UDK: 004.65, 8, 81’2, 81’28

XORAZM SHEVASIDAGI MATNLARNI O‘ZBEK ADABIY TILIGA O‘GIRISH MASHINALI TARJIMA MODELINI QURISH UCHUN PARALLEL KORPUS YARATISH

Sharipov Maqsud Siddiqovich,
dotsent
maqsbek72@gmail.com
Urganch davlat universiteti

Kurbanova Lola Ulug‘bek qizi,
o‘qituvchi
lolakompyuterlingvistikasi@gmail.com
Urganch davlat universiteti

Annotatsiya. Ushbu maqolada Xorazm shevasidagi matnlarni o‘zbek adabiy tiliga mashinali tarjima qilish imkonini beruvchi tizimni qurish uchun zarur bo‘lgan **parallel korpus yaratish** masalasi yoritilgan. Dialektal tilni standart tilga moslashtirishda dastlabki va muhim bosqich sifatida sheva-adabiy til juftliklaridan iborat parallel korpus tuzildi. Korpusni shakllantirishda Xorazm shevasida yozilgan ijtimoiy tarmoq matnlari, og‘zaki xalq ijodi namunalarining matnga ko‘chirilgan variantlari, hamda hududiy OAV materiallari asos sifatida tanlandi. Har bir shevada berilgan gapga mos holda o‘zbek adabiy tilida qo‘lda tarjima tayyorlandi. Korpus XML formatida belgilandi va u mashinali tarjima modellarini kelgusida o‘rgatish uchun tayyorlangan ma’lumotlar bazasini tashkil etadi. Tadqiqot natijasi sifatida mashinaviy tarjima va dialektologik izlanishlar uchun ochiq parallel resurs yaratildi.

Abstract. This article discusses the creation of a **parallel corpus** necessary for developing a system capable of performing machine translation from the Khorezm dialect into the Uzbek literary language**.** As an initial and essential stage in adapting dialectal language to the standard form, a parallel corpus consisting of dialect–standard sentence pairs was compiled. In the process of forming the corpus, texts written in the Khorezm dialect were collected from social media**,** transcribed samples of oral folk literature, and regional media sources. Each dialectal sentence was manually translated into its equivalent in the literary Uzbek language. The corpus was annotated in XML format and serves as a structured database prepared for training machine translation models in the future. As a result of the study, an open-access parallel resource was created to support further research in machine translation and dialectology.

Аннотация. В данной статье рассматривается вопрос создания **параллельного корпуса**, необходимого для построения системы машинного

перевода, переводящей тексты на хорезмском диалекте на узбекский литературный язык. На первом и важном этапе адаптации диалектной речи к стандартному языку был сформирован параллельный корпус, состоящий из пар предложений на диалекте и их ручных переводов на литературный узбекский язык. При формировании корпуса в качестве основного источника использовались тексты на хорезмском диалекте из социальных сетей**,** записанные версии устного народного творчества, а также материалы региональных СМИ. Каждому диалектному предложению был сопоставлен эквивалент на литературном языке, подготовленный вручную. Корпус был размечен в формате XML и представляет собой базу данных, предназначенную для последующего обучения моделей машинного перевода. В результате исследования был создан открытый параллельный ресурс, предназначенный для задач машинного перевода и диалектологических исследований.

Kalit so‘zlar: *Xorazm shevasi, o‘zbek adabiy tili, parallel korpus, mashinali tarjima, dialektologiya, NLP.*

Zamonaviy kompyuter lingvistikasi va tabiiy tilni qayta ishlash (NLP) sohasining jadal rivojlanishi natijasida mashinali tarjima texnologiyalari dunyo tillarining turli variantlari, shevalari va lahjalarini avtomatik tarzda standart tilga konvertatsiya qilish imkonini bermoqda. O‘zbek tilida ham hududiy shevalarning ko‘pligi, ularning leksik, fonetik va grammatik xususiyatlari tilshunoslar, tarjimonlar va NLP mutaxassislari uchun yangi ilmiy vazifalarni yuzaga keltirmoqda.

O‘zbekiston hududida keng tarqalgan Xorazm shevasi o‘zining boy iboralari, o‘ziga xos talaffuz va uslubiy vositalari bilan ajralib turadi. Ayni paytda, rasmiy yozishmalar, ta’lim tizimi va ommaviy axborot vositalarida adabiy tilning ustuvorligi sababli, Xorazm shevasida yaratilgan matnlarning adabiy tilga avtomatik tarjima qilinishi dolzarb masalaga aylanmoqda. Biroq, ushbu jarayonni amalga oshirish uchun eng asosiy shartlardan biri bu – **sheva-adabiy til juftliklaridan iborat parallel korpus mavjud bo‘lishidir.**

Ushbu maqolada Xorazm shevasidagi matnlarni o‘zbek adabiy tiliga tarjima qilishga mo‘ljallangan mashinali tarjima modelini qurish uchun zarur bo‘lgan **parallel korpusni yaratish bosqichlari,** manbalar tanlovi, matnlarni normalizatsiya qilish, XML formatida belgilash, va kelgusidagi tadqiqotlar uchun yaroqli ma’lumotlar bazasini shakllantirish masalalari keng yoritiladi. Korpusda aks ettirilgan matnlar turli janrlarga oid bo‘lib, ijtimoiy tarmoqlar, og‘zaki ijodiy materiallar va mahalliy nashrlardan olingan. Har bir sheva gapiga mos tarzda adabiy tarjima qo‘lda tuzilgan bo‘lib, bu o‘zbek tilida ilk marta dialektal parallel korpus yaratish yo‘lidagi muhim qadamlardan biridir. [2].

Tabiiy tilni qayta ishlash (NLP) va mashinali tarjima sohalarida dunyoda keng ko‘lamli tadqiqotlar olib borilmoqda. Ayniqsa, shevalar va lahjalardan adabiy tilga tarjima qilishga doir tadqiqotlar oxirgi yillarda dolzarb mavzulardan biriga aylangan. Arab, xitoy, hind va nemis tillarida bu borada keng qamrovli korpuslar yaratilgan va maxsus mashinali tarjima modellarida testdan o‘tkazilgan. Biroq, o‘zbek tilida dialektologik yondashuv asosidagi NLP ishlanmalari hanuz yetarli emas.

E. Kuriyozov, U. Salaev, va S. Matlatipov[3] tomonidan olib borilgan boshqa bir tadqiqotda o‘zbek tilidagi **matnlar asosida tasniflash datasetlari, nomuhim so‘zlar to‘plamlari, lemmatizatsiya modullari,** va **sentiment tahlil korpuslari** ishlab chiqilgan. Ular ko‘pincha umumiy adabiy tilga asoslangan bo‘lib, dialektal variantlar, xususan Xorazm shevasi uchun maxsus parallel korpus yaratilmagan. [1]

Shuningdek, **Madatov, Bekchanov va Vičič** [4] o‘z tadqiqotlarida o‘zbek tilida stop-so‘zlar va ularni aniqlash bo‘yicha statistik tahlillarni taqdim etishgan. Biroq, bu ishlar asosan yozma adabiy tilga asoslangan bo‘lib, shevalar, og‘zaki nutq va mintaqaviy lahjalarni qamrab olmaydi.

Xalqaro tajribaga e’tibor qaratilsa, arab tilining misrlik, livanlik va boshqa lahjalari uchun yaratilgan parallel korpuslar yordamida **sheva–standart til mashinali tarjimasi** muvaffaqiyatli amalga oshirilmoqda. Bu modeldan ilhomlanib, o‘zbek tilining Xorazm shevasini o‘rganish va tarjima qilishda ham xuddi shunday parallel juftliklar asosida korpus yaratish samarali yechim sifatida ko‘rilmoqda.

Xulosa qilib aytganda, mavjud adabiyotlar o‘zbek tili uchun korpuslar, morfologik va sintaktik modellar, lemmatizatsiya va transliteratsiya tizimlari borasida asos yaratib bergan bo‘lsa-da, **dialektal matnlar asosida parallel korpus tuzish** va ularni mashinali tarjima uchun tayyorlash borasida hali ilmiy bo‘shliq mavjud. Mazkur tadqiqot aynan shu bo‘shliqni to‘ldirishga qaratilgan bo‘lib, Xorazm shevasidagi matnlarni o‘zbek adabiy tiliga o‘girishda birinchi parallel korpus namunasi bo‘lib xizmat qiladi.

Ushbu tadqiqotning asosiy metodologik yondashuvi – Xorazm shevasidagi matnlarni o‘zbek adabiy tiliga mos tarzda juftlab, mashinali tarjima modellarini qurish uchun tayyor parallel korpus yaratishdan iborat. Tadqiqot quyidagi bosqichlarda amalga oshirildi:

Korpusni shakllantirish uchun Xorazm shevasida yozilgan turli janrdagi va shakldagi matnlar yig‘ildi. Asosiy manbalar quyidagilardan iborat:

- **Ijtimoiy tarmoqlardagi matnlar** (Telegram, Facebook postlari, izohlar);

- **Xalq og‘zaki ijodi namunalari** (ertaklar, latifalar, maqollar va hikoyalar);
- **Hududiy nashrlar va intervyular** (Xorazm viloyatidagi gazeta maqolalari va bloglar);
- **O‘zaro suhbatlardan olingan transkriptlar** (dialektal og‘zaki matnlarning yozma ko‘rinishlari).

Yig‘ilgan Xorazm shevasidagi matnlarning har bir gapiga mos tarzda o‘zbek adabiy tilidagi tarjimalar qo‘lda tayyorlandi. Tarjimalar ona tili bo‘yicha mutaxassislar va Xorazm shevasini chuqur biladigan tarjimonlar tomonidan amalga oshirildi. Har bir gaplik juftlikda semantik moslik va grammatik to‘g‘rilikni ta’minlashga alohida e’tibor qaratildi.

Matnlar va ularning tarjimalari XML formatida belgilanib, korpus holatiga keltirildi.

XML formatdan foydalanish:

- matnlarning strukturaviy tahlilini osonlashtirish;
- hujjatning uslubi, turi, manbasi, yozilgan sanasi, muallifi kabi **meta-ma’lumotlarni** saqlash;
- mashinali tahlilga tayyor shaklda saqlash imkonini beradi.

Misol uchun:

```
<CORPUS NAME="Xorazm-Uzbek">
  <TEXT ID="001" TYPE="og‘zaki" STYLE="sheva">
    <AUTHOR>Anonim</AUTHOR>
    <SOURCE>Telegram</SOURCE>
    <DATE>2024-12-01</DATE>
    <CATEGORY>Folklor</CATEGORY>
    <CONTENT TITLE="Qishloq hikoyasi">Opom galdi, bog‘ tarafni sirip qo‘ydim.</CONTENT>
    <STANDARD>Onam keldi, hovlini supurib qo‘ydim.</STANDARD>
  </TEXT>
</CORPUS>
```

Korpusdagi har bir sheva-adabiy gaplik juftlik .tsv (tab bilan ajratilgan qiymatlar) yoki .csv formatda saqlandi. Bu format mashinali tarjima modellarini (masalan, Transformer, mBART) o‘rgatishda keng qo‘llaniladi.

Format namunasi:

sheva_matni	adabiy_matn
Doyim oytdi, boroli.	Amakim aytdi, boraylik.

Ma’lumotlar ustidan qo‘lda va dasturiy tozalash ishlari olib borildi:

- noto‘g‘ri imloviy yozuvlar tuzatildi;
- orfografik va grammatik xatoliklar tuzatildi;
- takrorlanuvchi yoki kontekstdan tashqaridagi gaplar chiqarib tashlandi;
- semantik jihatdan nomuvofiq juftliklar belgilandi va ajratib qo‘yildi.

Umumiy natija sifatida, taxminan **10 000 dan ortiq gaplik parallel korpus** yaratildi. Ushbu korpus mashinali tarjima modellarini keyingi bosqichda o‘rgatish uchun tayyor holatga keltirildi. Korpus o‘zbek tilida ilk marta hududiy dialekt (Xorazm shevasi) asosida tayyorlangan parallel resurs sifatida ilmiy va amaliy ahamiyatga ega.

Ushbu tadqiqot doirasida Xorazm shevasini o‘zbek adabiy tiliga tarjima qilish imkonini beruvchi mashinali tarjima modelini yaratishga tayyorlov bosqichi sifatida **parallel korpus** shakllantirildi. Korpus yaratish jarayoni davomida to‘plangan materiallar, ularning tahlili va tuzilgan juftliklarning umumiy statistikasi quyida bayon etiladi.

Yig‘ilgan matnlar **leksik boyligi, morfologik xususiyatlari** va **sintaktik tuzilmasi** jihatidan adabiy o‘zbek tilidan ancha farq qilishi kuzatildi.

Bu xususiyatlar tarjima modelini qurishda e’tiborga olinadigan muhim lingvistik ko‘rsatkichlardir.

Yaratilgan parallel korpus quyidagi asosiy sonli ko‘rsatkichlarga ega:

Ko‘rsatkich	**Qiymat**
Jami gaplar soni	10 000+
O‘rtacha gap uzunligi	8–12 so‘z
Foydalanilgan manbalar	Telegram, Facebook, matbuot
Adabiy tarjima usuli	Qo‘lda, mutaxassislar tomonidan
Saqlash formati	XML, TSV
Belgilov formatlari	<TEXT>, <AUTHOR>, <STANDARD> va h.k.

XML formatda korpusni belgilash natijasida har bir matnga oid meta-ma’lumotlar (muallif, janr, manba, sana, sarlavha) ham aniqlashtirilib saqlandi.

Yaratilgan parallel korpus quyidagi jihatlarda muhim ahamiyat kasb etadi:

- **Mashinali tarjima modellarini o‘qitish** uchun tayyor va tozalangan resurs sifatida xizmat qiladi.
- **Dialektologik tadqiqotlar**da shevalarning morfologik va sintaktik tuzilmasini o‘rganish imkonini beradi.

- O‘zbek tilining **hududiy variantlari bo‘yicha lingvistik vositalar** yaratishga asos bo‘ladi.
- **Raqamli lingvistika va ta’lim tizimi**da og‘zaki va dialektal matnlarni standart tilga moslashtirishda qo‘llanishi mumkin.

Xulosa tarzida, tadqiqot natijasida o‘zbek tilida ilk bor Xorazm shevasidan adabiy tilga o‘girishga mo‘ljallangan **parallel korpus** yaratildi. Bu korpus nafaqat mashinali tarjima modellarini yaratish, balki dialektologiya, leksikografiya va NLP tadqiqotlari uchun keng imkoniyatlar eshigini ochadi.

Ushbu tadqiqot natijasida Xorazm shevasidagi matnlarni o‘zbek adabiy tiliga avtomatik tarzda tarjima qilish imkonini beruvchi mashinali tarjima tizimini yaratish uchun zarur bo‘lgan **parallel korpus** muvaffaqiyatli shakllantirildi. Korpus turli manbalardan yig‘ilgan dialektal matnlar va ularning qo‘lda tuzilgan adabiy tarjimalaridan iborat bo‘lib, mashinali tarjima modellarini o‘qitish uchun mustahkam poydevor yaratdi. Korpusning XML va TSV formatlarida tuzilgani uni keyingi bosqichda avtomatlashtirish va modellashtirish jarayonlarida yengillik yaratadi. Bu loyiha o‘zbek tilining hududiy variantlarini standartlashtirish va NLP yondashuvlari orqali integratsiyalashishiga xizmat qiladi.

Xorazm shevasi bilan cheklanib qolmasdan, boshqa hududiy shevalar (masalan, Farg‘ona, Qashqadaryo, Andijon) asosida ham parallel korpuslar yaratish zarur. Bu butun o‘zbek tilining dialektal xaritasini qamrab olgan mashinali tarjima tizimlarini ishlab chiqishga yo‘l ochadi.

Yaratilgan parallel korpusni asos qilib olib, mashinali tarjima modellarini (masalan, Transformer, mBART, mT5) o‘qitish va baholash bo‘yicha eksperimentlar o‘tkazish tavsiya etiladi. Bu orqali korpusning amaliy samaradorligi va NLP vositalarida qo‘llanilish darajasi aniqlanadi.

Yaratilgan korpusni ochiq manbada joylashtirib, ilmiy hamjamiyat va til texnologiyalari bilan shug‘ullanuvchi mutaxassislar uchun foydalanishga taqdim etish kerak. Bu nafaqat ilmiy almashinuvni kuchaytiradi, balki korpusning doimiy boyitilishini ham ta’minlaydi.

Foydalanilgan adabiyotlar:

1. E. Kuriyozov, U. Salaev, S. Matlatipov, and G. Matlatipov, “Text classification dataset and analysis for Uzbek language,” arXiv preprint arXiv:*2302.14494*, 2023.

2. Sharipov M.S., Kurbanova L.U., Qurbanova R.U. O‘zbek tili korpusi dasturiy ta’minotini yaratish // CTCL-2023.

3. Analysis for Uzbek language,” arXiv preprint arXiv:*2302.14494*, 2023.

4. S. Matlatipov, H. Rahimboeva, J. Rajabov, and E. Kuriyozov, “Uzbek Sentiment Analysis Based on Local Restaurant Reviews,” in *CEUR Workshop Proceedings*, 2022, pp. 126– 136. [Online]. Available: www.scopus.com

5. M. Sharipov, E. Kuriyozov, O. Yuldashev, and O. Sobirov, “UzbekTagger: The rule-based POS tagger for Uzbek language,” arXiv preprint arXiv:2301.12711, 2023.

6. K. Madatov, S. Bekchanov, and J. Vičič, “Accuracy of the Uzbek stop words detection: a case study on” School corpus»,” arXiv preprint arXiv:2209.07053, 2022.

7. K. Madatov, S. Bekchanov, and J. Vičič, “Dataset of stopwords extracted from Uzbek texts,” Data Brief, vol. 43, p. 108351, 2022.

8. X. Madatov, M. Sharipov, and S. Bekchanov, “O‘zbek tili matnlaridagi nomuhim so‘zlar,” *Computer linguistics: problems, solutions, prospects*, vol. 1, no. 1, 2021.

9. M. Sharipov and O. Sobirov, “Development of a rule-based lemmatization algorithm through Finite State Machine for Uzbek language,” arXiv preprint arXiv*:2210.16006*, 2022.

10. M. Sharipov, J. Mattiev, J. Sobirov, and R. Baltayev, “Creating a morphological and syntactic tagged corpus for the Uzbek language,” arXiv preprint arXiv:2210.15234, 2022.

11. E. Kuriyozov, S. Matlatipov, M. A. Alonso, and C. Gómez-Rodriguez, “Construction and evaluation of sentiment datasets for low-resource languages: The case of Uzbek,” in Human Language Technology. Challenges for Computer Science and Linguistics: 9th

12. Language and Technology Conference, LTC 2019, Poznan, Poland, May 17–19, 2019, Revised Selected Papers, 2022, pp. 232–243.

13. U. Salaev, E. Kuriyozov, and C. Gómez-Rodriguez, “A machine transliteration tool between Uzbek alphabets,” in CEUR Workshop Proceedings, 2022, pp. 42–50. [Online]. Available: www.scopus.com

14. M. Sharipov and U. Salaev, “Uzbek affix finite state machine for stemming,” arXivpreprint arXiv:2205.10078, 2022.

15. Tiedemann J. “Parallel Data, Tools and Interfaces in OPUS,” in Proc. of the 5 th International Workshop on Language Resources and Evaluation, 2012, pp. 221-225.

16. Goldwasser D., “Statistical Machine Translation: A Survey,” Journal of Computer Science, vol. 56, № 3, 2013, pp. 450-463.