

MUALLIFLIK KORPUSINI SHAKLLANTIRISH VA RAQAMLASHTIRISH BOSQICHLARI: AMALIY PLATFORMA ASOSIDAGI YONDASHUV

Akramova Shohista Islom qizi,
3-bosqich tayanch doktorant
ToshDO‘TAU

Annotatsiya: Ushbu maqolada mualliflik korpusini yaratish, uni raqamli muhitga moslashtirish va ilmiy-tahliliy foydalanish uchun tayyorlash bosqichlari ketma-ket yoritiladi. Tadqiqot amaliy jihatdan ishlab chiqilayotgan veb-platforma imkoniyatlariga tayangan holda bayon qilinadi: matnlarni yig‘ish va metadata bilan belgilash, formatlarni standartlash, tokenizatsiya va indekslash, konkordans qidiruvi, harf kesimidagi lug‘at ko‘rinishi, statistik panel, etimologik lug‘at moduli hamda media-integratsiya jarayonlari izchil tahlil qilinadi. Maqolada, ayniqsa, o‘zbek lotin yozuvidagi apostrof variantlarini (o‘/g‘) normallashtirish, PDF/TXT manbalaridan matn ajratish va korpus sifatini oshirishga qaratilgan texnik yechimlarning lingvistik tadqiqotlar uchun ahamiyati ochib beriladi. Natijada muallif uslubini, leksik qatlamni va vaqt bo‘yicha til dinamikasini raqamli metodlar asosida o‘rganish imkonini beruvchi korpus modeli taklif etiladi.

Kalit so‘zlar: *mualliflik korpusi, raqamlashtirish, konkordans, tokenizatsiya, o‘zbek tili, etimologik lug‘at, korpus lingvistikasi, metadata, statistik tahlil.*

Abstract: This article sequentially describes the stages of creating an author corpus, adapting it to a digital environment, and preparing it for scientific and analytical use. The study is presented on the basis of the practical capabilities of the developed web platform: collecting texts and annotating them with metadata, standardizing formats, tokenization and indexing, concordance search, dictionary view at the letter level, statistical dashboard, etymological dictionary module, and media integration processes are consistently analyzed. Particular attention is paid to the normalization of apostrophe variants in the Uzbek Latin script (o‘/g‘), text extraction from PDF/TXT sources, and the importance of technical solutions aimed

at improving corpus quality for linguistic research. As a result, a corpus model is proposed that enables the study of authorial style, lexical layers, and the dynamics of language over time based on digital methods.

Keywords: *author corpus, digitization, concordance, tokenization, Uzbek language, etymological dictionary, corpus linguistics, metadata, statistical analysis.*

Kirish

Zamonaviy filologiya va korpus lingvistikasi sharoitida muallif merosini tizimli o'rganish uchun oddiy matn to'plami yetarli bo'lmay qolmoqda. Bugungi tadqiqotchi nafaqat matnni o'qish, balki uni statistik, qiyosiy va kontekstual jihatdan tahlil qilish imkoniyatiga ham ega bo'lishi zarur. Shu nuqtai nazardan mualliflik korpusi muayyan ijodkorning asarlari, maqolalari va boshqa matnlarini yagona elektron tizimga birlashtiruvchi ilmiy platforma sifatida muhim ahamiyat kasb etadi [1:105].

Mualliflik korpusi umumiy milliy korpuslardan farqli ravishda tor obyekt bitta muallif ijodiga yo'naltiriladi. Bu esa muallif uslubini chuqurroq o'rganish, leksik qatlamlarni aniqlash, davrlar bo'yicha til dinamikasini kuzatish va kollokatsion birliklarni tahlil qilish imkonini beradi. Shu sababli korpus yaratishda matnlarni shunchaki yig'ish emas, balki ularni ilmiy mezonlar asosida saralash va belgilash asosiy vazifa hisoblanadi [3:105].

Tadqiqot davomida ishlab chiqilayotgan veb-platforma asosida mualliflik korpusining amaliy modeli shakllantirildi. Platformada maqolalar, kitoblar, mediateka materiallari, lug'atlar va statistik modullar yagona tizimga integratsiya qilindi. Har bir matn uchun metadata maydonlari – muallif, nashr yili, janr, uslub va manba turi kabi atributlar biriktirildi. Metadata korpusning ilmiy qiymatini oshiruvchi asosiy omillardan biri bo'lib, keyingi qidiruv va statistik tahlil jarayonlarini samarali tashkil etishga xizmat qiladi [1:12-15; 5:68].

Mualliflik korpusini shakllantirishning dastlabki bosqichi manbalarni yig'ish va raqamlashtirishdan iborat. Bu bosqichda maqolalar, monografiyalar, intervyular va boshqa matnlar pdf hamda txt formatlarida tizimga yuklandi. PDF manbalardan matn ajratishda satr uzilishlari, noto'g'ri kodlangan belgilar va apostrof variantlari kabi texnik muammolar kuzatildi. Ayniqsa o'zbek lotin yozuvidagi o' va g' harflarining turli Unicode shakllarda yozilishi qidiruv natijalariga salbiy ta'sir ko'rsatishi aniqlandi [6:45].

Shu sababli korpusni yaratish jarayonida matnni normallashtirish va tozalash alohida bosqich sifatida tashkil etildi. Normallashtirish jarayonida turli apostrof variantlari yagona standartga keltirildi, ortiqcha belgilar olib tashlandi va satr bo'linishlari tiklandi. Bu esa keyingi tokenizatsiya va indekslash bosqichlarining sifatini sezilarli darajada oshirdi [7:82].

Tokenizatsiya jarayonida matn alohida birliklarga tokenlarga ajratildi. Tinish belgilari, raqamli birliklar va ortiqcha bo'shliqlar filtrlandi. Natijada faqat ilmiy tahlil uchun muhim bo'lgan leksik birliklar korpus yadrosiga kiritildi. Ushbu yondashuv konkordans va chastota tahlilida yuqori aniqlikni ta'minlaydi [8:56].

Korpus lingvistikasining asosiy metodlaridan biri bo'lgan konkordans qidiruvi platformaning markaziy modullaridan biri sifatida ishlab chiqildi. Konkordans orqali foydalanuvchi muayyan so'z yoki birikmaning barcha qo'llanish holatlarini kontekst bilan ko'rishi mumkin. Bu esa muallifning semantik maydoni, uslubiy registri va kollokatsion xususiyatlarini aniqlashda muhim vosita hisoblanadi [2].

Platformada statistik modul ham ishlab chiqilgan bo'lib, unda tokenlar soni, eng faol leksemalar, yillar kesimidagi matnlar soni va leksik dinamikani vizual tarzda ko'rsatish imkoniyati yaratilgan. Statistik panel yordamida muallif ijodidagi faol davrlar va dominant mavzularni aniqlash mumkin bo'ladi [4].

Mualliflik korpusining muhim jihatlaridan biri etimologik lug‘at moduli bilan integratsiya qilinganidadir. Ushbu modul orqali foydalanuvchi ma’lum bir leksemaning kelib chiqishi, morfemik tuzilishi va manba tiliga oid ma’lumotlarni olishi mumkin. Bu esa korpusni oddiy elektron kutubxona emas, balki ilmiy-analitik platforma darajasiga olib chiqadi [9:9-12].

Shuningdek, platformada foydalanuvchi interfeysi ilmiy izlanishlarga mos tarzda ishlab chiqildi. Bosh sahifa, maqolalar katalogi, statistika, konkordans va lug‘at modullari o‘zaro integratsiyalashgan holda ishlaydi. Bu tadqiqotchiga bir vaqtning o‘zida bir nechta tahlil usullaridan foydalanish imkonini beradi [3:105].

Demak, mualliflik korpusini yaratish va raqamlashtirish jarayoni ko‘p bosqichli hamda murakkab metodik tizim hisoblanadi. Manbalarni yig‘ish, metadata yaratish, matnni normallashtirish, tokenizatsiya, indekslash, konkordans va statistik tahlil bosqichlari o‘zaro uzviy bog‘liq holda ishlaydi. Ushbu yondashuv muallif uslubini raqamli metodlar asosida chuqur tahlil qilish, o‘zbek tilining leksik qatlamini o‘rganish va korpus lingvistikasini rivojlantirish uchun muhim ilmiy asos yaratadi

1. Mualliflik korpusining konseptual asoslari

Mualliflik korpusi umumiy milliy korpusdan farqli ravishda tor obyektga bir muallif ijodiga yo‘naltiriladi. Bu esa quyidagi ustunliklarni beradi:

- muallif idiolektini aniq ko‘rish;
- davrlar bo‘yicha uslubiy o‘zgarishni kuzatish;
- janrlararo leksik tafovutlarni aniqlash;
- kalit birliklar (keywords), kollokatsiyalar va dominant obrazlarni topish;
- ilmiy-tanqidiy nashrlar uchun dalillangan til materialini tayyorlash.

Shu sababli korpusni yaratishda “ko‘p matn yig‘ish” emas, balki “to‘g‘ri tanlash va to‘g‘ri belgilash” birinchi o‘ringa chiqadi. Bizning platformada ham ayni tamoyil asosida ma’lumot modeli ishlab chiqilgan: matn birligi sifatida maqola va kitob

alohida obyektlar sifatida saqlanadi, ularga muallif, nashr yili, janr, adabiy tur, uslub kabi atributlar biriktiriladi. Natijada qidiruv natijasi oddiy satrlar emas, balki ilmiy filtrlashga tayyor korpus segmentlari ko‘rinishida qaytariladi.

2. Korpusni shakllantirishning birinchi bosqichi: manbalarni tanlash va yig‘ish

Mualliflik korpusi sifatli bo‘lishi uchun manbalar representativ va tekshirilgan bo‘lishi kerak. Amaliy ishda quyidagi mezonlar asos qilib olinadi:

1. **Asosiy korpus manbalari:** maqolalar, risolalar, monografik bo‘limlar, kitoblar.
2. **Qo‘shimcha manbalar:** intervyu matnlari, mediateka transkriptlari, izohli matnlar.
3. **Nashriy va bibliografik ishonchlilik:** nashr yili, nashriyot, mualliflik tasdig‘i.
4. **Format xilma-xilligi:** TXT, PDF, qo‘lda kiritilgan matn.

Bizning dasturda matnlar asosan ikki asosiy obyektga ajratilgan: Article va Book. Bu ajratish juda muhim, chunki keyingi bosqichlarda qidiruv yoki statistika “faqat maqolalar”, “faqat kitoblar” kabi filtrlarda ishlashi talab etiladi.

Amaliy jihatdan bu bosqichda raqamlashtirishga keladigan asosiy muammo formatlardan kelib chiqadi:

- PDF dan ajratilgan matnda satr uzilishlari, tire bilan bo‘lingan so‘zlar, yashirin belgilar bo‘lishi mumkin;
- turli manbalarda apostrof belgisi turlicha kodlarda saqlangan bo‘ladi (‘, ‘, ’ va boshqa Unicode variantlari).

Shu sababli yig‘ish bosqichi darhol tozalash bosqichiga ulanadi.

3. Ikkinchi bosqich: metadata tizimini yaratish va ilmiy belgilash

Korpus matnini “ilmiy aylanmaga” kiritadigan kuch metadata hisoblanadi. Metadata bo‘lmasa, korpus oddiy elektron kutubxona bo‘lib qoladi.

Platformada har bir matn uchun quyidagi atributlar ko‘zda tutilgan:

- asar nomi;
- muallif;
- yozilgan yoki nashr yili;
- janr va adabiy tur;
- uslub va qo'llanish sohasi;
- nashriy ma'lumotlar;
- fayl manbasi (PDF/TXT).

Bu atributlar korpusga quyidagi ilmiy imkoniyatlarni beradi:

- **xronologik tahlil:** muallifning yillar bo'yicha faolligi;
- **janriy tahlil:** badiiy va ilmiy matnlardagi leksik tafovut;
- **uslubiy tahlil:** ma'lum davrda ustun bo'lgan birliklar;
- **matnlararo solishtirish:** bir xil lemma yoki birikmaning turli asarlardagi funksiyasi.

Demak, metadata bosqichi korpusning keyingi barcha hisob-kitob funksiyalari uchun “skelet” vazifasini bajaradi.

4. Uchinchi bosqich: matnni tozalash, normallashtirish va tokenizatsiya

Raqamlashtirishning eng nozik bosqichi matnni mashina o'qiydigan ko'rinishga keltirishdir. Bu jarayon odatda uch yo'nalishda kechadi:

4.1. Texnik tozalash

- yashirin belgilarni olib tashlash;
- satr bo'ylab bo'linib ketgan so'zlarni tiklash (kel-\ngan → kelgan);
- ortiqcha bo'shliqlarni standartlash.

4.2. Grafemik normallashtirish

O'zbek lotin matnlarida ayniqsa o' va g harflari turli Unicode belgilar orqali yozilishi korpus qidiruviga jiddiy xalaqit beradi. Bizning platformada bu muammo maxsus normallashtirish funksiyalari orqali bartaraf qilinadi: turli apostrof

variantlari yagona standartga keltiriladi, harf bo'yicha qidiruvda noto'g'ri ajralishlar kamaytiriladi.

4.3. Tokenizatsiya (so'zlarga ajratish)

Tokenizatsiya jarayonida:

- matn bo'shliq asosida so'zlarga bo'linadi;
- chetdagi tinish belgilar olib tashlanadi;
- juda qisqa va raqamli birliklar filtrlanadi;
- faqat lotin-o'zbek yozuviga mos birliklar qabul qilinadi.

Bunday yondashuv konkordans va chastota tahlilining aniqligini oshiradi. Amaliy tajriba shuni ko'rsatdiki, tozalash/tokenizatsiya sifati past bo'lsa, eng zo'r vizual statistika ham ilmiy jihatdan ishonchli bo'lmaydi.

5. To'rtinchi bosqich: indekslash va korpus yadrosini qurish

Tokenlar olingach, ular korpus yadro tizimiga joylanadi. Indekslashdan maqsad qidiruvni tezlashtirish va statistik hisob-kitobni qayta-qayta noldan boshlamaslik.

Bizning dasturda indekslash quyidagi natijalarni beradi:

- umumiy tokenlar soni;
- unikal so'zlar ro'yxati;
- har bir tokenning chastotasi;
- harf kesimidagi so'zlar to'plami;
- keyingi sahifalash va yuklab olish imkoniyati.

Bunda kesh mexanizmlari ham qo'llanib, ko'p ishlatiladigan so'rovlar tezroq qaytariladi. Bu ayniqsa katta hajmli korpuslarda foydalanuvchi tajribasini sezilarli yaxshilaydi.

6. Beshinchi bosqich: konkordans moduli orqali kontekstual tahlil

Korpus lingvistikasining markaziy usuli – konkordans. Konkordans qidiruvi so'zning matndagi barcha qo'llanishlarini kontekst bilan beradi. Shu orqali tadqiqotchi quyidagilarni kuzata oladi:

- soʻzning semantik muhitini;
- kollokatsion sheriklarini;
- uslubiy registrini;
- janrlar boʻyicha funksional farqini.

Platformadagi konkordans sahifasi qidiruv maydoni, qidiruv turi, filtrlash va natijalarni koʻrish mexanizmlarini birlashtiradi. Shuningdek, harf kesimida soʻzlar roʻyxatini olish va kerak boʻlsa ularni yuklab olish mumkin. Bu imkoniyat ilmiy izlanishlarda “qoʻlda kartoteka” usulini zamonaviylashtiradi.

Masalan, tadqiqotchi muallif matnlarida muayyan leksemaning faol davrini aniqlash uchun avval konkordansdan barcha kontekstlarni oladi, keyin metadata orqali ularni yillar boʻyicha guruhlaydi. Natijada semantik maydonning evolyutsiyasi ham koʻrinadi.

7. Oltinchi bosqich: statistik modul va vizual analitika

Korpusdan ilmiy xulosa chiqarishda sonli koʻrsatkichlar muhim. Dasturdagi statistik modul ikki asosiy yoʻnalishda ishlaydi:

1. Top soʻzlar (chastota reytingi)

- korpusdagi eng koʻp qoʻllangan birliklar;
- jami tokenlar soni;
- maksimal chastota.

2. Yillar kesimidagi dinamika

- maqolalar sonining yillar boʻyicha taqsimoti;
- kontentning davriy “pik” nuqtalari;
- muallif ijodidagi intensiv davrlarni aniqlash.

Vizual koʻrinish (diagrammalar) tadqiqotchiga tez orientatsiya beradi, ammo bu modulning ilmiy qiymati shundaki, u keyingi sifatiy sharh uchun dalillangan asos yaratadi. Yaʼni “muallif bu mavzuni koʻp ishlatgan” degan fikr endi raqamlar bilan tasdiqlanadi.

8. Yettinchi bosqich: etimologik lugʻatni korpusga integratsiya qilish

Mualliflik korpusi faqat “soʻz necha marta uchradi?” savoliga javob bermasligi kerak; “soʻzning kelib chiqishi nima?”, “qaysi manba tilga boradi?”, “qanday morfemik yasash orqali shakllangan?” kabi savollar ham muhim.

Shu ehtiyojdan kelib chiqib platformaga LugatEntry modeli asosidagi lugʻat moduli qoʻshilgan. Excel jadvalidan lugʻat yozuvlarini ommaviy import qilish imkoniyati mavjud. Bu modulda lemma, qisqa etimologik izoh, manba til, asos/ildiz, morfemik yasash va usul kabi maydonlar saqlanadi.

Natijada korpus tahlili quyidagi bosqichga koʻtariladi:

- konkordansdan topilgan birlik boʻyicha darhol leksik-etimologik maʼlumot olish;
- manba tillar statistikasi orqali leksik qatlamni baholash;
- muallif uslubida qaysi etimologik qatlam faolroq ekanini koʻrish.

Bu yondashuv korpusni oddiy qidiruv tizimidan ilmiy-analitik platformaga aylantiradi.

9. Sakkizinchi bosqich: foydalanuvchi interfeysi va ilmiy foydalanish ssenariylari

Ilmiy platforma texnik jihatdan kuchli boʻlishi bilan birga, foydalanishga qulay boʻlishi ham zarur. Bizning tizimda asosiy sahifalar mantiqan ajratilgan:

- bosh sahifa (korpus haqida umumiy maʼlumot);
- muallif haqida sahifa;
- maqolalar va kitoblar katalogi;
- konkordans;
- statistika;
- lugʻatlar va lugʻat statistikasi;
- mediateka.

Bunday modullilik ilmiy ish jarayonini tezlashtiradi. Masalan:

1. tadqiqotchi avval maqola/kitob tanlaydi;
2. kerakli leksema bo'yicha konkordans qidiradi;
3. statistik paneldan chastotani tekshiradi;
4. lug'atdan etimologik qatlamni ko'radi;
5. natijani maqolada dalil sifatida ishlatadi.

Demak, interfeys dizayni ham metodikaning bir qismi sifatida qaraladi.

10. To'qqizinchi bosqich: sifat nazorati va barqaror yangilash

Korpus bir martalik loyiha emas, balki doimiy ravishda boyitilib boradigan tizimdir.

Shu sababli sifat nazorati alohida bosqich sifatida yo'lga qo'yiladi:

- noto'g'ri OCR natijalarini tahrirlash;
- dublikat matnlarni aniqlash;
- metadata bo'sh maydonlarini to'ldirish;
- normallashtirish qoidalarini yangilash;
- indeksni qayta hisoblash.

Amaliy jihatdan bu ishlar boshqaruv paneli, import skriptlari va qidiruv natijalarini tekshirish orqali bajariladi. Ayniqsa yangi manbalar qo'shilganda tokenizatsiya va statistika qayta ishga tushirilishi korpusning izchilligini ta'minlaydi.

Ilmiy-amaliy ahamiyat

Tayyorlanayotgan platforma asosidagi mualliflik korpusi quyidagi yo'nalishlarda yuqori amaliy qiymatga ega:

- **Filologik tadqiqotlar:** uslub, leksika, semantika, kollokatsiya.
- **Ta'lim:** magistratura va doktorantura ishlari uchun dalillangan material bazasi.
- **Leksikografiya:** yangi izohli/terminologik lug'atlar tuzish.
- **Raqamli gumanitar fanlar:** til birliklarini vizual va statistik talqin qilish.
- **Madaniy merosni saqlash:** muallif matnlarining raqamli arxivi va uzoq muddatli foydalanish.

Alohida ta'kidlash kerakki, o'zbek tilidagi grafemik nozikliklar (apostrof variantlari, lotin yozuvi muammolari) inobatga olingan holda ishlab chiqilgan algoritmlar boshqa o'zbek korpus loyihalari uchun ham metodik asos bo'la oladi.

Xulosa

Mualliflik korpusini shakllantirish va raqamlashtirish jarayoni ko'p bosqichli, ammo izchil metodika asosida bajarilganda yuqori samaradorlik beradi. Ushbu maqolada ishlab chiqilayotgan dastur tajribasi asosida quyidagi ketma-ketlik asoslandi: manba yig'ish, metadata belgilash, matnni tozalash va normallashtirish, tokenizatsiya, indekslash, konkordans tahlili, statistik vizualizatsiya, etimologik lug'at integratsiyasi, interfeysga joylash hamda sifat nazorati.

Natijada muallif merosi nafaqat elektron shaklda saqlanadi, balki ilmiy tahlil uchun faol “raqamli laboratoriya”ga aylanadi. Bu yondashuv o'zbek tilshunosligi va adabiyotshunosligida korpus metodlarini keng joriy etishga, ilmiy xulosalarning dalillilik darajasini oshirishga va raqamli gumanitar tadqiqotlar uchun barqaror platforma yaratishga xizmat qiladi.

Kelgusida korpusni rivojlantirishning istiqbolli yo'nalishlari sifatida morfologik teglash, avtomatik lemmatizatsiya, semantik tarmoqlar, n-gram tahlili va API orqali tashqi ilmiy tizimlar bilan integratsiya taklif etiladi. Shunda mualliflik korpusi milliy raqamli filologiyaning tayanch infratuzilmasiga aylanishi mumkin.

Foydalanilgan adabiyotlar ro'yxati

1. Abdullayev, F., Ibrohimova, F. O'zbek tilida boshqaruv: monografiya. – Toshkent, 1962.
2. Boysinov, S., Mamirov, O. Kompyuter lingvistikasida modellashtirish metodidan foydalanish. – URL: www.jdpu.uz
3. Bushuy, T., Safarov, Sh. Til qurilishi tahlil metodlari va metodologiyasi. – Toshkent, 2007.



4. Ibragimov, J. O'zbek tili matnlarida so'zlarning bevosita (kontakt) va bilvosita (distant) birikuvchanligini dastlabki dasturiy modellashtirish // Zamonaviy dunyoda innovatsion tadqiqotlar: nazariya va amaliyot nomli ilmiy, masofaviy, onlayn konferensiya materiallari. – DOI: <https://doi.org/10.5281/zenodo.7515431>
5. O'zbek tili grammatikasi. II tom / Mualliflar jamoasi. – Toshkent, 1978.
6. Rahimov, A. Kompyuter lingvistikasi asoslari. – Toshkent, 2011.
7. Sayfullayeva, R., Mengliyev, B. va b. Hozirgi o'zbek adabiy tili. – Toshkent, 2009.
8. Tillayeva, R. Lisoniy birliklarni modellashtirish: magistrlik dissertatsiyasi ishi. – Namangan, 2016.
9. Ревзин, И. И. Модели языка. – Москва, 1962. – С. 9–12.