

ОСОБЛИВОСТІ ЦИФРОВІЗАЦІЇ ЦЕРКОВНОСЛОВ'ЯНСЬКИХ ТЕКСТІВ

Семернін Віктор Вікторович

магістрант факультету інформаційних технологій
ДВНЗ «Ужгородський національний університет», Україна

Тар Віталій Вікторович

магістрант факультету інформаційних технологій
ДВНЗ «Ужгородський національний університет», Україна

Науковий керівник: Міца Олександр Володимирович

ORCID ID: 0000-0002-6958-0870

д-р. техн. наук, професор,
завідувач кафедри інформаційних управляючих систем та технологій
ДВНЗ «Ужгородський національний університет», Україна

Потреба у створенні інформаційної платформи для аналізу текстів і співів в українській редакції церковнослов'янської мови (УРЦМ), яка переважно включатиме паралельні та порівнювані текстові корпуси, а також корпуси співів, є терміновою та важливою з кількох причин:

1. УРЦМ швидко вимирає з нинішнього церковного (літургійного) вжитку.
2. Відбувається постійне придушення та планова заміна її російською редакцією церковнослов'янської представниками (предстоятелями церков і духовенством) Московського патріархату РПЦ в Україні.
3. Збереження унікальної спадщини УРЦМ у тексті та співі, яка все ще живе в церковному вжитку зараз, має вирішальне значення для майбутніх поколінь.

Платформа надає підтримку та ресурси для користувачів, які, можливо, не знайомі з мовою УРЦМ або дослідженням цифрового корпусу [1]. Це може включати посібники користувача, навчальні посібники та службу підтримки для технічної підтримки, а також багато інструментів корпусу і платформу корпусу.

Розробка платформи корпусу УРЦМ має враховувати потреби та очікування користувачів. Тому в процесі розробки важливо враховувати сторону користувача. Для цього було використано досвід інноваційних технологічних розробок науковців Ужгородського національного університету, таких як інтерактивна карта діалектів [2] та інтерактивна карта в рамках реалізації актуальних державних та громадських напрямків [3].

По-перше, платформа УРЦМ має бути зручною, враховувати різні типи користувачів (звичайних людей, які цікавляться лексикою та текстами УРЦМ, експертів середнього рівня та лінгвістів) і легкою в навігації. Розробивши власне програмне забезпечення, яке ми перевірили, створивши власний корпус УРЦМ. Дотримуючись чітких інструкцій, ми завантажили на платформу один із текстів УРЦМ у форматі PDF (відскановані зображення OCR перед завантаженням). Текст написаний мовою УРЦМ, але транслітеровано літерами сучасного українського алфавіту за фонетичним принципом («пиши як чуєш»). Єдиною відмінністю цього тексту УРЦМ від сучасного українського письма є наявність наголосу майже в кожному слові, крім службових (стоп-слів). Приклад результату роботи можна побачити на рисунку 1.

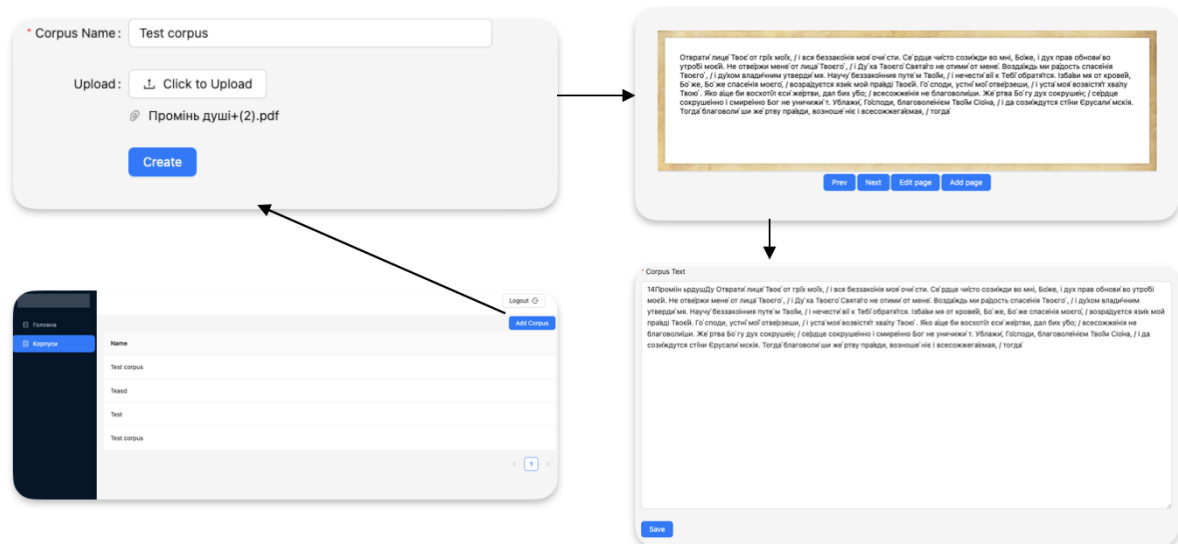


Рис. 1. Приклад роботи платформи

По-друге, платформа має надавати різноманітні тексти мовою УРЦМ, зокрема літургійні тексти, літературні твори та історичні документи. Користувачі повинні мати доступ до широкого кола текстів для цілей лінгвістичних досліджень. Ці тексти зазвичай друкуються алфавітом УРЦМ.

Враховуючи різне походження текстів УРЦМ з точки зору часу та місця публікації, стандартизований алфавіт і послідовне використання діакритичних знаків створює серйозну проблему для розвитку платформи корпусу УРЦМ. Єдиним рішенням може бути розділення текстів на окремі корпуси за їхніми історичними періодами. Цей підхід дозволить вченим і дослідникам аналізувати та порівнювати різні версії мови УРЦМ шляхом створення паралельних корпусів, які також допоможуть відстежити розвиток її мовних особливостей з часом. Тут можна також задіяти такі засоби, як штучні нейронні мережі [4] та машинне навчання [5].

Основні очікування від платформи такі:

- різні типи прийому текстових даних (УРЦМ різних періодів публікації);
- опрацювання тексту;
- семантичне тегування кожної частини корпусу;
- якісний та кількісний аналізи, які базуються на різних статистичних характеристиках;
- порівняння з різними перекладами того самого тексту та картографічних термінів різними мовами;
- тексти, які зберігатимуться та аналізуватимуться в корпусі, повинні вибиратися лише лінгвістами для побудови належних залежностей і ведення належної статистики для запобігання побічним ефектам у статистиці;
- лінгвісти можуть вибрати належні методи обчислення попередньої обробки та аналізу тексту, і ці методи повинні бути налаштованими;
- прийом звукових даних і створення корпусу співу.

Висновки. Підсумовуючи викладене можна стверджувати, – що створення платформи для аналізу текстів і співів мови УРЦМ є критично важливим завданням. Ця платформа надає можливість працювати з різноманітними текстами, забезпечувати можливості розділення текстів та задовольняти потреби різних типів користувачів, включаючи лінгвістів і дослідників. Збереження цієї унікальної мови та

спадщини УРЦМ у текстах і співах є важливим завданням для майбутніх поколінь та культурного збагачення.

Список використаних джерел:

1. Popovych, N., Lutskiv, A., Mitsa, O., Lyntvar, O., & Ivanova, A. (2023). Ukrainian Redaction of Church Slavonic (URCS): Needs for Digitalization and Text Corpora Platform Generation. Part I.
2. Mitsa, O., Sharkan, V., Maksymchuk, V., Varha, S., & Shkurko, H. (2023). Ethnocultural, Educational and Scientific Potential of the Interactive Dialects Map. In 2023 IEEE International Conference on Smart Information Systems and Technologies (SIST) (pp. 226-231). Astana, Kazakhstan. doi: 10.1109/SIST58284.2023.10223544.
3. Lupei, M., Shlahta, M., Mitsa, O., Horoshko, Y., Tsybko, H., & Gorbachuk, V. (2022, September). Development of an Interactive Map Within the Implementation of Actual State and Public Directions. In 2022 12th International Conference on Advanced Computer Information Technologies (ACIT) (pp. 384-387). IEEE. doi: 10.1109/ACIT54803.2022.9913191.
4. Lupei, M., Mitsa, A., Povkhan, I., & Sharkan, V. (2020, August). Determining the eligibility of candidates for a vacancy using artificial neural networks. In 2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP) (pp. 18-22). IEEE. doi: 10.1109/DSMP47368.2020.9204020.
5. Lupei, M., Mitsa, O., Sharkan, V., Vargha, S., Lupei, N. (2023). Analyzing Ukrainian Media Texts by Means of Support Vector Machines: Aspects of Language and Copyright. In: Hu, Z., Dychka, I., He, M. (eds) *Advances in Computer Science for Engineering and Education VI. ICCSEEA 2023. Lecture Notes on Data Engineering and Communications Technologies*, vol 181. Springer, Cham. doi: 10.1007/978-3-031-36118-0_16.