

JADID DAVRI ADABIY MUHITIGA DOIR ASARLARDAN KORPUSDA FOYDALANISH

Nuritdinov Abrorbek Sayfiddin o'g'li

nuriddinovabrorbek@gmail.com

University of business and science
nodavlat oliy ta'lim muassasasi o'qituvchisi

Annotatsiya. O'zbek tili korpusida matn bilan ishlash, korpusga kiritilayotgan manbaning tadqiq imkoniyatlari va unda tug'iluvchi muammolar haqida so'z yuritildi.

Abstract. The Uzbek language consisted of working with the text, the research capabilities of the source entered into the corps, and the problems that arise there.

Аннотация. В корпусе узбекского языка обсуждались возможности работы с текстом, исследовательские возможности источника, входящего в корпус, и проблемы, которые в нем возникают.

Kalit so'zlar: *qidiruv tizimi, korpus, korpus interfeysi, korpus annotatsiyasi, O'zbek tili korpusi, token, lemma, konkordans.*

Korpus turli xildagi tadqiqot uchun asos, manba, vosita vazifasini o'tay olganligi uchun ham dunyodagi juda ko'p sohalar korpusga tayanib ish ko'radi [Hamroyeva, 2020:16]. Aksariyat qidiruv tizimlari uchun eng murakkab muammolaridan biri – bu matnli qidirish. Ko'pgina qidiruv tizimlari uncha katta bo'lmagan massivlarda ishlaydi, lekin korpus hajmi ortgan sayin ishlash tezligi va qidirish sifati kamayadi. Bunda qidiruv tezligi nafaqat korpus hajmi, balki talabning murakkabligiga, so'zlar soniga uzviy bog'liq.

O'zbek tili korpusining qidiruv tizimiga bo'lgan asosiy talablarni ifodalaymiz:

- 1) so'z va so'z birikmasini ularning belgilari (grammatik, semantik va b.)ga qarab qidirish;
- 2) matn (nutq yoki asarning tugal fikr anglatuvchi parchasi) va so'zlar orasidagi masofani hisobga olish;
- 3) metamatnli ma'lumotni qidirish;
- 4) taraqqiy etgan til talablariga xos, o'z ichiga mantiqiy bog'lamalar, qavs va matn operatorlarni qamrab olish;
- 5) indeksatsiyalash samaradorligi;
- 6) yetarli darajadagi murakkab savolga yetarli darajadagi tezlikda javob topish;
- 7) keng ko'lamlilik, yetarli darajadagi katta hajmgacha (yuz millionlab so'zlarni ishlatish).



Korpus bilan ishlashda yuqori samaradorlikka erishish uchun, to'liq matnli qidiruv tizimi zarur. Mazkur tizim ulkan matnlar massivida matn qidiruv vositasi uchun maxsus optimallashtirilgan. Bunday tizimlar o'rtasida eng mashhurlari – Yandeks-server va Google internet qidiruvi, shuningdek, indeksatsiya va internet-manbalar qidiruvi kabi tizimlar ham. Tabiiy ravishda o'zbek tili korpusi uchun Yandeks-server tizimi juda mos. Yandeks-server tizimi juda yuqori samaradorlik va keng ko'lamlilik, matn qidiruvni to'liq amalga oshirish, qudratli til so'rovnomasiga ega bo'lish, matnlarning nostandart turlari uchun egiluvchan indeksator sozlagichini o'rnatish, o'zbek tilini to'liq ko'tarish kabi imkoniyatlariga ega. Yandeks-server qisqa fursat ichida ulkan matnlar massivida murakkab so'rovlarni amalga oshirish imkoniyatiga ega bo'lib, bunda qidiruv tezligi korpus hajmiga umuman bog'liq emas.

Bilamizki, matnlarni qayta ishlashda matndan foydalanuvchi qaysi maqsad bilan yondashganiga qarab matnlarni avtomatik qayta ishlovchi dasturlar mavjud. Bu dasturlar turlicha bo'lgani uchun ularni topishimiz va ulardan qanday foydalanishimiz haqida ma'lumotlarga ega bo'lishimiz kerak. Bu esa bizdan ko'proq vaqt sarflashimizni talab qiladi. Bularning barchasi korpusda jamlangan holda o'z yechimini topmoqda. Korpus bilan qanday ishlashni esa korpus interfeysi ko'rsatib beradi. Interfeys asli inglizcha so'z bo'lib, “tashqi ko'rinish” degan ma'noni anglatadi. Asosan kompyuter texnologiyasida qo'llanadi. Korpus interfeysi korpusning dastlabki kirganimizdagi ko'rinishi hisoblanadi. Bu vizual interfeys hisoblanadi. O'zbek tili korpusining interfeysini ko'rishimiz uchun internet tarmog'idagi <https://uzbekcorpus.uz> havolasiga kiramiz. Hozirda O'zbek tili korpusi online holda ishlaydi. O'zbek tili korpusi interfeysida foydalanuvchiga qulay bo'lgan til (o'zbek, rus, ingliz) va korpus annotatsiyasi (unda: bosh sahifa, loyiha a'zolari, izlash, morfoanalizator, parallel korpus, ta'limiy korpus, mualliflik korpusi, lingvistik resurslar, tezaurus, foydalanuvchi yo'riqnomasi, korpus haqida, ilmiy nashrlar)dan iborat. Interfeys orqali berilgan buyruqlardan ichki interfeyslar paydo bo'ladi.

Korpusdan izlash to'rt xil usul bilan (token, lemma, birikma bo'yicha va konkordans) amalga oshiriladi. Qidiruv davomida qaysi usulda qidirish buyrug'i bilan birgalikda uslub va davrni ham tanlash zarur. Token bo'yicha qidiruv so'z shakllarini topib beradi. Token bo'yicha qidirishda izlashga so'zning leksema shaklini kiritish zarur. Chunki Abdurauf Fitrat asarlarida, xususan, jadid davri adabiyotida grammatik qo'shimchalarning hozirgidan ancha farqi bo'lgan. Masalan, token bo'yicha qidirishda *ot* so'zini izlamoqchimiz. Foydalanuvchi biror manbada jadid davrida bu so'zning *otg'a* shaklini ko'rib, shunday shaklda qidirishi bu davr adabiyoti uchun natija doirasini toraytiradi. Bu so'zni u davrda *otqa* kabi shakllarda ham qo'llanganligini uchratish mumkin (manba kirillda): *Қараш белгиси бир **отқа** қўшилгач, ундан кейин келган бир **отға** учинчи турли “олмош излари” қўшиладур* [Abdurauf Fitrat, 2006:171]. Yoki *to'ldiruvchi* terminini hozirgi shu nom bilan izlashi mumkin. Lekin Abdurauf Fitrat asarlarida *to'ldirg'ich* shaklida



berilgan: *Бориш тўлдиригичи гапнинг шундай бир бўлагидирким, гапнинг бошқа бир бўлмидан ўзига томон бир ҳаракат, бир узаниш борлигини билдирадир. Бунинг белгиси “га — қа — га — ка” дир. Бориш тўлдиригичидан белгисининг тушириб қолгани ҳам бўладир: Бозор бордим — бозорга бордим. Бориш тўлдиригичига “га — га, давур — довур” қўшилганда ҳаракат ва узанишнинг сўнг ўзи бўладир. Бузоқнинг югургани оғилгача. Бозоргача бордим. Бу йил қишқача далада қолдим. Бу иш қиёматқа довур бўлмайди. Шаҳарга давур айланиб келдим* [Abdurauf Fitrat, 2006:182].

Bunday hozirgi kundagi farqli shakllarni izlashda ham munosib yechim topish zarur. Dastlabki barchaga ma'lum usul so'zshaklni yanada qisqartirish orqali izlash, bunda *to'ldir* shakli kiritiladi va kerakli natijaga erishiladi. Agar asos tarkibida farq bo'lsa-chi, masalan, *sifat* so'zi *sufat* shaklida uchraydi: *Сўз: от, суфат, феъл, олмош, кўмакчи сўзлари каби қисимларга ажраладир* [Abdurauf Fitrat, 2006:145]. Bu holat tezausda o'z yechimini topadi. Tezaus shakllantirilguncha qidirilishi lozim bo'lgan so'zlarning hozirgi va jadid davri mos ravishda variantlashtirilib sinonimlar qatori ishlab chiqiladi. Bu orqali so'zning hozirgi davridagi shakliga jadid davridagi shakli sinonim qilib so'z haqidagi metama'lumotga kiritiladi. Jadid davridagi shaklini ko'rgandan so'ng shu shaklda qidiriladi.

Lemma bo'yicha izlashda so'zlarning lug'atdagi shakli topiladi. Birikma bo'yicha izlashda birikma shaklida kelgan shakllar qidiriladi. Bu izlash jarayoni Abdurauf Fitratning asarlarida birikma holda kelgan tilshunoslik terminlarini topishda yordam beradi. Konkordansda esa so'zning boshqa so'zlar bilan qo'llanish qurshovini belgilab beradi.

Abdurauf Fitrat asarlari O'zbek tili korpusiga mualliflik korpusi sifatida joylandi. Asli tanlab olingan matnlar chegaralanganligi uchun ular maxsus matnlar korpus bazasi hisoblanadi. Maxsus matnlar korpusi – balansga keltirilgan korpus, u hajman u qadar katta bo'lmagan, alohida tadqiqiy topshiriqqa xizmat qiluvchi, ma'lum maqsadni ko'zlagan, tuzuvchi rejasiga muvofiq tayyorlanadi [Hamroyeva, 2020:55]. Bular sirasiga mualliflik va ta'limiy korpuslarni kiritish mumkin. Bu korpuslarning milliy korpuslardan farqli jihatlari ma'lum bir davrga yoki muallifga tegishli o'ziga xos jihatlari, so'zning zamonga, yozuvchiga mos ma'nosi, ajratilgan manbani til xususiyatlari, so'zlari orqali nimarni nazarda tutganligi chuqurroq tahlili namoyon bo'ladi. Abdurauf Fitratning asarlarini mualliflik korpusi qatorida tadqiq etish davr nuqtayi nazaridan ham, til nuqtayi nazaridan ham, mazmun nuqtayi nazaridan ham ahamiyatlidir. Avvalo, yuqorida ko'rib o'tganimizdek, qidiruv tizimida chigalliklarni keltirib chiqaradi. Jadid davri va hozirgi davr tili to'liq mos emasligi foydalanuvchiga kerakli natijani bermaydi. Mualliflik korpusi sifatida foydalanganda esa aynan jadid davri tili nuqtayi nazaridan yondashadi. So'zlarni shu davrga mos ravishda izlaydi. Mazmun jihatidan Abdurauf Fitratning asarlarini o'zini ham ajratgan holda tadqiq etish zarur. Asarlar badiiy adabiyotga, publitsistikaga, tarixga, adabiyotshunoslikka, tilshunoslikka oid ekanligini ham hisobga olinadi.

Masalan, tilshunoslikka oid asarlarini tahlil qilishda ko‘proq tilshunoslik terminlariga e‘tiborni qaratish mumkin. Terminlarning alohida lug‘atini shakllantirib foydalanuvchiga havola etiladi. Korpusdan foydalanuvchi jadid davri so‘zlarining shaklini bilishi unga korpusdan foydalanish imkoniyatini kengaytirib beradi. Shuning uchun morfologik teglash jarayonida so‘zning jadid davridagi har bir shaklini ko‘rsatish zarur. Agar korpus foydalanuvchi qidirgan so‘zni topa olmasa, mos variantlarni taqdim etadi. Shu joyga foydalanuvchi so‘zni hozirgi shakl bilan qidirsam, korpus jadid davridagi variantini taqdim etishi lozimligini hisobga olish zarur. Shularni hisobga olib Abdurauf Fitratning asarlarini mualliflik korpusi sifatida tadqiq etdik va O‘zbek tili korpusining mualliflik korpusi bazasiga joyladik. Tajribalar asosida keyinchalik jadid tilshunoslari asarlarining, balki barcha jadid ma’rifatparvarlarining asarlari asosida korpus yaratishimiz mumkin. Bunda bizga hozirgi tajribalarimiz dasturi amal vazifasini o‘taydi.

Shunday qilib, jadidlar korpusini quyidagicha interfeysda shakllantirish mumkin: (1-rasm)



1-rasm. Asosiy oyna

Jadid korpusini yaratishda e‘tibor beriladigan jihatlar quyidagilardan iborat:

- korpus nomini interfeysning yuqori qismiga joylanadi;
- korpusning annotatsiya qismida mundarija sifatida korpusga joylangan jadid ma’rifatparvarlarining ro‘yxati turadi. Bu korpusdan qanday ma’lumotlar topishimiz mumkinligini ko‘rsatadi;

— mundarijadagi har bir nomning ustiga kelib bosganimizda qo'shimcha ravishda yana uchta tugma (avtobiografiya, asarlari va tadqiqotlar) hosil bo'ladi. Bu tugmalarni bosish orqali birinchisidan muallif avtobiografiyasi, ikkinchisidan asarlar ro'yxati hamda uchinchisidan muallif va asarlari ustida olib borilgan keyingi tadqiqotlar ro'yxati haqida ma'lumotlar chiqadi;

— annotatsiya qismida jadidlar nomi tugagandan so'ng barcha korpuslar kabi korpusdan foydalanish, korpus haqida, loyiha a'zolari va boshqa qismlar joylanadi;

— asosiy oynada dastlab foydalanuvchiga qulay bo'lgan til (o'zbek, ingliz, rus), ikki qidiruv tizimi joylashadi;

— bu ikki qidiruv tizimi umumiy va maxsus qidiruvlardan iborat bo'ladi;

— umumiy qidiruv tizimi korpusning ma'lumotlar bazasidagi barcha manba bilan izlashga yordam beradi, aynan ma'lum manbani chegaralab olmaydi;

— maxsus qidiruv tizimida esa manbani chegaralab olishimiz mumkin bo'ladi. U uchta izlash qismi (so'zni kiriting, muallifni kiriting va asar nomini kiriting) orqali qidiradi. Bulardan so'z va muallif bo'yicha izlash majburiy, asar nomi bo'yicha qidirish esa ixtiyoriy, ya'ni bu qismga asar nomi kiritilmagan taqdirda ham topib bera olishi kerak. Chunki foydalanuvchi ba'zan muallifning barcha asari bo'yicha, ba'zan ma'lum asar bo'yicha izlashi mumkin;

— umumiy va maxsus qidiruv tizimida ham korpusda bashorat qila olish imkoniyati bo'lishi zarur. Qidiruv qismiga so'zni kiritish davomida uning bazadagi shakllarini (*ot*→*oti*, *otqa*, *ota*, *otadi*), bu holat mos ravishda lotin grafikasida kiritilsa, kirillda yoki aksincha (*o'xsha*→*yxuamulgan*, *yxuamuu*, *yxuaui*, *yxuaagan*) kabi ko'rsatishi zarur;

— qidiruvni lemma, token, birikma va konkordans bo'yicha ajratish natija chiqqandan so'ng qilinsa, qulaylik tug'diradi. Bunda dastlabki qidiruv token bo'yicha bo'ladi va natijalarning yuqori qismida yonma-yon token, lemma, birikma va konkordans tugmalari joylashadi. Ularni tanlash orqali moslanadi. Bu ko'rinish barcha qidiruv shakllarining natijalarini bir joyga jamlashga yordam beradi. Qidiruv natijalarining yuqori qismida joylashgan qidiruv shakllari o'zi bilan birga hujjatlar statistikasini ko'rsatib beradi: (2-rasm)

Jadid korpusi



[O'zbek](#) [English](#) [Русский](#)

Token	Lemma	Birikma	Konkordanser
17 ta hujjatda 28 ta so'z			
Abdurauf Fitrat. Tanlangan asarlar. IV jild. Oila yoki oila boshqarish tartiblari.			Zikr qilingan uchta tarbiyaning eng muhimi axloqiy tarbiya hisoblanadi.
Abdurauf Fitrat. Tanlangan asarlar. IV jild. Oila yoki oila boshqarish tartiblari.			O'zingiz xohlaganingizcha bolangizning aqliy va badan tarbiyasi bilan mashg'ul bo'lishingiz mumkin va qo'lingizdan ko'proq...
Abdurauf Fitrat. Tanlangan asarlar. IV jild. Oila yoki oila boshqarish tartiblari.			Binobarin, bolaning aqliy va jismoniy tarbiyasi samarasiz qolmasligi uchun farzandning axloqiy tarbiyasi jiddiy shug'ullanib ko'proq...
Abdurauf Fitrat. Tanlangan asarlar. IV jild. Oila yoki oila boshqarish tartiblari.			Axloqiy tarbiya — odam axloqini kamolga yetkazish demakdir, ya'ni odamni shunday tarbiya qilish kerakki, fe'li va amali o'ziga ham, ko'proq...

2-rasm. Qidiruv natijasining ko'rinishi

Shunday shaklda ko'rinish foydalanuvchiga qulay bo'ladi va osonlik bilan bir qidiruv shaklidan ikkinchi qidiruv shakliga o'tkaza oladi.

Bugungi kunda manbalarini elektron resursga aylanishi bilan birga uning so'zma-so'z qidiruv imkoniyatiga egaligi, bu jarayon tizimli ravishda bo'lishi korpusda matnlardan foydalanish imkoniyatining kengligini ko'rsatadi. Xususan, tilshunoslikda axborot tizimlarini yaratish ehtiyoji tadqiqotchi uchun kerakli bo'lgan ma'lumotlardan foydalana olish, o'z ishida unumliroq harakatlarni amalga oshirish va axborot oqimlarini samarali boshqarish imkonini beradi.

O'zbek tili korpusi qidiruv tizimi Abdurauf Fitratning asarlarini lemma, token, birikma va konkordans bo'yicha izlash imkoniyatini beradi. Bu orqali foydalanuvchiga qulay bo'lgan qidiruv shaklini tanlay oladi.

Korpusda kengaytirish va ixchamlik talabi mavjud. Kengaytirish talabi orqali ma'lumotlar bazasiga kirgan manbalarni yangi ma'lumotlar orqali boyitib boriladi. Bundan tashqari matn ustidagi yangi tahlillarni ham qo'shiladi. Ixchamlik talabi orqali korpus foydalanuvchiga zarur ma'lumotlarni o'zinigina taqdim etish, ortiqcha ma'lumotlarni saralay olish imkoniyatini ham yaratadi.



Foydalanilgan adabiyotlar:

1. Abduraxmonova N. Kompyuter lingvistikasi. – Toshkent, 2020.
2. Абдурауф Фитрат. Танланган асарлар. IV жилд. – Тошкент: Маънавият, 2006.
3. Ҳамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари. Монография. – Тошкент, 2020.
4. Ҳамроева Ш. Корпус лингвистикаси атамаларининг қисқача изоҳли луғати. – Бухоро, 2018.
5. Xolmonova Z. Kompyuter lingvistikasi. – Toshkent, 2019.
6. <https://uzbekcorpus.uz> — O‘zbek tili korpusi.
7. <http://uzschoolcorpara.uz> — O‘zbek tilining ta’limiy korpusi.