше використання у лінгвістиці, освіті, дослідженнях і повсякденній мовній практиці.

4.2 Оптичне розпізнавання тексту (OCR)

Здається, що оцифрувати книжку — це просто: відсканував сторінки, зберіг як PDF і готово. Але для живої мови цього недостатньо. Уявіть собі скан газети початку XX століття: ви не зможете шукати по тексту, не зможете виділити слово, скопіювати уривок чи порахувати, скільки разів зустрічається певне поняття. Щоби мова справді «ожила» в цифровому просторі, текст має бути редагованим, пошуковим і аналізованим. Саме це й робить технологія OCR — оптичне розпізнавання символів.

Для багатьох великих мов — англійської, німецької, французької — OCR давно вже працює автоматизовано. Але коли йдеться про мову меншості, ще й таку, яка за своє життя змінила кілька графічних систем, — вам не пощастить знайти готову кнопку «розпізнати кримськотатарською». І тут починається справжня магія — але не та, що трапляється сама собою, а та, що створюється руками спільноти.

У нашому випадку — під час роботи над Національним корпусом кримськотатарської мови — OCR став першим і найтривалішим технічним викликом. До нас надходили документи у найрізноманітніших форматах: старі підручники, фольклорні збірки, газети, листівки, машинописні й рукописні архіви. Частина з них була вже в PDF, але просто як картинка. Інші доводилося сканувати вручну, часто з пошкоджених або вигорілих сторінок. Рівень якості — від «ідеально» до «взагалі нечитабельне». До цього додайте кілька шрифтів, які давно вийшли з ужитку, нестандартну орфографію та літери, яких немає на вашій

клавіатурі — і от вам щоденна реальність невеликої мовної команди.

Першим інструментом OCR, з яким ми почали працювати, став ABBYY FineReader — комерційна програма, яка має чималий потенціал і часто використовується в бібліотечних архівах. Вона дозволяє не лише розпізнавати текст, а й вбудовувати його у PDF, створюючи searchable-файл — такий, у якому можна робити пошук по слову. FineReader має один великий плюс — він зручний у роботі з інтерфейсом «drag-and-drop», а також дозволяє вручну коригувати текст просто в процесі розпізнавання. Але є й мінус: вбудована підтримка мов обмежена, і кримськотатарська — звісно, не входить до переліку.

Тому паралельно ми почали експериментувати з Tesseract OCR — безплатним рушієм з відкритим кодом, який розробляє спільнота за підтримки Google. Його можна навчити будь-якої мови. Але — і тут важливий момент — Tesseract нічого не зробить без вас. Ви маєте вручну зібрати корпус сторінок, створити до кожної текстову відповідність, поєднати це все у спеціальні файли з розміткою, запустити тренування моделі, перевірити результат і — найімовірніше — зробити це ще кілька разів, бо з першого разу все не запрацює.

Так і було у нас. Перші результати були кумедні: система плутала «і» з «ı», «с» з «е», зливала два слова в одне, не бачила апострофів або, навпаки, знаходила їх там, де їх не було. Але з кожною ітерацією модель покращувалась. Ми зрозуміли: головне — не чекати ідеального результату, а запустити живий цикл: скан → розпізнавання → перевірка → повторне тренування → скан наступного документа.

Окремим викликом стали діакритичні знаки. Багато кримськотатарських літер (ğ, ş, ñ, â). Щоб Tesseract розпізнавав їх правильно, ми мусили перевірити, чи ці символи взагалі входять до Unicode і чи підтримуються в шрифтах, які ми використовували. Часто доводилось створювати власні кастомні шрифти або вручну коригувати бокси символів у редакторі.

Щоб систематизувати роботу, ми розробили власну інструкцію для волонтерів. Вона містила:

- приклади назв файлів (напр.: NoAuthor__Terciman--1883.pdf);
- шаблони для FineReader-проєктів;
- списки типових помилок OCR;
- регулярні вирази для автоматичного пошуку й заміни помилок;
- чітку інструкцію, як зберігати проміжні версії.

Ми також створили мініглосарій типових OCR-проблем:

- розпад слова: коли кінець слова переноситься на наступний рядок, але програма думає, що це два окремих слова;
- злиття слів: коли два сусідні слова зливаються в одне без пробілу;
- плутанина символів: латинська c і кирилична с, латинська a і кирилична а тощо;
- невидимі знаки: коли апостроф розпізнається як пробіл або зовсім зникає.

Для того щоб перевірити якість розпізнавання, ми тестували розпізнані документи на «сліпих» прикладах — тобто таких, які не входили до навчального набору. Це

дозволяло реально оцінити, наскільки добре модель узагальнює знання. Також ми створили окремі зразки для тестування рідкісних символів і «схожих літер» — наприклад, окремі сторінки лише з тюркськими діакритиками.

Коли модель почала давати стабільно добрий результат, ми перейшли до масштабування. Це означало, що ми вже могли запускати розпізнавання сотень сторінок без щоденного ручного втручання. На цьому етапі OCR став не просто інструментом, а двигуном — саме він забезпечив ті 250 000 речень, що лягли в основу корпусу.

Але й тоді ми не припинили вдосконалювати процес. Наприклад, ми експериментували з попередньою транслітерацією: якщо текст був кирилицею, ми могли спочатку розпізнати кириличний варіант, а потім вже латинізувати через окремий модуль. Це дало змогу скористатися сильнішою кириличною підтримкою FineReader і при цьому отримати результат у латинці.

І останнє — OCR відкрив двері до наступного рівня: лематизації. Щойно ми мали тексти в редагованому форматі, ми могли під'єднати морфологічний аналіз, почати «розбирати» кожне слово на частини, визначати початкову форму, граматичні ознаки тощо. Але про це — у наступному розділі.

Якщо сказати просто: OCR — це перетворення сторінки на текст. Але якщо сказати точно — це перетворення пам’яті на ресурс, а мови — на повноцінного учасника цифрового світу. Це саме той крок, який дає початок усьому: корпусу, словникам, машинному перекладу, освітнім програмам. І

саме тому він потребує терпіння, уваги та командної роботи. Однак ці зусилля виправдані, бо кожна просканована сторінка — це ще один рядок, якого не загубить час.

4.3 Створення корпусу

Після того як ви отримали «чистий» електронний текст — без помилок OCR, з виправленими символами й відсутніми артефактами — настає ключовий етап: організація цих текстів у корпус. Це не просто тека з файлами, а ретельно продумана база, з якої можна витягти соціолінгвістичні зміни, жанрові відмінності, частотні шаблони та граматичні закономірності. Як застерігають дослідники з Бірмінгемського університету, якщо корпус побудований хаотично, то результати аналізу можуть ввести в оману.

Спочатку — про розмір. Існує спокуса зібрати максимально великий корпус, однак обсяг має відповідати вашим цілям: якщо вам потрібно дослідити рідкісні граматичні конструкції — великий корпус дає більше шансів їх зустріти. Але навіть корпус із кількох десятків тисяч слів може бути дуже інформативним для частотних аналізів і словникової роботи — виводити найуживаніші слова, показувати типові колокації, давати початкову підставу для словника або навчального матеріалу.

Другий принцип — баланс і репрезентативність. Найкращі корпуси включають різні жанри, стилі, часові періоди та тематики. Наприклад: художні тексти та газетні статті, науково-популярні вставки та усні інтерв’ю, сучасні блоги й історичні документи. Ідея в тому, щоб корпус був мікрокосмом мови — у масштабі. Важливо також обирати

тексти таким чином, щоб не було однотипних джерел, які перекреслюють баланс. У корпусах мов-меншин, якщо включені усні дані, варто додавати метадані про вік, стать, соціальне становище інформанта, щоби уникнути спотворення зразка реальної мовної ситуації.

Третій момент — електронний формат. Корпусні інструменти працюють із простими текстовими файлами (.txt) або з XML/TEI-форматами, де структура документів чітко описана. PDF, DOCX, форматування з таблицями й картинками — все це руйнує структуру корпусу та ускладнює аналіз. Тому перед включенням ваш текст повинен бути очищеним, без зайвих символів, без маркерів сторінок, без PDF-позначок. У корпусі важливо зберігати розбиття на параграфи, речення, але у форматі, зрозумілому автомату.

І останнє — анотація. Кожен текст має бути позначений не лише його змістом, а й контекстом: автор, дата створення, жанр, джерело, регіон, орфографічна редакція, тип шрифту (якщо релевантно). Таку метадану можна включити у файл поряд із текстом — як окремий .meta або й у самому txt через хедер. А потім — лінгвістичне тегування: частини мови (POS-теги), леми, граматичні категорії (рід, число, відмінок, час), можливо — розмітка синтаксису, названі сутності (імена, локації), корпусні позначення. Це значно розширює можливості: ви можете шукати, наприклад, по всіх прислівниках у текстах 1930-х або по родових формах іменників у сучасних блогах.

Уявіть тепер — корпус кримськотатарських текстів, де кожен документ має:

- назву, автора, рік,
- жанр (газета, фольклор, інтерв'ю, історичний документ),
- фонетичну або орфографічну редакцію,
- усякий механічний POS-тег (наприклад, N-noun, V-verb тощо),
- окремо виділені леми.

Це дає надзвичайне поле для аналізу: від частотного словника до словоформ, частин мови, синтаксичних моделей, умовних конструкцій тощо.

Ключовий момент — повторюваність і прозорість процесу. У теорії та рекомендаціях з Бірмінгема часто згадується принцип реплікації: документуйте рішення, шаблони, код, скрипти. Наприклад, як ви вибирали тексти, який розмір зрізу ви брали (перші 2000 слів, середина чи full-text), як нормалізували перенесення слів, як виправляли апострофи чи дефіси. В описі корпусу має бути чітко вказано: «ми взяли 50 газетних текстів 1990–2000 рр., по 3000 слів кожен, вибір перший фрагмент» — це додає наукової цінності та дозволяє іншим оцінити ваш ресурс об'єктивно.

British National Corpus вважають взірцевим — 100 мільйонів слів із голосами з історичних газет, романів, технічних текстів, усних інтерв'ю — і все ретельно збалансовано за жанрами та періодами. І хоча мова кримськотатарська значно менша, але принципи ті самі: хоча б невеликий, але репрезентативний корпус дає значно більше користі, ніж великий хаотичний набір текстів.

Технологічно — хороший корпус починається з простої структури:

```
/corpus/
    /texts/
       doc1.txt
       doc2.txt
    /meta/
       doc1.meta
       doc2.meta
    /annotation/
       doc1.conll
       doc2.conll
```

У `.meta` — YAML або TSV зі стовпцями: назва, жанр, рік, джерело, мовна редакція. У `.conll` або іншому форматі — тегування слів і лем.

Важливо також протестувати корпус перед масовим аналізом: перевірити узгодженість тегів, порівняти частоти POS-категорій між жанрами, побачити, чи не домінує один жанр. Хорошою практикою є застосувати простий інструмент, як-от **AntConc** або **SketchEngine**, щоб хлопці з волонтерської групи могли бачити частотні списки, колокації, n-грами — і на основі цього скоригувати включення або тегування текстів.

Для усного корпусу — коли у вас є записи інтерв'ю — необхідно визначити рівень транскрипції: чи ви позначаєте паузи, повтори, хіхікання, невизначені

звуки. Такі рішення слід прийняти на початку і документувати. Якщо працюють кілька транскрибаторів — важливо провести віктим-тестування на консистентність (inter-annotator agreement). Це гарантує, що транскрипція в різних текстах подібна і порівнянна.

Отже, корпус — це не просто набір текстів, а структурована, документована, позначена база, де розмір відповідає вашій меті, баланс гарантує репрезентативність, чистий формат дозволяє аналіз, а анотація робить корпус ресурсом для лінгвістичної та соціокультурної інтерпретації.

Саме такий корпус стає справжнім фундаментом: він живе, розвивається, з ним можна запускати лематизацію, POS-тегінг, синтаксичний аналіз, обчислювати частотні словники, знаходити тенденції, робити навчальні вправи.

4.4 Лексикографічні ресурси

Створення сучасного електронного словника є одним із найважливіших напрямів цифрової підтримки мови. Словники — це не лише зручний інструмент для користувачів, а й основа для функціонування систем автоматичного перекладу, перевірки орфографії, навчання мов, побудови лінгвістичних моделей. Особливо у випадку мов, що перебувають під загрозою зникнення, електронні лексикографічні ресурси стають ключовими для забезпечення їх життєздатності в цифровому середовищі.

Процес створення таких ресурсів є багаторівневим і потребує як технічних навичок, так і глибокого

розуміння мовної структури. На першому етапі вирішальним є питання джерел лексичних даних. У випадку кримськотатарської мови одним з основних джерел стали паперові словники, видані в різні періоди XX-XXI століть, включно з регіональними діалектними збірками, зокрема з румунської Добруджі. Такі матеріали мають значну культурну та мовознавчу цінність, але вони часто існують лише в одиничних паперових примірниках, що ускладнює доступ до них. Тому першим кроком стало оцифрування — сканування сторінок і застосування технології оптичного розпізнавання тексту (OCR). Отримані в такий спосіб текстові версії є вихідним матеріалом для лексикографічної обробки.

Наступний важливий крок — розмітка структури словникових статей. Будь-який словник — це не просто перелік слів, а структурований масив даних, де кожен елемент має своє місце та функцію. У типовій словниковій статті виділяють: заголовкове слово (лему), частину мови, транскрипцію, переклади, приклади вживання, фразеологізми, варіанти, синоніми, антоніми, етимологічні відомості. Щоб ці дані були придатні для цифрової обробки, кожен з них повинен бути явно розпізнаний і структурований — вручну або за допомогою автоматизованих скриптів. У процесі підготовки кримськотатарських словників для цифрової платформи Lexonomy структура кожної статті була уніфікована: сформовано єдину логіку подання, що дозволило інтегрувати лексикографічні бази зі словниками, створеними в інший час і за різними принципами.

Водночас після OCR-розпізнавання текст містить чимало механічних помилок, які потрібно виявити й

усунути. Це завдання реалізується через етап нормалізації та перевірки даних. У ході цього етапу команда проєкту проводила вичитку розпізнаних словників, виправляючи типові помилки, такі як плутанина символів (наприклад, кириличне «с» замість латинського «c»), неправильні розриви слів, дублювання або пропущення діакритичних знаків.

Коли словникові дані пройшли етап очищення та структуризації, постало питання публікації. Важливо, щоби результати лексикографічної праці не залишалися у вигляді локальних файлів чи експортів у Excel. Тому для публікації було обрано платформу Lexonomy — відкриту систему для створення, редагування та пошуку в електронних словниках. Lexonomy дозволяє працювати зі словниками прямо в браузері, підтримує імпорт і експорт даних у форматах XML та JSON, має гнучкий інтерфейс для користувача та можливість оформлення словникових статей відповідно до потреб конкретної мови. Для кримськотатарської було розгорнуто кілька пілотних версій словника, зокрема добруджанського варіанту, з відображенням значень румунською, кримськотатарською та українською мовами. Така тримовна модель дозволяє одночасно зберігати оригінальну структуру джерела та забезпечувати зручний доступ для користувачів з різними мовними вподобаннями.

Окрему увагу команда приділила впровадженню механізму зворотного зв’язку. Для цього в інтерфейсі Lexonomy передбачено форму для коментарів, за допомогою якої користувачі можуть повідомляти про помилки, подавати альтернативні варіанти перекладу або ділитися контекстами вживання слова. Цей

двосторонній підхід перетворює словник на живий інструмент, який постійно вдосконалюється завдяки залученню носіїв мови, вчителів, студентів, перекладачів. Такий підхід виявився особливо корисним під час опрацювання діалектних і регіональних слів: часто саме зворотний зв’язок дозволяв уточнити семантику чи сучасне вживання маловідомих лексем.

Значну роль у лексикографічному процесі відіграє текстовий корпус — зібрання автентичних текстів, розмічених і структурованих для лінгвістичного аналізу. Корпус дозволяє додавати до словника реальні приклади вживання слів у контексті, що значно підвищує якість словникових статей. Наприклад, замість абстрактного перекладу слово можна супроводити уривком з художнього тексту, публіцистики або інтерв’ю. Це допомагає краще зрозуміти стилістичні й прагматичні відтінки значення, а також виявити типові словосполучення. У проєкті Національного корпусу кримськотатарської мови саме корпус, зібраний і оброблений під час попередніх етапів, став основою для наповнення словника прикладами та для виявлення нових слів, відсутніх у наявних паперових джерелах.

У перспективі корпусна підтримка відкриває шлях до напівавтоматизованого створення словників. Інструменти, як-от Sketch Engine, дозволяють будувати частотні словники, виявляти колокації, автоматично визначати частину мови, знаходити контексти вживання і навіть прогнозувати лексичні зв’язки між словами. У випадку кримськотатарської мови такий підхід вже використовується для створення майбутніх ресурсів — спеціалізованих

тематичних словників (наприклад, для освіти, туризму, права) та інтеграції у системи машинного перекладу.

Отже, лексикографічна робота в умовах цифрової епохи не обмежується скануванням і викладенням словника в PDF. Це комплексний процес, що передбачає створення структурованої бази, її очищення, розмітку, публікацію, підтримку інтерфейсу взаємодії з користувачем і постійне збагачення живою мовою, представленою в корпусі. Такий підхід дозволяє не лише зберігати словникову спадщину, а й активно розвивати мову в нових сферах — освітній, науковій, цифровій. У випадку кримськотатарської мови — це не просто лексикографія, це форма мовної дії, яка дає змогу мові звучати та працювати в цифровому просторі нарівні з іншими мовами світу.

5. Мовні технології та цифрові сервіси

5.1 Машинний переклад і мовні інструменти

У сучасному цифровому світі, де спілкування, навчання й комунікація дедалі більше переходять у віртуальний простір, наявність мовних технологій — це не розкіш, а необхідність. Для мов, що перебувають у загрозливому становищі, таких як кримськотатарська, інтеграція у сферу цифрових сервісів є важливою умовою їхнього виживання. Однією з ключових цілей мовного відродження є не лише збереження пам'яті про мову, а й забезпечення її активного функціонування у нових контекстах — зокрема, в онлайн-перекладачах, мобільних додатках, освітніх платформах, голосових асистентах. Саме тому створення цифрових мовних інструментів — таких як машинний переклад, спел-чекери, автодоповнення, синтез і розпізнавання мовлення — стало важливим фокусом кримськотатарських цифрових проєктів.

Почати варто з найбільш помітного для користувачів інструменту — машинного перекладу. Його запуск потребує не лише політичної волі, а передусім наявності великих двомовних корпусів — тобто текстів, які паралельно існують двома мовами. Алгоритми сучасного машинного перекладу, особливо нейромережеві моделі, навчаються саме на таких даних, вивчаючи закономірності відповідності між словами, фразами та структурами речень. У випадку кримськотатарської мови систематичний збір і підготовка таких корпусів були здійснені командою Національного корпусу кримськотатарської мови у співпраці з Міністерством цифрової трансформації

України. Це дозволило розпочати діалог із компанією Google щодо включення мови до системи Google Translate — і за підтримки урядових структур, зокрема Мінцифри та Офісу Президента, цей крок був реалізований. У червні 2024 року кримськотатарська стала однією зі 110 нових мов, які з'явилися у Google Translate.

Поява кримськотатарської в Google Translate — реальна зміна мовного ландшафту: відтепер мова отримала шанс функціонувати у сфері міжкультурної комунікації, в освіті, міжнародних зв'язках, а також у повсякденному житті — наприклад, для туристів, переселенців або журналістів. Проте це лише початок. Щоб якість перекладу поліпшувалась, потрібне **постійне розширення корпусу**, включення нових тем, жанрів, стилів — від фольклору до інструкцій користувача. Також бажано створити **систему постредагування** перекладів носіями мови, аби враховувати стилістичні та граматичні особливості, не помічені моделлю.

Другим напрямом, що базується на наявності корпусу та словника, є **розробка інструментів автоматичної перевірки правопису (спел-чекерів)**. Їх призначення — виявлення помилок у словах, пропозиція правильних форм, іноді — корекція у реальному часі. У випадку кримськотатарської мови такий інструмент можна створити на основі **частотного словника**, отриманого з корпусу, доповненого інформацією про леми, словоформи та граматичні особливості. У проєкті НККМ вже було сформовано понад 50 мільйонів словоформ, що дає змогу перейти до етапу створення прототипів перевірки орфографії. Багато мовних

спільнот використовують для цього платформу **Hunspell** (відомий рушій для перевірки правопису в LibreOffice, Firefox, Chrome), або open-source бібліотеки на основі Python, які легко інтегруються в вебдодатки та мобільні системи. Наприклад, у Грузії, Вірменії, Уельсі такі інструменти стали невіддільною частиною освітніх платформ.

Ще одним пов'язаним інструментом є **автодоповнення** — функція, яка прогнозує наступне слово або завершує поточне. Її застосовують у текстових редакторах, месенджерах, інтерфейсах смартфонів. Для цього потрібна **мовна модель**, навчена на великих корпусах, що знає не лише словоформи, а й частоти їх сполучення. Наприклад, якщо користувач вводить «yaş», система може запропонувати «yaşlı» (старий) або «yaşamaq» (жити), залежно від контексту. Розвиток такої системи можливий на базі вже зібраного корпусу, за умови формування граматичних тегів і статистики частот.

Окремим викликом для малих мов є **розпізнавання та синтез мовлення** — тобто перетворення усного мовлення в текст і навпаки. Ці технології стали звичними для користувачів Google Assistant, Siri, Alexa — але за ними стоїть колосальна інфраструктура: тисячі годин записів, транскрипцій, моделей. Для підтримки малих мов, таких як кримськотатарська, велику роль відіграють ініціативи відкритого коду — зокрема **Mozilla Common Voice**. Це глобальний проєкт, який дозволяє будь-якій спільноті зібрати й опублікувати власний аудіокорпус для подальшого використання у TTS (text-to-speech) і ASR (automatic

speech recognition). Наразі у Common Voice вже представлені понад 100 мов, зокрема баскська, тувинська, абхазька. Для участі достатньо зібрати записи читання стандартних фраз (близько 5000 речень), дотримуючись балансу за статтю, віком, вимовою. У проєкті кримськотатарської мови вже зроблено перші кроки до створення аудіокорпусу: записано кілька десятків годин читаного тексту, підготовлено інструкції для волонтерів, сформовано базовий набір фраз. У перспективі ці дані можуть бути використані для голосових інтерфейсів, автоматичного дублювання відео, освітніх застосунків для вивчення мови.

Важливим аспектом мовної цифровізації є **локалізація інтерфейсів** — переклад користувацьких меню, налаштувань, повідомлень у програмному забезпеченні. Цей напрям дозволяє не лише зробити цифрові сервіси доступними, а й **закріпити мову в повсякденному вжитку**. Переклад інтерфейсів телефонів, комп'ютерів, банківських додатків, сервісів освіти й охорони здоров'я значно підвищує статус мови та забезпечує її функціонування у сферах, які раніше були недоступними. Багато міжнародних компаній (Google, Microsoft, Signal) використовують **платформи колективної локалізації**, такі як **Crowdin**, **Transifex**, **Localization Lab**, де будь-який користувач може приєднатись до перекладацької команди, запропонувати варіант і пройти валідацію. У проєкті кримськотатарської мови вже були ініційовані локалізаційні ініціативи: перекладено кілька open-source додатків, створено термінологічні глосарії, протестовано локалізовані версії інтерфейсів. Це

напрям, що потребує постійної підтримки, але має дуже швидкий практичний ефект — адже після запуску інтерфейсу тисячі користувачів можуть бачити свою мову щодня на екрані.

Усі ці інструменти — від машинного перекладу до автодоповнення — функціонують не окремо, а як **екосистема**, що базується на корпусі, словниках, розмічених текстах, аудіозаписах. Саме тому першочерговим завданням є **створення та підтримка якісного мовного ресурсу** — текстового та аудіо корпусу, лексикографічної бази, морфологічного аналізатора. Без цього не запрацює жоден сервіс. Але щойно така основа створена, відкриваються нові горизонти: ігрові застосунки для дітей, голосові помічники, автоматичні субтитри, системи пошуку, перекладу, навчання — все це стає можливим для будь-якої мови, незалежно від її кількості носіїв чи рівня офіційного визнання.

У випадку кримськотатарської мови ми вже бачимо перші успіхи: запуск у Google Translate, створення словникових модулів, перші тести аудіоаналізу. Попереду — багато технічної та редакторської роботи, але головне — є **фундамент і команда**, яка здатна рухатися далі. Усе це свідчить: сучасні мовні технології — не лише справа великих компаній чи держав, а також сфера, до якої можуть долучитися активісти, волонтери, викладачі, студенти. І кожен новий запис, кожна нова розмічена стаття, кожен перекладений інтерфейс — це ще один крок до того, щоб мова зазвучала на повну силу в цифрову епоху.

5.2 Інші мови як приклад

Досвід європейських регіональних мов демонструє, що навіть мови з обмеженою чисельністю носіїв можуть досягти високого рівня цифрової представленості, якщо спільнота та держава інвестують у технологічні ресурси. За результатами доповіді META-NET, понад 20 європейських мов опинилися у групі тих, що мають «слабку» або навіть «відсутню» підтримку мовних технологій — тобто великі ризики цифрової відсутності або занепаду. Утім, винятковою позицією серед цих мов є баскська, каталонська, галісійська та валлійська: вони демонструють помітні успіхи у створенні мовно-технологічної інфраструктури та надалі розвиваються як цифрово-достатні мови.

Баскська та каталонська мови отримали технологічну підтримку завдяки активному залученню дослідницьких груп, освітніх інституцій і спільнот. Наприклад, іспанське товариство з обробки природної мови SEPLN координує розробку та впровадження моделей NLP для баскської, каталонської, галісійської та іспанської, включаючи створення корпусів, POS-тегінг, машинний переклад, розпізнавання мови й інтерфейси національних значень. Корпус сучасної валлійської мови (CorCenCC) охоплює близько 11 млн слів, що включають писемні, усні та електронні жанри, з балансом за віком, гендером, регіоном і контекстом. Додано інструменти для аналізу: пошук, колокації, N-грами, частотні списки й освітній модуль Y Tiwtiadur для учнів і викладачів валлійської.

Галісійська мова стала об’єктом розробки великих мовних моделей у заходах Open Source: дослідники створили генеративні LLM-моделі, адаптовані до галісійської, на основі вже наявних архітектур, що дозволило працювати з лінгвістично мало забезпеченою мовою у форматі інструкцій і чатів (fine-tuned LLaMA-7B і Alpaca-галисійський дата-сет).

Це дозволяє зробити висновок: **наявність корпусів, словників і технологічних інструментів створює ланцюгову реакцію**, яка стимулює створення нових даних, сервісів, застосунків і користувацького контенту у мові. Соціологи називають це «цифровим циклом розвитку»: коли зростає цифрова інфраструктура мови — зростає її використання, що породжує нові мовні дані, які живлять алгоритми та продукти.

Навпаки, мови без технологій потрапляють у **порочне коло цифрового занепаду**: відсутність сервісів спричиняє низьке використання мови онлайн, що унеможливлює збір корпусів, словників, локалізації — і таким чином поглиблення технологічного дефіциту.

У доповіді Digital Language Diversity Project (DLDP) наголошується: бути цифрово-активною мовою означає не лише мати вебсайт або соцмережу, а й **використовувати власну мову у документах, медіа, поштових службах, месенджерах**. Результати опитування серед спікерів баскської, бретонської, карельської та сардинської мов показують: більшість користувачів бажають користуватися мовою електронно, але зіштовхуються з технічними

бар’єрами, браком контенту, страхом бути незрозумілими або незахищеними. Це підкреслює необхідність системного цифрового планування за участю спільноти.

Успішні мовні приклади демонструють конкретні інструменти: для баскської мови розроблено QA-датасет EuSQuAD — автоматично перекладений і синхронізований SQuAD2.0 із понад 142 000 питань-відповідей, що слугує основою для моделей обробки природної мови і запитів у рідній мові. У випадку валлійської мови CorCenCC надає не лише корпус, а й інструментарій анотації, POS-тегування, семантичного маркування, а також інтегрований освітній сервіс для вивчення трендових лінгвістичних шаблонів.

Отже, урок, який можна винести для кримськотатарських цифрових ініціатив: навіть мова з обмеженими ресурсами здатна стати технологічно представлена на рівні з меншими регіональними мовами ЄС. Ключовим чинником є **комбінована стратегія**: створення корпусів, словників, частотних та паралельних корпусів; розвиток спільноти, залучення мовознавців, активістів, студентів; побудова технологічного стека — від локалізації інтерфейсів до QA-систем та генеративних мовних моделей. При цьому важливо документувати процеси, стандартизувати підходи та передбачити механізми зворотного зв’язку. Саме такий підхід створить умови для того, щоб кримськотатарська мова неабияк розвивалася електронно, подібно до успішних

прикладів: баскської, каталонської, галісійської чи валлійської.

6. Організація роботи спільноти та управління проєктом

Цифрове відродження мови — це не лише питання технологій чи інфраструктури, а насамперед питання людей. Жоден лексикографічний модуль, жодна платформа корпусу чи мобільний застосунок не з'являються самі собою. За кожною успішною мовною ініціативою стоїть команда: організована, натхненна, методична. Саме тому ефективна організація роботи спільноти, управління процесами та стале залучення учасників — це фундамент цифрових проєктів для мов, що потребують підтримки.

На прикладі кримськотатарського цифрового проєкту, а також відповідно до рекомендацій Digital Language Diversity Project (DLDP), можна виокремити кілька стратегічних етапів, що забезпечують успішну реалізацію подібних ініціатив.

6.1. Діагностика цифрової життєздатності мови

Першим і часто найменш помітним етапом є початкове оцінювання — наскільки мова вже присутня в цифровому просторі, які інструменти вже існують, що треба створити або вдосконалити. Процес діагностики включає аналіз таких параметрів: наявність стандарту Unicode для письма, підтримка мовної розкладки клавіатури, присутність мови у Вікіпедії, Google Translate, спел-чекерах, соціальних мережах, освітніх платформах, а також оцінка того, чи існує корпус, словники, аудіоресурси, мобільні застосунки. За методикою DLDP пропонується використовувати шкалу цифрового розвитку мови, яка

включає рівні: передцифровий (або «нулевий»), дрімаючий, базовий, розвинутий, активний. У рамках проєкту НККМ (Національний корпус кримськотатарської мови) така діагностика проводилася з урахуванням близько 30 параметрів, що дозволило виявити як сильні сторони (наявність письма, активність у соцмережах, існування кількох словників), так і критичні прогалини (відсутність синтезу мовлення, слабка присутність у системах локалізації, відсутність онлайн-граматик).

6.2. Формування міждисциплінарної команди

Наступним кроком є створення команди, яка поєднує різні компетенції: мовознавчу, технічну, проєктну, редакторську, комунікаційну. У кримськотатарському кейсі до основної команди увійшли носії мови, філологи, викладачі, редактори, IT-спеціалісти, студенти, волонтери з технічним бекграундом. У пікові моменти до роботи над корпусом було залучено понад 30 осіб — кожен із яких мав чітко визначену роль: від оцифрування й перевірки OCR до форматування метаданих, технічного супроводу сайту та адміністрування платформи Lexonomy. Для уникнення дублювання обов’язків та розпорошення відповідальності було створено просту таблицю з поділом завдань, інструментів і контактних осіб. Командна взаємодія здійснювалася через Google Workspace, Slack, Trello та локальні канали зв’язку.

6.3. Підготовка та збір даних

Один з найбільш ресурсозатратних етапів — це пошук, збір, сканування й обробка текстових матеріалів. У цифрових мовних проєктах важливо не лише «дістати» потрібний документ, а й обробити його відповідно до технічних вимог: сканування з роздільною здатністю не менше ніж 300 dpi (а для фрагментів зі шрифтами з діакритикою — 600 dpi), збереження у форматах TIFF або PDF-A, попереднє очищення сканів від шуму, згортання сторінок, рамок тощо. Всі документи було збережено з уніфікованими назвами (наприклад, `<автор>__<назва>--<рік>.pdf`), що дозволило зручно структурувати архів. Одночасно створювались резервні копії на декількох хмарних сховищах, із поділом доступів за рівнем — це убезпечило від втрат у разі технічних збоїв. На цьому етапі активну роль відігравали як волонтери, які приносили приватні архіви, так і освітні партнери, зокрема викладачі, що допомагали верифікувати автентичність джерел.

6.4. Оцифрування, структурна обробка та уніфікація форматів

Зібрані тексти проходили кілька хвиль обробки. Спочатку — OCR-розпізнавання (за допомогою ABBYY FineReader, іноді Tesseract), потім — ручна перевірка, редагування, стандартизація правопису. Всі тексти переводились у формат .txt з єдиним кодуванням (UTF-8), без прихованого форматування. Особливу увагу приділяли структурі: поділу на абзаци, речення, збереженню пунктуації та маркерів

мовлення. Паралельно формувалась база метаданих: жанр, дата публікації, автор, джерело, регіон походження, оригінальний шрифт, примітки. Ці дані збирались у табличному вигляді (Google Sheets), а потім експортувались для завантаження у корпусну систему. Відповідно до сучасних стандартів, використовувались полегшені формати XML або TSV.

6.5. Створення мовних продуктів і платформ

На базі очищених і структурованих текстів створювалися лінгвістичні продукти: Національний корпус (на Sketch Engine), електронні словники (на Lexonomy), навчальні модулі, прототипи машинного перекладу. Також був запущений окремий вебсайт, що агрегував ресурси, забезпечував доступ до словника та мав форму зворотного зв'язку. Для інтеграції з системами машинного перекладу було підготовлено паралельні корпуси (тексти з відповідниками українською, англійською, турецькою), що уможливило запуск кримськотатарської мови в Google Translate у 2024 році. Крім того, підготовлено попередні модулі для спел-чекера та основи для синтезу мовлення.

6.6. Популяризація, освітня робота та залучення громади

Успіх будь-якого мовного ресурсу залежить не лише від якості контенту, а й від того, наскільки активно ним користуються. Саме тому наступним стратегічним кроком стала популяризація результатів. Команда організовувала майстер-класи, вебінари, презентації,

онлайн-воркшопи для вчителів, студентів, активістів. Було створено серію навчальних відеоінструкцій (зокрема — як користуватись корпусом, як додавати статті до Вікіпедії, як перевіряти розпізнані тексти). Проводились конкурси на найкращу статтю у Вікіпедії кримськотатарською, марафони перекладу інтерфейсів, флешмоби у соцмережах. Уся ця діяльність дозволила сформувати **живу мережу прихильників мови**, які не просто споживають ресурс, а беруть участь у його розвитку.

6.7. Стійкість, розвиток і управління ресурсами

Останній, але без перебільшення вирішальний етап — підтримка й розвиток. Тут ідеться не лише про технічні оновлення або бекапи. Перш за все — про управління ритмом роботи, уникнення вигорання, планування реалістичних, досяжних цілей. Команда свідомо працювала малими ітераціями: один документ — одне завдання — один продукт. Такий підхід дозволив уникати перевантаження, мотивувати учасників і створювати постійне відчуття прогресу. Також важливою була зовнішня комунікація — регулярні звіти, публікації, участь у міжнародних ініціативах (наприклад, Google i18n, Mozilla Common Voice), пошук партнерств та фінансування.

Водночас команда проєкту зробила висновок: для збереження стійкості цифрової ініціативи потрібно не лише утримувати ресурс, а й **вибудовувати інституційну рамку**. Саме тому було створено Національний корпус кримськотатарської мови як окрему публічну платформу з відкритим кодом і

документацією, а також розпочато процес створення асоційованої мережі партнерів серед бібліотек, освітніх установ, архівів, неурядових організацій. Це створює підґрунтя для наступних етапів — інтеграції до систем освіти, формування державної підтримки, розширення технологічного стека мови.

У підсумку: організація цифрового мовного проєкту — це передусім організація довіри. Довіри до того, що мова має значення. Довіри до спільноти, яка її розвиває. Довіри до того, що навіть обмеженими ресурсами, але системно, прозоро й натхненно — можна створити те, що залишиться. І якщо кожен учасник проєкту вірить у це — навіть невелика команда може здійснити прорив, який змінить статус мови в цифровому світі.

7. Висновки та рекомендації

Цифрове середовище сьогодні одночасно є викликом і можливістю для мов, які перебувають у тендітному стані. З одного боку, більшість регіональних і корінних мов багато в чому залежать від домінування великих мов у мережі — зокрема англійської, китайської, іспанської та французької. Дослідження META-NET 2012 року засвідчили: понад **21 європейська мова** має технічну підтримку, що оцінюється експертами як «мала» або «відсутня» — включно з áreа машинного перекладу, синтезу мовлення, аналізу тексту або корпусів. Цей висновок підтвердився й у проєкті European Language Equality (ELE), що у 2023 році знову переконав: лише незначне число мов має належний рівень цифрової готовності — решта все ще знаходяться на периферії мовних технологій.

Водночас досвід успішних мов — таких як баскська, каталонська, галісійська, валлійська — демонструє протилежний сценарій. Ці мови, попри відносно невелику кількість носіїв, досягли серйозної технологічної підтримки, мають повноцінні корпуси, онлайн-словники, локалізаційні системи та навіть генеративні модельні рішення. Цей шлях доводить: **настав час здійснити стрибок у цифрову сучасність**, і зробити це під силу навіть малим мовним спільнотам — за наявності стратегії, ресурсів і співпраці.

Кейс кримськотатарської мови є вдалим прикладом такої трансформації. Упродовж кількох років спільнотою було сформовано текстовий і паралельний корпус, розроблено електронний

словник, адаптовано OCR-інструменти, підготовлено підґрунтя для машинного перекладу. Завдяки цьому кримськотатарська мова у липні 2024 року стала однією з нових мов в Google Translate, що стало одним із перших значних успіхів української мовної політики в цифровій сфері. Це свідчить: навіть небезпечно загрожена мова зможе швидко набрати цифрову життєздатність — за умови узгодженої, комплексної роботи команди та підтримки держави.

На підставі озвученого, можна сформулювати низку рекомендацій для мовних спільнот і державних структур.

Перш за все — **усі цифрові ініціативи треба починати з основ**. Вкрай важливо впевнитися, що мова:

1. повністю підтримується в Unicode,
2. існують розкладки клавіатури,
3. правопис стандартизовано або узятий домовлений орфографічний стандарт. Якщо такі речі не врегульовані — саме на цьому етапі слід розв'язати ці питання, ще до будь-яких технічних рішень.

Далі — **створення корпусу**. Спільнота повинна зібрати всі доступні тексти, застосувати сканування згідно з технічними стандартами (300–600 dpi), провести OCR-розпізнавання і ручну корекцію. Важливо супроводжувати тексти метаданими (жанр, автор, дата, джерело) і анотаціями (POS-теги, леми, граматичні ознаки). Це фундамент, без якого неможливо побудувати подальші інструменти.

Третій крок — **легалізація і розвиток словників**. Оцифруйте наявні паперові словники, уніфікуйте структуру статей, нормалізуйте орфографію і розширюйте словникові статті прикладами з корпусу. Використайте онлайн-платформи, такі як Lexonomy, щоб забезпечити доступність та можливість зворотного зв'язку від користувачів.

Наступним важливим напрямом є **партнерство з технологічними організаціями**, університетами, IT-компаніями. Це дозволяє отримати підтримку для створення машинного перекладу, спел-чекерів, TTS/ASR систем.

Усе це неможливо без **успішної спільнотної мобілізації**. Організовуйте майстер-класи, воркшопи, презентації, конкурси, залучайте молодь до створення контенту — blog, Wikipedia, відео. Цифрова участь ефективніша за будь-яку рекламу: вона прищеплює активне користування мовою і створює фактичне цифрове життя.

Для **державних інституцій і міжнародних донорів** також є важливі напрямки. Слід підтримувати фінансово та організаційно проєкти з корпусотворення, словникової бази, технологічної інфраструктури мов. Не менш важливо створити **центри лінгвістичного різноманіття** і розробити стратегічні документи цифрової рівноправності — як це робить Європейський парламент згідно з резолюцією 2018 року щодо технологічної підтримки всіх європейських мов (Language Equality in the Digital Age). Поширення «Digital Language Survival Kit» DLDP (інструмент для самооцінки цифрової готовності і

планування розвитку) має стати невіддільною частиною таких практик.

Також доцільно створити тренінги для мовних активістів і спільнот, які навчать користуватися шкалами цифрової готовності, оцінювати потреби та порівнювати надбання — це дозволить проводити регулярні аудити та моніторинг прогресу.

У підсумку, реалізація цих кроків дає змогу не тільки **зберегти мову**, а й **піднести її до цифрового сучасного статусу**. Мови України — кримськотатарська, караїмська, урумська — мають потенціал не лише вижити, але й бути повноцінними гравцями цифрового середовища. За умови чіткої стратегії, системного планування, партнерств, активного залучення громади й підтримки держави — вони можуть стати прикладом мовного технологічного розвитку, що натхне інші спільноти та змінить мовну карту світу.

Це не лише рекомендації; це **дорожня карта до цифрової жвавості мови**, яка живе, розвивається й розширює свою сферу поширення. Якщо слідувати цій стратегії — результат не забариться.