NAMANGAN DAVLAT UNIVERSITETI
ILMIY AXBOROTNOMASI, [2026-8]
ISSN:2181-1458
ISSN:2181-0427

10.00.00-FILOLOGIYA
journal.namdu.uz
PHILOLOGICAL

UO‘K 81’322:811.512.1

TURKIGO‘Y MUMTOZ ADIBLAR ILMIY ELEKTRON KORPUSINI YARATISH: TAMOYILLAR, TUZILISH VA METODOLOGIYA

Sulaymonov Mo‘minjon Yusubjonovich
Filologiya fanlari nomzodi, NamDU dotsenti
E-mail: msulaymon62@mail.ru
ORCID: https://orcid.org/009-0008-7078-5962
Xoldorova Gulnoza Husniddinovna
Filologiya fanlari bo‘yicha falsafa doktori, PhD
E-mail: gulnozaholdarova7@gmail.com
ORCID: https://orcid.org/0009-0005-9993-0279

Annotatsiya: *Maqolada turkigo‘y mumtoz adiblar merosini raqamlashtirish va ilmiy elektron korpusda jamlashning nazariy hamda amaliy asoslari tahlil qilinadi. Tadqiqotning maqsadi XIV-XV asrlarga oid turkiy yozma yodgorliklarni qamrab oluvchi manba-tanqidiy, ko‘p yozuvli va ko‘p qatlamli korpusni yaratish tamoyillari, tuzilishi va metodologiyasini asoslashdan iborat. Ish jahon (VATEC, Turk milliy korpusi) hamda milliy (o‘zbek tili milliy korpusi, Alisher Navoiy korpusi) tajribasining qiyosiy tahliliga tayanadi. Natijada korpusning besh qatlamli arxitekturasi, annotatsiyaning token, lemma, morfologik teg, aruz va janr darajalari, shuningdek arab yozuvidagi manbalarni transliteratsiya va transkripsiya bilan parallellashtirish modeli taklif etiladi. Ilmiy yangilik sifatida mumtoz she’riy matnlar uchun aruz-vazn annotatsiyasi qatlamining kiritilishi hamda manba ishonchliligini ta’minlovchi metama’lumot tizimi ilgari suriladi. Korpusning tekstologik tadqiqot, tarixiy leksikografiya, mualliflik uslubini lingvostatistik aniqlash va lingvodidaktika uchun amaliy ahamiyati ko‘rsatiladi.*

Kalit so‘zlar: *lingvistik korpus, turkiygo‘y adiblar, korpus lingvistikasi, transliteratsiya, morfologik annotatsiya, aruz, TEI, tarixiy leksikografiya.*

СОЗДАНИЕ НАУЧНОГО ЭЛЕКТРОННОГО КОРПУСА ТЮРКОЯЗЫЧНЫХ КЛАССИЧЕСКИХ ПИСАТЕЛЕЙ: ПРИНЦИПЫ, СТРУКТУРА И МЕТОДОЛОГИЯ

Аннотация: *В статье анализируются теоретические и практические основы цифровизации наследия тюркоязычных классиков и его объединения в научном электронном корпусе. Цель исследования — обосновать принципы, структуру и методологию создания источниковедчески выверенного, многописьменного и многоуровневого корпуса, охватывающего тюркские памятники XIV-XV вв. Работа опирается на сравнительный анализ мирового (VATEC, TNC) и национального опыта. Предложена пятиуровневая архитектура корпуса с уровнями токенизации, лемматизации, морфологической разметки, аруза и жанра.*

Ключевые слова: *лингвистический корпус, тюркоязычная классика, корпусная лингвистика, транслитерация, морфологическая аннотация, аруз, TEI, историческая лексикография.*

CREATING A SCHOLARLY ELECTRONIC CORPUS OF TURKIC CLASSICAL WRITERS: PRINCIPLES, STRUCTURE AND METHODOLOGY

Abstract: *The article analyses the theoretical and practical foundations of digitising the heritage of Turkic classical writers and consolidating it in a scholarly electronic corpus. The aim of the study is to substantiate the principles, structure and methodology for building a source-critical, multi-script and multi-layered corpus covering Turkic written monuments of the XIVth–XVth centuries. The study relies on a comparative analysis of international (VATEC, the Turkish National Corpus) and national (the Uzbek National Corpus, the Alisher Navoi corpus) experience. As a result, a five-layer corpus architecture is proposed, together with annotation levels for tokens, lemmas, morphological tags, aruz and genre, and a model that aligns Arabic-script sources with their transliteration and transcription. The scientific novelty lies in introducing an aruz annotation layer for classical poetic texts and a metadata system ensuring source reliability. The corpus’s applied value for textual studies, historical lexicography, stylometric authorship analysis and language teaching is demonstrated.*

Keywords: *linguistic corpus, Turkic classical literature, corpus linguistics, transliteration, morphological annotation, aruz, TEI, historical lexicography.*

Kirish

Zamonaviy filologiya raqamli texnologiyalar bilan uzviy bog‘langan sohaga aylanib bormoqda. Katta hajmdagi yozma manbalarni jamlash, saqlash va tahlil qilishda lingvistik korpus eng samarali vositalardan biri sifatida e’tirof etilmoqda. Korpus lingvistikasi ta’limotiga ko‘ra, tabiiy tilning tanlab olingan, muvozanatlashtirilgan va lingvistik jihatdan izohlangan matnlar to‘plami tadqiqotchiga til birliklarining amaldagi qo‘llanishini ilmiy asosda kuzatish imkonini beradi [1]. Ayni shu yondashuv jahon tilshunosligida so‘nggi o‘n yilliklarda til hodisalarini o‘rganishning yetakchi metodologiyasiga aylandi [2]. Bu metadologiya adabiyotshunoslik, xususan, mumtoz adabiyot namunalari uchun ham birdek o‘rinlidir.

Turkiy xalqlarning yozma merosi nihoyatda boy va qadimiy. O‘rxun-Enasoy bitiklaridan boshlanib, Yusuf Xos Hojibning “Qutadg‘u bilig”, Ahmad Yugnakiyning “Hibat ul-haqoyiq”, Ahmad Yassaviy hikmatlari, Rabg‘uziy qissalari, Xorazmiy “Muhabbatnoma”i, Sayfi Saroyi, Lutfiy, Alisher Navoiy, Zahiriddin Muhammad Bobur, Boborahim Mashrab va Muhammad Rizo Ogahiy ijodigacha bo‘lgan ulkan adabiy xazina turkiygo‘y mumtoz adabiyotning tarixiy taraqqiyotini aks ettiradi. Ushbu meros nafaqat badiiy qadriyat, balki eski turkiy va chig‘atoy tilining tarixiy holatini o‘rganish uchun ham, shu davrdagi adabiyiy jarayonlarni tahlil qilish uchun ham beqiyos manba hisoblanadi.

Biroq bu merosning aksariyati arab yozuvidagi qo‘lyozma va toshbosmalarda saqlangan bo‘lib, ular turli fondlarga tarqoq holda joylashgan. An’anaviy filologik tadqiqot usullari bilan bunday manbalarni qiyosiy o‘rganish, muayyan so‘z yoki grammatik shaklning ko‘p asrlik matnlar bo‘ylab qo‘llanishini kuzatish mashaqqatli va sekin kechadi. Elektron korpus esa qidiruv, konkordans, chastota va lingvostatistik tahlil imkoniyatlarini beruvchi vosita sifatida ushbu muammoni hal etishga xizmat qiladi.

O‘zbekistonda so‘nggi yillarda milliy-madaniy merosni raqamlashtirish, qo‘lyozma xazinani ilmiy muomalaga jalb etish va ma’naviy qadriyatlarni asrash davlat siyosatining ustuvor yo‘nalishlaridan biriga aylandi. Mumtoz adabiyot merosini raqamli korpus shaklida tashkil etish ushbu vazifaning eng samarali yechimlaridan hisoblanadi, chunki korpus bir vaqtning o‘zida ham saqlash, ham qidiruv, ham ilmiy tahlil funksiyalarini birlashtiradi. Shu ma’noda turkiygo‘y mumtoz adiblar korpusi milliy g‘oyaning raqamli infratuzilmasi sifatida ham ahamiyat kasb etadi.

Jahon tajribasida raqamli filologiya (digital humanities) yo‘nalishi milliy yozma merosni saqlash, ilmiy muomalaga jalb etish va xalqaro miqyosda targ‘ib qilishning zamonaviy shakliga aylandi. Ko‘plab davlatlar o‘z klassik adabiyotini raqamli korpuslar va tanqidiy elektron nashrlar orqali umumbashariy ilm-fan xizmatiga qo‘ymoqda. Turkiy mumtoz meros ham xuddi shunday raqamli infratuzilmaga muhtoj: aks holda, u ilmiy jamoatchilik uchun qidiriladigan, tahlil qilinadigan va qiyoslanadigan manba bo‘lmasdan qolaveradi. Aynan shu ehtiyoj turkiygo‘y mumtoz adiblar korpusini yaratishni hayotiy zaruriy vazifaga aylantiradi.

O‘zbekistonda va turkiy dunyoda so‘nggi yillarda milliy hamda mualliflik korpuslari yaratildi. Shunga qaramay, turkiygo‘y mumtoz adiblar merosini yagona, manba-tanqidiy asosda, ko‘p yozuvli va ko‘p qatlamli tarzda qamrab oluvchi maxsus ilmiy korpus hanuz to‘liq shakllangani yo‘q. Mavjud resurslar yo zamonaviy til bilan, yo faqat bitta muallif ijodi bilan cheklanadi. Mazkur maqolaning maqsadi turkiygo‘y mumtoz adiblar ilmiy elektron korpusini yaratishning nazariy tamoyillari, tarkibiy tuzilishi va metodologik asoslarini ishlab chiqishdan iborat. Bunda korpusning ilmiy ishonchliligi, reprezentativligi va tarixiy matnlarga xos filologik talablarga muvofiqligiga alohida e’tibor qaratiladi.

Tadqiqotning obyekti turkiygo‘y mumtoz adiblar yozma merosini raqamlashtirish va korpus shaklida tashkil etish jarayoni; predmeti esa korpusni yaratishning tamoyillari, tarkibiy tuzilishi va annotatsiya metodologiyasidir. Maqsaddan kelib chiqib quyidagi vazifalar belgilandi: mavjud milliy va xalqaro korpuslar tajribasini qiyosiy o‘rganish; matn tanlashning manba-tanqidiy mezonlarini asoslash; korpusning ko‘p qatlamli arxitekturasini modellashtirish; annotatsiya darajalarini, jumladan mumtoz she’riyat uchun aruz-vazn qatlamini ishlab chiqish; hamda korpusning ilmiy va lingvodidaktik ahamiyatini ochib berish.

Adabiyotlar tahlili

Korpus lingvistikasining nazariy asoslari J. Sinclair, T. McEnery, A. Wilson kabi olimlar ishlarida ishlab chiqilgan. Sinclair konkordans va kollokatsiya tahlilining metodologik tamoyillarini asoslab bergan [1], McEnery va Wilson esa korpusning til tadqiqotidagi o‘rni hamda annotatsiya darajalarini tizimlashtirgan [2]. D. Biber korpus dizaynining markaziy masalasi sifatida reprezentativlik va muvozanat muammosini ko‘tarib, matn tanlashning nazariy mezonlarini belgilagan [3]. P. Baker, A. Hardie va T. McEnery esa soha atamashunosligini bir tizimga solib, korpus lingvistikasining tushunchaviy apparatini shakllantirgan [4]. Ushbu tadqiqotlar korpusning umumiy nazariy poydevorini tashkil etadi.

Korpus lingvistikasi nazariyasida korpuslar bir necha turga ajratiladi: milliy (umumiy) korpus tilning butun tizimini vakillik qiladi; mualliflik korpusi bir ijodkor tilini qamraydi; diaxronik korpus tilning tarixiy o‘zgarishini kuzatishga

mo‘ljallangan; ixtisoslashgan korpus esa muayyan soha yoki janr matnlarini birlashtiradi. Turkiygo‘y mumtoz adiblar korpusi o‘z mohiyatiga ko‘ra diaxronik va ko‘p muallifli ixtisoslashgan korpus turlarini uyg‘unlashtiradi hamda matnshunoslik talablari tufayli tanqidiy (manba-variantli) korpus xususiyatlarini ham o‘zida mujassam etadi. Ushbu murakkab tabiat uni mavjud korpus turlaridan ajratib turadi va maxsus metodologik yechimlarni taqozo qiladi.

Turkiy tillar korpuslari orasida eng dastlabki va tarixiy jihatdan ahamiyatli loyiha — islomgacha bo‘lgan eski turkiy matnlarning elektron korpusi VATEC (Vorislamische Alttürkische Texte). U 1999–2003-yillarda M. Erdal, J. Gippert, K. Röhrborn va P. Zieme rahbarligida yaratilgan bo‘lib, buddaviy va moniy mazmundagi yodgorliklarni transliteratsiya, transkripsiya, tarjima va glossalash bilan qamrab oladi [9]. Eski turkiy tilning grammatik tavsifi M. Erdalning fundamental asarida mukammal bayon etilgan [5]. Zamonaviy turkiy til uchun esa Y. Aksan va hamkasblari tomonidan Turk milliy korpusi (TNC) ishlab chiqilib, u Britaniya milliy korpusi (BNC) modeliga asoslangan, 50 million so‘zdan iborat muvozanatli korpus sifatida shakllantirilgan [6].

O‘zbek tilshunosligida korpus masalasi 2010-yillardan jadal o‘rganila boshlandi. N. Abdurahmonova o‘zbek tili elektron korpusini yaratish g‘oyasini ilgari surib, uning konseptual asoslarini ishlab chiqdi va uzbekcorpora.uz milliy korpusining shakllanishiga asos soldi. B. Elov va uning hamkasblari korpusni dasturiy jihatdan hayotga tatbiq etishda, Sh. Hamroyeva esa o‘zbek tilida mualliflik korpusini tuzish mezonlarini belgilashda muhim hissa qo‘shdi. Milliy korpusni loyihalash tamoyillari A. Qarshiyev, M. Tursunov va S. Maxmudov ishlarida bayon etilgan [7]. Mumtoz adabiyotga oid alohida tajriba sifatida Alisher Navoiy mualliflik korpusini keltirish mumkin: unda XV asr til uslubi, iste’moldan chiqqan tarixiy va arxaik so‘zlar lingvostatistik tahlil asosida o‘rganilgan [10].

Metodologik yondashuv nuqtai nazaridan Rus milliy korpusi (RNC) hamda Leipzig Corpora Collection (LCC) tajribasi ham diqqatga sazovor. RNC tarkibida she’riy matnlar subkorpusi, sintaktik razmetka va parallel korpuslar mavjudligi ko‘p komponentli korpus qurishning namunasi bo‘la oladi; LCC esa ikki yuzdan ortiq til uchun bir xillashtirilgan statistik korpuslar tuzish tajribasini beradi. Bu tajribalar turkiy mumtoz korpusda parallel (masalan, chig‘atoy–zamonaviy o‘zbek) subkorpuslar hamda she’riy matnlar uchun alohida qatlamini loyihalashda yo‘l ko‘rsatadi. Shu bilan birga, tarixiy matnlar korpusi diaxronik korpus sifatida alohida yondashuvni davriylashtirish, orfografik variantlilik va morfologik ko‘p ma’nolilikni hisobga olishni — talab etadi.

Matnlarni elektron kodlash va almashish sohasida xalqaro TEI (Text Encoding Initiative) standarti keng qo‘llaniladi; u matn tarkibi va metama’lumotini XML asosida bir xillashtirilgan tarzda ifodalash imkonini beradi [8]. Yuqoridagi manbalar tahlili shuni ko‘rsatadiki, mavjud korpuslar yo faqat eski turkiy davr (VATEC), yo zamonaviy til (TNC, o‘zbek milliy korpusi), yoxud yakka muallif ijodi (Navoiy korpusi) doirasi bilan cheklangan. Turkiygo‘y mumtoz adiblarni ko‘p muallifli, ko‘p asrli, manba-tanqidiy va ko‘p yozuvli asosda birlashtiruvchi yaxlit ilmiy korpus modeli hozircha ishlab chiqilmagan. Aynan shu bo‘shliq ushbu tadqiqotning dolzarbligini belgilaydi.

Tadqiqot metodologiyasi

Tadqiqot korpus lingvistikasi, tarixiy leksikografiya va raqamli filologiya usullariga tayanadi. Metodologik jihatdan quyidagi yondashuvlar qo‘llanildi: mavjud milliy va xalqaro korpuslarni qiyosiy tahlil qilish; korpusning tarkibiy tuzilishini modellashtirish; matn tanlashning manba-tanqidiy mezonlarini ishlab chiqish; hamda annotatsiya bosqichlarini tizimli tavsiflash.

Korpus dizayni Biber ko‘rsatgan reprezentativlik va muvozanat tamoyiliga [3] hamda BNC/TNC amaliyotiga [6] asoslanadi, ammo tarixiy turkiy matnlarga xos xususiyatlar hisobga olingan holda moslashtiriladi. Matn tanlashda muallif, janr, davr va manba turi bo‘yicha muvozanat ta’minlanadi. Har bir matn birlamchi manbaga (qo‘lyozma, toshbosma yoki tanqidiy nashr) bog‘lanib, uning ilmiy ishonchliligi hujjatlashtiriladi.

Korpus tuzish jarayoni ketma-ket bosqichlardan iborat qilib loyihalandi: loyihalash va mezonlarni belgilash; birlamchi manbalarni yig‘ish va tavsiflash; matnlarni raqamlashtirish; TEI/XML asosida kodlash; ko‘p bosqichli annotatsiya; qidiruv tizimini o‘rnatish; va sifat nazorati. Har bir bosqichda natijalar hujjatlashtirilib, jarayonning takrorlanuvchanligi va tekshiriluvchanligi ta’minlanadi. Bunday bosqichli ish oqimi korpusning ilmiy standartlarga muvofiqligini kafolatlaydi.

Matnni raqamli qayta ishlash bosqichi arab yozuvidagi manbalar uchun optik belgi tanish (OCR), so‘ng qo‘lda tekshirish va transliteratsiyani o‘z ichiga oladi. Kodlashda Unicode standarti va TEI/XML sxemasi qo‘llaniladi [8]. Lingvistik ishlov token ajratish, lemmatizatsiya va morfologik teglashdan iborat bo‘lib, tarixiy so‘z shakllarini asos (lemma)ga keltirishda “Abushqa”, “Sanglox” singari tarixiy lug‘atlar va Navoiy asarlari tilining izohli lug‘atlari leksikografik tayanch vazifasini bajaradi. Qidiruv va tahlil bosqichida konkordans, N-grammlar, chastota ro‘yxatlari va kollokatsiya ko‘rsatkichlaridan foydalaniladi. Ishonchlilikni ta’minlash uchun avtomatik teglash natijalari mutaxassis tomonidan qo‘lda tekshiriladi va me’yoriylashtiriladi.

Natijalar va ularning tahlili

Tadqiqot natijasida turkiygoʻy mumtoz adiblar ilmiy elektron korpusining konseptual modeli ishlab chiqildi. Ushbu korpus XIV-XV asrlarga oid turkiy yozma yodgorliklarni manba-tanqidiy asosda jamlagan, koʻp yozuvli va koʻp qatlamli izohga ega raqamli tizim sifatida taʼriflanadi. Model quyidagi asosiy tamoyillarga tayanadi.

Birinchidan, ilmiylik va manba ishonchliligi. Mumtoz matnlarning turli qoʻlyozma nusxalari orasidagi tafovutlar tufayli har bir matn birligi aniq manbaga bogʻlanadi; metamaʼlumotda qoʻlyozma yoki nashr, uning saqlanish joyi va sanasi qayd etiladi. Bu tekstologik ishonchlilikni taʼminlaydi.

Ikkinchidan, reprezentativlik va muvozanat. Korpus tarkibi muallif, davr va janr boʻyicha muvozanatlashtiriladi; nazm va nasr, diniy va dunyoviy matnlar mutanosib ravishda qamrab olinadi [3].

Uchinchidan, koʻp yozuvlilik. Arab yozuvidagi asl matn (faksimile), uning ilmiy transliteratsiyasi va fonetik transkripsiyasi oʻzaro parallellashtiriladi. Bu yondashuv VATEC amaliyotidan oʻzlashtirilgan boʻlib [9], foydalanuvchiga bir vaqtning oʻzida asl yozuv va zamonaviy oʻqilishni qiyoslash imkonini beradi.

Toʻrtinchidan, koʻp qatlamlilik va standartlashtirish. Annotatsiya token, lemma, morfologik teg, aruz-vazn va janr darajalarida amalga oshirilib, butun tizim TEI/XML va Unicode standartlariga muvofiqlashtiriladi [8].

Mavjud korpuslar bilan qiyosiy tahlil taklif etilayotgan modelning oʻziga xos oʻrnini aniq koʻrsatadi (1-jadval).

1-jadval

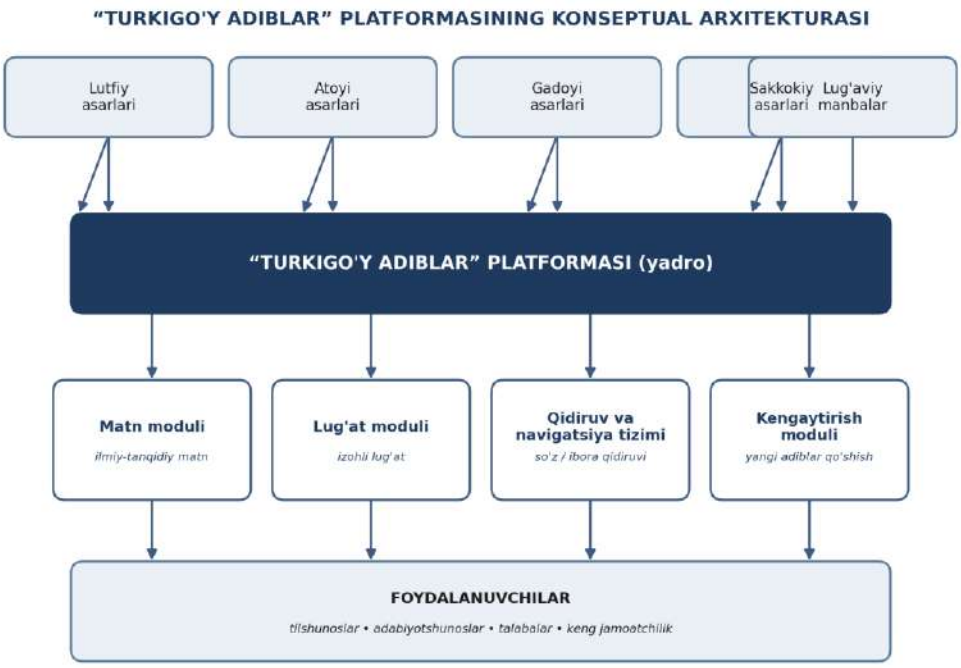
Turkiy tillarga oid mavjud korpuslarning qiyosiy tavsifi

Korpus	Qamrov (til/davr)	Hajm/xususiyat	Annotatsiya darajasi
VATEC	islomgacha eski turkiy (VIII–XIII asr)	buddaviy va moniy matnlar	transliteratsiya, transkripsiya, tarjima, glossa
Turk milliy korpusi (TNC)	zamonaviy turk tili	~50 mln soʻz, muvozanatli	morfologik teglash, lemmatizatsiya
Oʻzbek tili milliy korpusi	zamonaviy oʻzbek tili	Koʻp uslubli, milliy qamrov	token, lemma, morfologik razmetka
Alisher Navoiy korpusi	XV asr, bir muallif	mualliflik korpusi	leksik-statistik, izohli
Taklif etilayotgan korpus	XIV-XV asr, koʻp muallif	manba-tanqidiy, koʻp yozuvli	token–lemma–morfologiya–aruz–janr

Jadvaldan koʻrinadiki, mavjud resurslar yo faqat qadimiy davrni, yo zamonaviy tilni, yoxud yakka ijodkorni qamraydi. Taklif etilayotgan model esa koʻp asrli va koʻp muallifli qamrovni manba ishonchliligi hamda aruz annotatsiyasi bilan birlashtirishi bilan farqlanadi. Dastlab korpusga XIV-XV asarda Alisher Navoiy davrigacha ijod qilgan turkigoʻy adiblar merosi jamlanadi. Keyinchlik korpus tarkibini shakllantirishda matn tanlash mezonlari alohida ahamiyat kasb etadi. Tanlov turkiy yozma merosning butun tarixiy davomiyligini qamrab olishi koʻzda tutiladi: eng qadimgi qatlamda Oʻrxun-Enasoy bitiklari va islomgacha boʻlgan yodgorliklar; qoraxoniylar davri obidalari sifatida Yusuf Xos Hojibning “Qutadgʻu bilig” va Ahmad Yugnakiyning “Hibat ul-haqoyiq” asarlari; tasavvufiy adabiyot namunasi sifatida Ahmad Yassaviy hikmatlari va Rabgʻuziy qissalari; oltin Oʻrda va Xorazm adabiy muhitidan Xorazmiy “Muhabbatnoma”i hamda Sayfi Saroyi merosi; chigʻatoy adabiyotining klassik bosqichidan Lutfiy va Alisher Navoiy ijodi; keyingi davrlardan Zahiriddin Muhammad Bobur, Boborahim Mashrab va Muhammad Rizo Ogahiy asarlari kiritiladi. Bunday tanlov korpusga tarixiy va janriy vakillik baxsh etadi.

Matnlar davriylashtirilib joylashtirilishi korpusning diaxronik qimmatini oshiradi. Til holatining tarixiy oʻzgarishini kuzatish maqsadida matnlar eski turkiy, ilk chigʻatoy, klassik chigʻatoy ilk zamonaviy oʻzbek bosqichlariga ajratiladi. Bu tabaqalash foydalanuvchiga muayyan soʻz, grammatik shakl yoki uslubiy hodisaning asrlar davomida qanday evrilganini aniq davr kesimida kuzatish imkonini beradi. Shu tarzda korpus nafaqat qidiruv vositasi, balki tarixiy tilshunoslik uchun empirik laboratoriya vazifasini oʻtaydi.

Korpusning ichki tuzilishi besh qatlamli arxitektura ko ʻrinishida loyihalandi (1-rasm). Quyi qatlam manbalar bazasidan, undan yuqorisi matnni qayta ishlash bosqichidan iborat; markaziy qatlamda matn variantlari va metamaʼlumot joylashadi; ular ustidan koʻp bosqichli annotatsiya qatlami quriladi; eng yuqori qatlam esa foydalanuvchi interfeysi va qidiruv tizimini oʻz ichiga oladi.



1-rasm. Turkiygoʻy mumtoz adiblar ilmiy elektron korpusining koʻp qatlamli arxitekturasi

Korpusning ilmiy qiymati k'p jihatdan uning annotatsiya qatlamlariga bogʻliq. Har bir qatlam ma'lum bir tadqiqot vazifasini bajaradi (2-jadval). Tokenlash va lemmatizatsiya qidiruvni tarixiy soʻz shakllari boʻylab birlashtirib beradi; morfologik teglash grammatik tahlilni osonlashtiradi; aruz-vazn qatlami esa mumtoz she'riyatni poetik jihatdan oʻrganishga xizmat qiladi.

2-jadval

Korpusning annotatsiya qatlamlari

Qatlam	Mazmuni	Vazifasi
Tokenlash	matnni soʻz/birlikka ajratish	qidiruv birligini shakllantirish
Lemmatizatsiya	Soʻz shaklini asos (lemma)ga keltirish	tarixiy shakllarni birlashtirib qidirish
Morfologik teglash	Soʻz turkumi va grammatik shakl	grammatik va statistik tahlil
Aruz-vazn qatlami	she'riy oʻlchov va taqte'	poetik va uslubiy tadqiqot
Janr va uslub izohi	asar janri, izohtalab soʻzlar	textologik va lingvodidaktik tahlil
Metama'lumot	muallif, asar, manba, sana	ishonchlilik va filtrlash

Annotatsiya qatlamlarining oʻzaro bogʻliqligini yagona bayt misolida koʻrsatish mumkin. Mumtoz she'rdan olingan bir misra avval soʻz va turgʻun birliklarga ajratiladi (tokenlash); har bir soʻz shakli oʻz asosiga keltiriladi (lemmatizatsiya), bunda chigʻatoy tiliga xos qadimiy shakl zamonaviy lemma bilan bogʻlanadi; soʻngra soʻz turkumi va grammatik shakl belgilanadi (morfologik teg); nihoyat, misraning aruz vazni va taqte'i qayd etiladi (aruz qatlami). Natijada bir misra boʻyicha foydalanuvchi ham leksik, ham grammatik, ham poetik ma'lumotni yagona soʻrov orqali oladi. Ana shu koʻp qatlamli bogʻlanish korpusning oddiy matn omboridan tubdan farqli, tahliliy vositaga aylanishini ta'minlaydi.

Modelni amalga oshirishda bir qator ilmiy-texnik muammolar aniqlandi va ularning yechimlari taklif etildi. Arab yozuvidagi matnlarni OCR orqali oʻqish hozircha yuqori aniqlik bermaydi; shuning uchun avtomatik tanish natijalari mutaxassis tomonidan qoʻlda tekshiriladi va asl faksimile bilan solishtiriladi.

Orfografik variantlilik tarixiy matnlar korpusining oʻziga xos qiyinchiligidir. Arab yozuvi unlilarni toʻliq ifodalamagani va turli kotiblar bir soʻzni turlicha yozgani sababli, bir leksema matnlarda bir necha yozuv variantida uchraydi. Agar bunday variantlar birlashtirilmasa, qidiruv natijalari toʻkis chiqmaydi. Buning yechimi lemmatizatsiya bosqichida barcha variantlarni yagona asos ostiga jamlash, ayni paytda asl yozilishni saqlab qolishdan iborat. Shu tarzda foydalanuvchi ham me'yoriy lemma boʻyicha, ham aniq matn variantiga koʻra qidiruv oʻtkazishi mumkin boʻladi.

Foydalanuvchi interfeysi va qidiruv tizimi korpusning eng yuqori qatlamini tashkil etadi. Korpus menejeri konkordans qidiruvi, so‘z shakli va lemma bo‘yicha izlash, muallif, asar, davr hamda janr bo‘yicha filtrlash imkoniyatlarini beradi. Qidiruv natijalarini chastota jadvali, kontekst ko‘rinishi (KWIC) va statistik ko‘rsatkichlar shaklida taqdim etish, shuningdek natijalarni tashqi tahlil uchun eksport qilish nazarda tutiladi. Faksimile, transliteratsiya va transkripsiyani yonma-yon ko‘rsatuvchi parallel oyna esa foydalanuvchiga asl manba bilan zamonaviy o‘qishni bevosita qiyoslash imkonini beradi. Interfeysning qulayligi korpusni nafaqat tor doiradagi mutaxassislar, balki keng o‘quv auditoriyasi uchun ham foydalanishga yaroqli qiladi.

Korpusning uzoq muddatli barqarorligi va ochiqligi alohida e’tibor talab qiladigan masaladir. Xalqaro tajriba shuni ko‘rsatadiki, til resurslarining amaliy qiymati ularning uzluksiz saqlanishi, yangilanishi va boshqa tizimlar bilan o‘zaro moslashuviga (interoperability) bevosita bog‘liq. Shu bois korpus ochiq va bir xillashtirilgan formatlarda (Unicode, TEI/XML) qurilib, uning ma’lumotlar bazasi modulli tarzda kengaytiriladigan qilib loyihalanadi. Bu yangi matn, muallif yoki annotatsiya qatlamini tizimning umumiy tuzilishini buzmasdan qo‘shish imkonini beradi va korpusning kelgusi tadqiqotlar bilan birga o‘sib borishini ta’minlaydi.

Modelning yana bir muhim jihati – uning matnshunoslik (tekstologiya) an’anasi bilan uzviy bog‘liqligidir. Mumtoz asarlarning turli qo‘lyozma nusxalari o‘rtasidagi tafovutlar korpusda variant sifatida qayd etilishi mumkin, bu esa raqamli tanqidiy nashr yaratish uchun zamin hozirlaydi. An’anaviy matnshunoslikda mashaqqatli va sekin kechadigan nusxalararo qiyoslash korpus vositalari yordamida tezlashadi va aniqlashadi. Shu tariqa korpus filologik tadqiqotning raqamli davomiga aylanib, klassik matnshunoslik usullarini zamonaviy hisoblash imkoniyatlari bilan uyg‘unlashtiradi.

Alohida ilmiy yangilik sifatida she’riy matnlar uchun aruz-vazn annotatsiyasi qatlamining kiritilishini ta’kidlash lozim. Mumtoz turkiy nazm aruz vaznida yaratilgani bois har bir baytning vazni va taqte’i korpusda qayd etilsa, bu poetika va uslubshunoslik uchun mutlaqo yangi tadqiqot imkoniyatlarini ochadi. Bunday qatlam mavjud turkiy korpuslarning hech birida uchramaydi.

Bundan tashqari, korpus turkiy xalqlarning umumiy adabiy merosini qamragani bois turkiy davlatlararo ilmiy hamkorlik uchun ham qulay maydon ochadi. Umumiy yozuv, umumiy klassik matnlar va o‘zaro yaqin til qatlami turli mamlakat tadqiqotchilariga yagona metodologik asosda ish olib borish, ma’lumot almashish va qiyosiy turkiyshunoslik tadqiqotlarini birgalikda amalga oshirish imkonini beradi. Bu jihat korpusning ilmiy-integratsion salohiyatini yanada oshiradi.

Korpusning amaliy ahamiyati keng qamrovli. U tekstologik tadqiqotda nusxalararo tafovutlarni qiyoslash, tarixiy leksikografiyada so‘zlarning davriy qo‘llanishini kuzatish, mualliflik uslubini lingvostatistik usulda aniqlash hamda qiyosiy turkiyshunoslik uchun ishonchli empirik baza yaratadi. Lingvodidaktik jihatdan esa korpus XV–XIX asr til uslubini zamonaviy til bilan qiyoslab o‘rgatish, mumtoz adabiyot darslarida izohtalab so‘zlar ustida ishlash uchun boy o‘quv materiali beradi. Shu tariqa korpus turkiy mumtoz merosni ham saqlash, ham ilmiy muomalaga to‘liq jalb etish vositasiga aylanadi.

Xulosa

Tadqiqot natijalari turkiygo‘y mumtoz adiblar ilmiy elektron korpusini yaratish real va ilmiy jihatdan asoslangan vazifa ekanini ko‘rsatadi. Ishda korpusning nazariy tamoyillari (ilmiylik va manba ishonchliligi, reprezentativlik va muvozanat, ko‘p yozuvlilik, ko‘p qatlamlilik hamda standartlashtirish), besh qatlamli tarkibiy tuzilishi va bosqichma-bosqich metodologiyasi asoslab berildi.

Tadqiqotning ilmiy yangiligi shundan iboratki, birinchi marta turkiygo‘y mumtoz adiblar merosini keng qamrovda, manba-tanqidiy asosda hamda arab yozuvi–transliteratsiya–transkripsiya parallelligida birlashtiruvchi yaxlit korpus modeli taklif etildi. Ayniqsa, mumtoz she’riyat uchun aruz-vazn annotatsiyasi qatlamining kiritilishi va manba ishonchliligini ta’minlovchi metama’lumot tizimining ishlab chiqilishi modelni mavjud korpuslardan tubdan farqlaydi.

Amaliy jihatdan bunday korpus textologiya, tarixiy leksikografiya, lingvostatistik uslubshunoslik va lingvodidaktika sohalari uchun ishonchli empirik zamin bo‘ladi. Kelgusi tadqiqotlarda korpusning pilot versiyasini yaratish, annotatsiya vositalarini avtomatlashtirish va uni turkiy tillararo parallel korpuslar bilan bog‘lash istiqbolli yo‘nalish sifatida belgilanadi.

Mualliflar bayonotlari

Mualliflar hissasi: tadqiqot konsepsiyasi, tahlil va matnni tayyorlash muallif tomonidan bajarilgan. Jadvallar, konseptual xarita sun’iy intellekt yordamida tayyorlangan. Manfaatlar to‘qnashuvi: muallif manfaatlar to‘qnashuvi yo‘qligini bildiradi. Moliyalashtirish: tadqiqot nazariy xulosalarni jamlagani bois tashqi moliyaviy yordamsiz amalga oshirilgan. Amaliyotga bog‘lash uchun moliyaviy ko‘mak kerak bo‘ladi.

Foydalanilgan adabiyotlar

1. Sinclair J. Corpus, Concordance, Collocation. – Oxford: Oxford University Press, 1991.
2. McEnery T., Wilson A. Corpus Linguistics. – Edinburgh: Edinburgh University Press, 1996.

3. Biber D. Representativeness in Corpus Design // Literary and Linguistic Computing. – 1993. – Vol. 8, № 2. – P. 243–257.

4. Baker P., Hardie A., McEnery T. A Glossary of Corpus Linguistics. – Edinburgh: Edinburgh University Press, 2006.

5. Erdal M. A Grammar of Old Turkic. – Leiden – Boston: Brill, 2004.

6. Aksan Y., Aksan M., Koltuksuz A. et al. Construction of the Turkish National Corpus (TNC) // Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). – Istanbul: ELRA, 2012. – P. 3223–3227.

7. Qarshiyev A., Tursunov M., Maxmidov S. O'zbek tili milliy korpusini loyihalash // Computer Linguistics: Problems, Solutions, Prospects. – 2022. – Vol. 1, № 1.

8. TEI Consortium. TEI P5: Guidelines for Electronic Text Encoding and Interchange. – Text Encoding Initiative Consortium, 2023.

9. VATEC – Vorislamische Alttürkische Texte: Elektronisches Corpus / M. Erdal, J. Gippert, K. Röhrborn, P. Zieme. – Frankfurt am Main, 1999–2003. – URL: https://vatec2.fkidg1.uni-frankfurt.de.

10. Alisher Navoiy korpusi. – URL: http://alishernavoicorpus.uz.

11. O'zbek tilining milliy korpusi. – URL: https://uzbekcorpora.uz.

References

1. Sinclair J. Corpus, Concordance, Collocation. Oxford, Oxford University Press, 1991. (in English).

2. McEnery T., Wilson A. Corpus Linguistics. Edinburgh, Edinburgh University Press, 1996. (in English).

3. Biber D. Representativeness in Corpus Design. Literary and Linguistic Computing, 1993, vol. 8, no. 2, pp. 243–257. (in English).

4. Baker P., Hardie A., McEnery T. A Glossary of Corpus Linguistics. Edinburgh, Edinburgh University Press, 2006. (in English).

5. Erdal M. A Grammar of Old Turkic. Leiden – Boston, Brill, 2004. (in English).

6. Aksan Y., Aksan M., Koltuksuz A. et al. Construction of the Turkish National Corpus (TNC). Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). Istanbul, ELRA, 2012, pp. 3223–3227. (in English).

7. Qarshiyev A., Tursunov M., Maxmidov S. O'zbek tili milliy korpusini loyihalash [Designing the Uzbek National Corpus]. Computer Linguistics: Problems, Solutions, Prospects, 2022, vol. 1, no. 1. (in Uzbek).

8. TEI Consortium. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Text Encoding Initiative Consortium, 2023. (in English).

9. VATEC – Vorislamische Alttürkische Texte: Elektronisches Corpus. M. Erdal, J. Gippert, K. Röhrborn, P. Zieme. Frankfurt am Main, 1999–2003. Available at: https://vatec2.fkidg1.uni-frankfurt.de. (in German/Old Turkic).

10. Alisher Navoi Corpus. Available at: http://alishernavoicorpus.uz. (in Uzbek).

11. Uzbek National Corpus. Available at: https://uzbekcorpora.uz. (in Uzbek).