O‘ZBEK TILI KORPUSI: TARIXIY RIVOJLANISHI, MAQSAD VA VAZIFALARI

Botir Elov Boltayevich,
Texnika fanlari falsafa doktori, dotsent
elov@navoiy-uni.uz
ToshDO‘TAU

Primova Mastura Hakim qizi,
O‘qituvchi
primovamastura@navoiy-uni.uz
ToshDO‘TAU

Amirkulov Marufjon Alikulovich,
tayanch doktorant
amirkulov.edu01@gmail.com
ToshDO‘TAU

Annotatsiya. Korpus lingvistikasi zamonaviy tilshunoslikning muhim yo‘nalishlaridan biri bo‘lib, tilning haqiqiy foydalanishini aks ettiruvchi katta hajmdagi matnlar to‘plamini tahlil qilishga asoslanadi. Korpus tilning turli kontekstlarda qo‘llanilishini namoyon etuvchi, maxsus tanlangan va tizimlashtirilgan matnlar bazasiga aytiladi. Bu soha tilni o‘rganishda empirik ma’lumotlarga tayanish imkonini beradi va tilshunoslik nazariyalarini sinovdan o‘tkazishda muhim vosita sifatida xizmat qiladi. O‘zbek tili milliy korpusi esa ushbu umumiy yondashuvning mahalliy ilovasi sifatida o‘zbek tilining o‘ziga xos xususiyatlarini saqlash, hujjatlashtirish va zamonaviy texnologiyalar bilan integratsiya qilishda markaziy o‘rin tutadi. O‘zbek tili korpusi nafaqat tilshunoslik tadqiqotlari uchun muhim resurs, balki sun’iy intellekt (SI) va tabiiy tillarni qayta ishlash (NLP) texnologiyalari uchun ham asosiy ma’lumotlar bazasi hisoblanadi. Ushbu maqolada o‘zbek tili korpusining tarixiy rivojlanishi, uning maqsadlari va vazifalari batafsil tahlil qilinadi.

Kalit so‘zlar: *Til modellari, n-gram til modeli, unigram, modelni baholash, Laplas silliqlash, mashinali o‘qitish.*

Kirish

Korpus lingvistikasi – bu tabiiy tillarning yirik elektron ma’lumotlar bazasini (korpus) yaratish, tahlil qilish va ularni lingvistik tadqiqotlar uchun qo‘llashga qaratilgan fan sohasi. O‘zbek tili korpusi bu sohaning asosiy obyekti bo‘lib, u o‘zbek tilining leksik, grammatik, semantik va pragmatik xususiyatlarini o‘rganish uchun asos bo‘lib xizmat qiladi. **Til korpusi** – strukturallashtirilgan, annotatsiyalangan (teglangan) va statistik jihatdan ishonchli matnlar to‘plami. O‘zbek tili korpusidan turli dialektlar, tarixiy matnlar va zamonaviy matnlar

(matbuot, adabiyot, ilmiy asarlar)ni o‘z ichiga olgan holda tilning to‘liq spektrini aks ettirishda foydalaniladi.

Til korpusi tushunchasi jahon tilshunosligida 1960-yillardan shakllana boshlagan bo‘lib, dastlab ingliz tilida Braun korpusi (Brown Corpus) yaratilgan (1961). Keyinchalik yirik milliy korpuslar, masalan, Britaniya Milliy Korpusi (BNC, ~100 million so‘z) va boshqalar tuzilib, korpus lingvistikasi fan sifatida rivoj topdi. Korpuslar tilshunoslik tadqiqotlarida inqilob yasadi: ilgari olimlar misollarni qo‘lda yig‘ib, ko‘p vaqt sarflagan bo‘lsa, endilikda raqamli korpuslar yordamida soniyalar ichida yuzlab misollarni topish va tahlil qilish imkoni tug‘ildi. Shu tariqa, zamonaviy lingvistikada til korpuslari **empirik tadqiqotlar** uchun asosiy manba sifatida e’tirof etila boshlandi.

O‘zbek tilida korpus lingvistikasi bo‘yicha dastlabki tadqiqotlar 20-asrning ikkinchi yarmida tilshunoslik tadqiqotlari doirasida shakllana boshlagan. Bu davrda O‘zbek tilining leksikasi, grammatikasi va dialektlarini o‘rganishda kichik hajmdagi matnlar to‘plamlari qo‘lda tuzilgan edi. Biroq, zamonaviy ma’noda korpus yaratishga oid loyihalar 21-asr boshlarida, kompyuter texnologiyalarining rivojlanishi bilan birga boshlandi. Dastlabki loyihalar asosan ma’lum bir janrga (masalan, adabiy matnlar yoki rasmiy hujjatlar) yo‘naltirilgan bo‘lib, ularning hajmi va funksionalligi cheklangan edi.

O‘zbek tili korpusining shakllanishi quyidagi bosqichlarni o‘z ichiga oladi:

1. **Raqamlashtirish davri (2000-yillar):** Kompyuter texnologiyalarining kirib kelishi bilan matnlarning elektron formatda saqlanishi. Toshkent davlat universiteti va O‘zbekiston FA tilshunoslik instituti tomonidan dastlabki elektron korpus loyihalari.
2. **Tizimlashtirish va kengaytirish (2010-yillar):** Ushbu bosqichda matnlar to‘plami nafaqat hajman o‘sdi, balki turli janrlar (ilmiy, ommaviy axborot vositalari, og‘zaki nutq transkriptlari) bilan boyitildi. Shu bilan birga, matnlarni annotatsiya qilish (masalan, so‘z turlari bo‘yicha belgilash) bo‘yicha ilk tajribalar amalga oshirildi.
3. **Milliy korpusning shakllanishi (2010-yillardan boshlab):** O‘zbekiston Respublikasi Vazirlar Mahkamasining qarorlari asosida “O‘zbek tili milliy korpusi” loyihasining ishga tushirilishi. O‘zbek tili milliy korpusi loyihasi rasman boshlanib, davlat dasturlari va ilmiy grantlar qo‘llab-quvvatlovi bilan rivojlandi. Bu davrda korpusning hajmi millionlab so‘zlarga yetdi va undan turli ilmiy va texnologik maqsadlarda foydalanish imkoniyati kengaydi.
4. **BigData va AI integratsiyasi (2020-yillar):** Hadoop, Spark kabi BigData platformalarida korpus ma’lumotlarini qayta ishlash. GPT-3/4, BERT kabi neyron tarmoqlarning o‘zbek tili uchun moslashtirilishi.

Hozirgi kunda O‘zbek tili milliy korpusi turli manbalardan (kitoblar, gazetalar, veb-saytlar, ijtimoiy tarmoqlar) olingan matnlarning katta to‘plamini o‘z ichiga oladi. Korpusning umumiy hajmi millionlab so‘zlardan iborat bo‘lib, u turli davrlar (tarixiy matnlar, sovet davri, mustaqillik davri) va janrlarni (ilmiy, adabiy, publitsistik) qamrab oladi. Shu bilan birga, korpus foydalanuvchilar uchun qulay interfeys va qidiruv tizimi bilan ta’minlangan bo‘lib, bu uni tadqiqotchilar va texnologlar uchun qulay vositaga aylantiradi.
O‘zbek tili korpusining asosiy maqsadlari quyidagilardan iborat:

1) *O‘zbek tilining tarixiy va zamonaviy variantlarini arxivlash.*
2) *Yo‘qolib borayotgan dialektlarni (Qashqadaryo, Surxondaryo) hujjatlashtirish.*
3) *Lingvistik qonuniyatlarni statistik usullar bilan tekshirish (masalan, so‘z chastotasi, kolokatsiyalar).*
4) *Adabiyotshunoslik, tarix va sotsiolingvistika sohalarida qo‘llash.*
5) *Mashina tarjimasi, ovozli assistentlar, avtomatik tekshirish tizimlari uchun ma’lumotlar bazasi.*
6) *NLP (Tabiiy tilni qayta Ishlash) algoritmlarini o‘zbek tiliga moslashtirish.*
7) *Sun’iy intellekt va NLP texnologiyalarini rivojlantirish.*
8) *O‘quv qo‘llanmalari, lug‘atlar va onlayn platformalar yaratishda korpusdan foydalanish.*

Adabiyotlar sharhi

Tony McEnery va Andrew Hardie tomonidan yozilgan “Corpus Linguistics: Method, Theory and Practice”[1] kitobi korpus lingvistikasi sohasidagi metodologiya, nazariy asoslar va amaliy qo‘llanilishni har tomonlama yoritgan muhim ilmiy manba hisoblanadi. Mualliflar korpusga asoslangan tadqiqotlarning statistik va kvalitativ usullarini batafsil ko‘rib chiqib, lingvistik nazariyalarni empirik ma’lumotlar bilan sinab ko‘rishning ahamiyatini ta’kidlaydi (masalan, kolokatsiyalar, so‘z chastotasi). Ular monolingval, multilingval va ixtisoslashgan korpuslar kabi turli korpus turlarini, shuningdek, konkordans, kollokatsiya va kalit so‘z tahlili kabi analitik usullarni batafsil o‘rganib, tadqiqotchilarga o‘z ishlarini loyihalash va amalga oshirishda yo‘l-yo‘riq beradi. Bundan tashqari, korpus lingvistikasidagi cheklovlar, xususan, reprezentativlik va umumlashtirish muammolari tanqidiy baholanib, soha bo‘yicha chuqur va o‘ylangan yondashuvni rag‘batlantiradi. Umuman olganda, ushbu kitob yangi boshlayotgan va tajribali tadqiqotchilar uchun metodologik qat’iylik va refleksivlik bilan ilmiy tadqiqot olib borishda qimmatli qo‘llanma vazifasini o‘taydi.

O‘zbek tilshunosligida ham til korpusi g‘oyasi nisbatan kechroq shakllandi. Ilk marotaba 2010-yillar atrofida mahalliy olimlar korpus lingvistikasi masalalariga

e’tibor qaratib, ushbu yo‘nalishda tadqiqotlarni boshladilar. Jumladan, O‘zbekiston Milliy Universiteti professori N.Z. Abdurahmonova o‘zbek tili elektron korpusini yaratish g‘oyasini ilgari surib [2], uning konseptual va ilmiy asoslarini ishlab chiqdi. Abdurahmonova rahbarligidagi ilmiy guruh o‘zbek tili uchun ilk elektron korpus loyihasini boshladi va bu boradagi natijalar keyinchalik 2021-yilda himoya qilingan doktorlik dissertatsiyasida yoritildi. Shu bilan birga, ToshDO‘TAU olimasi Sh. Hamroyeva mustaqil tadqiqotchi sifatida o‘zbek tilida **mualliflik korpusi** yaratish masalasini o‘rganib, bu bo‘yicha ilk ilmiy izlanishlarni amalga oshirdi. Hamroyevaning tadqiqoti o‘zbek tilida mualliflik korpusini tuzish mezonlarini belgilab beruvchi muhim manba bo‘lib, milliy korpus g‘oyasining shakllanishida dastlabki qadamlardan biri sifatida e’tirof etildi.

Ilk ilmiy qarashlarda Abdurahmonova va Hamroyeva o‘zbek milliy korpusini yaratishning dolzarbligini asoslab berdilar. Ular o‘zbek tilini raqamlashtirish, katta hajmdagi matnlarni jamlab, zamonaviy til holatini aks ettiruvchi elektron baza yaratish nafaqat lingvistik tadqiqotlar, balki tilning taraqqiyoti va xalqaro maydonda tutgan o‘rnini mustahkamlash uchun zarur, degan xulosalarga kelishgan. Ayniqsa, Sh. Hamroyeva 2018-yilda chop etgan maqolasida milliy korpusni *“til ma’naviy merosni saqlash va boyitishning asosiy vositasi”* deb ta’riflab, **o‘zbek tilini jahon tillari qatoriga olib chiqishning yo‘li – uni raqamlashtirish va milliy korpusini yaratish** ekanini alohida ta’kidlagan. Shunday qilib, o‘tgan asr oxiri va XXI asr boshlariga kelib o‘zbek tilida milliy korpus yaratish g‘oyasi ilmiy adabiyotlarda shakllandi va dastlabki amaliy qadamlar tashlandi.

O‘zbek tilining Milliy Korpusini amalda shakllantirish va uni dasturiy jihatdan hayotga tatbiq etishda Toshkent davlat o‘zbek tili va adabiyoti universiteti olimlari, xususan, B. Elov va uning hamkorlari katta hissa qo‘shdilar. Abdurahmonova boshchiligidagi loyiha natijasi o‘laroq 2020-yilda internetda “O‘zbek tilining milliy korpusi” (https://uzschoolcorpara.uz/) ishga tushirildi. Mazkur korpus dasturiy jihatdan to‘liq funksionallikka ega bo‘lib, maxsus korpus menejeri yordamida matnlarni saqlash, izlash va tahlil qilish imkoniyatlarini beradi. Korpus kontenti juda boy: **u o‘zbek tilidagi lug‘atlar, internet sahifalari, o‘quv qo‘llanmalar, ilmiy, rasmiy va badiiy matnlar majmuasini** o‘z ichiga oladi[3]. Matnlar zamonaviy o‘zbek tilining xilma-xil uslub va janrlarini qamrab olishi uchun keng ko‘lamda to‘plangan. Natijada elektron korpus o‘zbek tilining hozirgi holatini aks ettiruvchi katta hajmdagi yozma ma’lumotlar bazasini yaratishga muvaffaq bo‘lindi.

B. Elov va jamoasi korpusning texnik arxitekturasini ishlab chiqishda zamonaviy korpus menejment tizimlariga tayandilar. Korpusning veb-interfeysi orqali foydalanuvchilar turli usullarda qidiruv o‘tkazishi mumkin: xususan, **so‘zning asl ko‘rinishi (token) bo‘yicha, so‘zning lug‘aviy shakli (lemma)** bo‘yicha, **so‘z birikmalar va konkordans bo‘yicha qidirish** imkoniyati mavjud.

Bunday qidiruv turlari korpusdan foydalanuvchilarga kerakli til birligini tez va aniq topish hamda keng kontekstda ko‘rish imkonini beradi. Korpus tizimi foydalanuvchiga natijalarni qulay shaklda taqdim etadi: bunda kontekstda topilgan so‘zlar ajratib ko‘rsatiladi, zarur hollarda statistik ma’lumotlar (so‘z chastotasi, kombinatsiyalanuvchi so‘zlar va h.k.) ham birgalikda ko‘rsatiladi. Korpus platformasi tarkibida maxsus **morfologik tahlilchi** va **stemming dasturlari ham integratsiya qilingan** bo‘lib, foydalanuvchi xohlagan so‘zning tuzilishini, uning boshi va qo‘shimchalarini ajratib tahlil qilishi mumkin. Bu esa o‘zbek tilidek boy morfologiyali til korpusida samarali navigatsiya qilish va chuqurroq lingvistik tahlil o‘tkazishga xizmat qiladi.

Milliy korpusda lingvistik annoatsiya va teglash ishlari bosqichma-bosqich amalga oshirilmoqda. Ma’lumki, korpusning qimmati uning ichidagi matnlarning ma’lum darajada teglangan, ya’ni lingvistik belgi qo‘yilgan bo‘lishida namoyon bo‘ladi. B. Elov va Sh. Hamroyeva ilmiy hamkorlikda o‘zbek tilining nutq qismlarini belgilash (PoS tagging) uchun maxsus teglar tizimini ishlab chiqdilar [4]. Ular 2022-yilgi maqolalarida o‘zbek tilidagi so‘z turkumlari uchun batafsil belgi to‘plamini (tegsetni) taklif qilib, **korpus matnlarini avtomatik teglash** masalasini ko‘tardilar [5]. Bu ish o‘zbek korpus lingvistikasi uchun muhim ahamiyatga ega, chunki agglutinativ xususiyatga ega o‘zbek tilida so‘zlarni morfologik jihatdan tahlil qilgan holda to‘g‘ri teglash murakkab vazifa hisoblanadi. Hozirgi kunda korpus matnlarining asosiy qismi **morfologik teglashdan o‘tkazilgan** va foydalanuvchi so‘zlarni **so‘z turkumi, grammatik kategoriyalari** bo‘yicha filtrlab qidirishi mumkin. Korpus jamoasi bundan tashqari **leksik-statistik ko‘rsatkichlarni** ham hisoblash ustida ishlamoqda. Misol uchun, TF-IDF kabi statistik usullar o‘zbek korpusi matnlarida muayyan so‘zning nisbiy ahamiyatini aniqlash uchun qo‘llanilgan bo‘lib, bunday tadqiqotlar natijalari ham e’lon qilindi[6]. Korpusda so‘zlarning chastotasi, eng ko‘p qo‘llanuvchi birliklar ro‘yxati, kollokatsiyalar (qo‘shma yasalmalarning chiqish tezligi) kabi ko‘rsatkichlar avtomatik tarzda olinadi va ilmiy izlanishlar uchun tayyor material sifatida xizmat qiladi.

O‘zbek milliy korpusining hajmi va qamrovi doimiy kengayib bormoqda. Dastlabki versiyada korpus hajmi bir necha million so‘zni tashkil etgan bo‘lsa, qisqa fursatda matnlar bazasi sezilarli ravishda oshirildi (hozirgi kunda o‘n millionlab so‘zlar). Matnlar **uslubiy jihatdan turli bo‘lgani** uchun korpus tarkibini statistik tahlil qilish qiziqarli natijalar bermoqda. Masalan, ilmiy uslubdagi matnlarda eng yuqori chastotali so‘zlar rasmiy va badiiy uslub matnlaridagidan farq qilishi kuzatilgan (bu boradagi aniq ma’lumotlar korpus asosida tuzilgan tezislarda keltirilmoqda). Bundan tashqari, korpusdagi matnlardan foydalanib **N-grammalar, so‘zlarning o‘rtacha uzunligi, gap qurilishidagi o‘rtacha elementlar soni** kabi statistik ko‘rsatkichlar ham avtomatik chiqarilishi yo‘lga qo‘yilgan. B. Elov va hamkorlarining izlanishlari natijasida kelgusida **korpusni intellektual tahlil qilish**

(masalan, matnlardan mavzuga oid axborotni avtomatik ajratish, ma’lumotlarni klasterlash) kabi funksiyalarni joriy etish rejalashtirilmoqda.

Qarshiyev, Tursunov va Maxmidov (2022)ning ilmiy tadqiqoritda o‘zbek tili milliy korpusini loyihalashning strukturaviy tamoyillari, lingvistik va texnologik jihatlari ko‘rib chiqilgan [5]. Mualliflar korpusni multidisiplinar yondashuv asosida shakllantirish zarurligini ta’kidlab, uning tarkibiga dialektal matnlar, tarixiy manbalar va zamonaviy matbuot materiallarini kiritishning ahamiyatini alohida urg‘ulaydi. Tadqiqotda o‘zbek tilining morfologik murakkabligi (masalan, so‘z yasalishi, izofa konstruksiyalari) va annotatsiya jarayonidagi qiyinchiliklar (tokenizatsiya, lemmatizatsiya) algoritmik yechimlar bilan tavsiflangan.

M. B. Kasimova (2024) [7] o‘zbek tilining milliy korpusida darajanish hodisasini ilmiy tahlil qilib, ushbu jarayonning til ichidagi ifodalanishini o‘rganadi. Muallif o‘zbek tilidagi nisbiy va orttirma darajalar (masalan, “ko‘proq”, “eng yaxshi”), shuningdek, so‘z birikmalari orqali intensivlikni ifodalash vositalarini korpus ma’lumotlari asosida tizimli tahlil qilgan. Tadqiqotda korpus annotatsiyasining lingvistik aniqligi (POS teglar, sintaktik bog‘lanishlar) va ma’lumotlarning reprezentativligi urg‘ulanib, darajalanishning dialektal va adabiy variantlari solishtirilgan. Bu ish o‘zbek tilining dinamik xususiyatlarini korpus lingvistikasi nuqtai nazaridan tushunishga muhim hissa qo‘shadi, shu bilan birga NLP tizimlarini milliy tilga moslashtirish uchun amaliy asos yaratadi.

Xudayberganov, N. (2024) [8] tomonidan yozilgan “O‘zbek tili korpusiga morfologik ishlov berish” maqolasi o‘zbek tili milliy korpusida morfologik tahlilning dolzarb masalalarini ilmiy jihatdan o‘rganadi. Muallif o‘zbek tilining murakkab morfologik tuzilishiga mos keladigan usullar va algoritmlarni tahlil qilib, so‘z shakllarining xilma-xilligi hamda affikslar tizimini aniqlashdagi muammolarga e’tibor qaratadi. Tadqiqotda korpus lingvistikasi va kompyuter lingvistikasi sohasidagi joriy qiyinchiliklar yoritilib, morfologik ishlov berishning til resurslari sifatini oshirishdagi ahamiyati ta’kidlanadi. Ushbu ish o‘zbek tilini qayta ishlashda yangi imkoniyatlar ochishi bilan birga, kelajakdagi lingvistik tadqiqotlar uchun muhim poydevor vazifasini o‘taydi.

Umuman, B. Elov boshchiligidagi jamoa tomonidan yaratilgan o‘zbek tili milliy korpusi *texnik jihatdan zamonaviy, lingvistik jihatdan boy annotatsiyalangan va statistik jihatdan foydali ma’lumotlar manbai* bo‘lib shakllandi [9]. Bu korpus O‘zbekistonda korpus lingvistikasi rivojini amalda yangi bosqichga olib chiqdi va endilikda turli ilmiy-amaliy maqsadlar uchun tayanch platforma bo‘lib xizmat qilmoqda.

Korpus tuzish jarayonidagi asosiy bosqichlar va yondashuvlar

Har qanday til korpusini yaratish bir necha muhim **bosqichlar** asosida amalga oshiriladi. O‘zbek tili korpusini tuzish tajribasi ham shuni ko‘rsatadiki, quyidagi bosqichlar izchil bajarilganda sifatli korpus shakllantirish mumkin:

1. **Maqsad va qamrovni belgilash:** Dastlab korpusning maqsadi va chegaralari aniqlanadi. Ya’ni, korpus *qanday til materialini* o‘z ichiga oladi, uning *janrlar va uslublar bo‘yicha tarkibi* qanday bo‘lishi lozim – shu kabi mezonlar belgilanadi. O‘zbek milliy korpusini tuzishda ham avvaliga zamonaviy adabiy tilni maksimal darajada qamrab oluvchi korpus yaratish maqsad qilindi. Bunda *badiiy adabiyot, ommaviy axborot vositalari, ilmiy adabiyot, rasmiy uslub* kabi barcha asosiy sohalardan matnlar jalb etilishi rejalashtirildi.
2. **Matnlar to‘plash va raqamlashtirish:** Korpus uchun zarur materiallarni yig‘ish bosqichi ko`p mehnat talab qilinadi. O‘zbek tili uchun matnlar turli manbalardan olindi: *elektron hujjatlar, internet saytlari, elektron kitoblar, gazetalar arxivi, raqamlashtirilgan adabiyotlar* va hokazo. Ba’zi kerakli manbalar faqat qog‘ozda mavjud bo‘lsa, ularni **skanerlash va OCR yordamida raqamli ko‘rinishga keltirish** ishlari qilindi. Shuningdek, o‘zbek tilida **lotin va kirill** yozuvlaridan ikkalasi ham qo‘llangani bois, matnlarni bir xil formatga keltirish masalasi hal etildi: kirillda bo‘lgan matnlar lotinga o‘girildi yoki aksincha, foydalanuvchi uchun qidiruvda qulay bo‘lishi uchun dasturiy yechimlar joriy qilindi. Bu bosqichda matnlarning mualliflik huquqi masalalariga ham e’tibor qaratilib, ochiq manbalar yoki ruxsat etilgan materiallar ishlatildi.
3. **Matnlarni tozalash va teglash:** Toplangan ma’lumotlar bazasi shakllantirilgach, ularni korpusga kiritishdan avval **tozalash (cleaning)** lozim bo‘ladi. Ya’ni, matnlardagi ortiqcha texnik axborot, reklama, kod parchalari kabi til uchun ahamiyatsiz qismlar chiqarib tashlanadi. O‘zbek korpusini tuzishda ham web-sahifalardan olingan matnlar filtrlanib, faqat lingvistik jihatdan **sof matn** qoldirildi. Keyin esa matnlarni korpusga yuklash jarayonida tokenlash, ya’ni matnni alohida so‘zlarga ajratish amalga oshirildi. Keyingi muhim qadam – annotatsiya (teg qo‘yish): bu yerda har bir so‘zga uning lug‘aviy ma’nosi (lemma) va grammatik xususiyatlarini belgilovchi tegar teglar biriktirildi. Annotatsiya darajasiga qarab korpus **teglanmagan korpus** yoki **teglangan korpus** deb ataladi. O‘zbek milliy korpusining ilk versiyasi asosan tokenizatsiya va lemmatizatsiya bilan chegaralangan bo‘lsa, hozirgi kunda morfologik teglash ham qo‘shilmoqda. Ushbu bosqichlarda lingvist-mutaxassislar va dasturchilar birgalikda ishlashi talab etiladi, chunki masalan, **yopiq qismlar** (“va”, “bu” kabi)ni to‘g‘ri aniqlash yoki birikmalarni yaxlit birlik sifatida belgilash kabi masalalar tilshunoslik bilimini ham, dasturlashni ham talab etadi.

4. **Ma’lumotlar bazasini shakllantirish:** Tozalangan va teglangan matnlar korpusning ichki bazasiga yuklanadi. Bunda maxsus korpus platformasi ishlab chiqildi. O‘zbek milliy korpusi uchun maxsus veb-platforma yaratildi va matnlar **janr, sana, muallif, manba kabi metama’lumotlar bilan birga bazaga kiritildi**. Bu kelgusida foydalanuvchilarga, masalan, faqat 2000-yillardan keyingi gazetalar tilini qidirish, yoki faqat badiiy adabiyotdan misollar topish kabi filtrlash imkonini beradi.

5. **Test sinov va tuzatishlar:** Korpus dastlabki shakliga keltirilgach, uni sinovdan o‘tkazish muhim. Bu bosqichda korpusdan turli namuna so‘rovlar qilinib, tizimning to‘g‘ri ishlashi, qidiruv natijalarining relevanti tekshiriladi. Test jarayonida aniqlangan xatolar – masalan, noto‘g‘ri tokenlarga bo‘linib ketgan so‘zlar, xato teglangan shakllar – tuzatiladi. O‘zbek korpusi yaratishda dasturchi va filologlardan iborat ishchi guruh bir necha bor sinov rejimini o‘tkazib, foydalanuvchi nuqtayi nazaridan noqulayliklar bo‘lsa, ularni bartaraf etdi.

6. **Yangilash va boyitib borish:** Korpus yaratish bir martalik jarayon emas, **uzluksiz jarayondir**. Tayyor korpus muntazam yangilanib va kengayib boradi. Masalan, har yili yangi gazeta maqolalari, yangi kitob matnlari qo‘shilishi mumkin, yoki korpusga ilgari kiritilmagan janrlar qo‘shilishi rejalashtiriladi (masalan, internet forumlaridagi yozishmalar, suhbat chatlaridan olingan matnlar va hokazo). O‘zbek tilining milliy korpusini rivojlantirish strategiyasida ham har yili uning hajmini oshirish, yangi **subkorpuslar** qo‘shish (masalan, dialektal nutq korpusi, she’riy matnlar korpusi va hokazo) ko‘zda tutilgan. Shu bilan birga, korpus foydalanuvchilaridan kelib tushayotgan fikr-mulohazalar asosida tizim funktsiyalari takomillashtirilib boriladi.

Korpus tuzish jarayonida ikki xil yondashuv mavjud: **statik korpus** va **dinamik (monitor) korpus** yondashuvlari. Statik korpus belgilangan davr va manbalardan tuzilib, tugallangan ma’lumotlar to‘plamini anglatadi (masalan, ma’lum yillargacha bo‘lgan adabiyotlar bazasi). Dinamik korpus esa doimiy to‘ldirilib boriladi, tilning o‘zgarishi va yangi birliklar paydo bo‘lishini doimiy kuzatish imkonini beradi. O‘zbek milliy korpusi dastlab statik modelda ishga tushgan bo‘lsa-da, hozirda u dinamik tarzda boyitilmoqda. Zero, til uzluksiz rivojlanishda davom etar ekan, korpus ham ushbu jarayonni monitor sifatida kuzatib borishi va aks ettirishi lozim. Xususan, yildan-yilga tilga kirib kelayotgan yangi so‘zlar (neologizmlar) yoki yangi matn janrlari (masalan, internet memlari tili) ham korpusga qo‘shib borilishi bilan uning ilmiy qimmati oshadi.

O‘zbek tilidagi korpusning asosiy maqsadi va vazifalari

O‘zbek tili milliy korpusi bir qator muhim maqsad va vazifalarni amalga oshirish uchun yaratilgan integral platformadir. Uning asosiy maqsadlari quyidagicha izohlanadi:

1. **Tilshunoslik tadqiqotlari:** Korpus, eng avvalo, nazariy va amaliy tilshunoslikda empirik bazani ta’minlashni maqsad qilgan. Ya’ni, o‘zbek tilining lug‘aviy boyligi, grammatik qurilishi, uslubiy xususiyatlari bo‘yicha **dalillarga tayangan tadqiqotlar** qilish uchun korpus asosiy manba bo‘ladi. Misol uchun, olimlar ma’lum bir so‘zning yangi ma’no kasb etgan-etmaganini aniqlash uchun korpusdan foydalanib, turli davrlardagi qo‘llan malarni solishtirishi mumkin. Shu ma’noda korpus til tizimidagi o‘zgarishlarni va bugungi holatni aniqlash vazifasini bajaradi. Korpus yordamida tilshunoslar uchun ilgari imkonsiz bo‘lib kelgan ko‘plab masalalarni (masalan, so‘z birikmalarning statistik xususiyatlarini) o‘rganish imkoni tug‘ildi. Korpusda to‘plangan ma’lumot til jarayonlarining real manzarasini taqdim etadi, bu esa har qanday lingvistik nazariyani ishonchli asoslarga tayantirish imkonini beradi.

2. **Kompyuter lingvistikasi va NLP (Tabiiy tilda ishlov berish):** Bugungi kunda NLP sohasida katta hajmdagi til ma’lumotlarisiz muvaffaqiyatga erishib bo‘lmaydi. O‘zbek tili korpusi shu sohada ham asosiy infratuzilma vazifasini o‘taydi. Misol uchun, mashina tarjimasi tizimlarini yaratishda parallel korpuslar talab etiladi – o‘zbek milliy korpusida maxsus parallel korpus bo‘limi ham yaratilmoqda, unda o‘zbek va xorijiy tillardagi tarjima juftliklari jamlanmoqda. Shuningdek, **matnlarni avtomatik tasniflash, so‘z tanib olish (speech recognition), matnni tahlil qilish (text analytics)** kabi ko‘plab NLP masalalarida korpusdan olinadigan til modellari muhim ahamiyatga ega. Milliy korpus negizida o‘zbek tilining n-gramma modellari, statistik til modellari ishlab chiqilmoqda. Bu modellardan kompyuter lingvistikasi amaliyotida – ovozli yordamchilar, intellektual qidiruv tizimlari, matnni avtomatik tuzatish vositalari kabi ko‘plab yo‘nalishlarda foydalanish mumkin. Korpusning yana bir vazifasi – **agglutinativ tillarga xos NLP muammolarini hal etish** uchun ilmiy maydon yaratishdir: masalan, o‘zbek tilida so‘zlarning turli shakllarini tanib olish (lemmatizatsiya) va ularni bir ma’noli formaga keltirish (disambiguation) bo‘yicha tadqiqotlar aynan korpus orqali sinovdan o‘tkaziladi. Shunday qilib, o‘zbek tili korpusi nafaqat lingvistika, balki sun’iy intellekt va dasturiy yechimlar sohasida ham asosiy resurs rolini o‘ynaydi.

3. **Raqamli lug‘atchilik va leksikografiya:** Korpusning muhim maqsadlaridan yana biri – zamonaviy **elektron lug‘atlar** uchun poydevor yaratish. An’anaviy lug‘atlar odatda tuzuvchilarning tajribasi va cheklangan materiallarga tayanib tuzilardi. Korpus esa lug‘at tuzishda boy va ishonchli material manbasi bo‘lib xizmat qiladi. Masalan, hozirgi kunda tayyorlanayotgan yangi avlod o‘zbek tilining izohli lug‘ati korpusdan olingan misollar bilan boyitilmoqda. Korpusdan **so‘zning eng tipik qo‘llanish misollari, ko‘p ma’noli so‘zlarning har bir ma’nosiga mos kontekstlar, so‘zning kollokatsion (qo‘shma) bog‘lanishlari** kabi ma’lumotlar avtomatik tarzda olinib, lug‘at moddalariga kiritilishi mumkin. Bu esa lug‘atlarni ancha hayotiy va foydali qiladi, foydalanuvchi lug‘atdan so‘zning

haqiqiy ishlatilishi haqida tasavvur oladi. Shuningdek, korpus tez-tez yangilanib boruvchi resurs bo‘lgani uchun, lug‘atlardagi ma’lumotlarni ham osonlik bilan yangilash imkoni mavjud – masalan, oxirgi o‘n yilda paydo bo‘lgan yangi so‘zlar korpusga qo‘shilgan bo‘lsa, lug‘atning navbatdagi nashrida ular aks ettiriladi. **Terminologik lug‘atlar** tuzishda ham korpusdan foydalanilmoqda: muayyan soha matnlari filter orqali ajratib olinib, o‘sha sohaga xos atamalar korpus ichida aniqlanadi va lug‘atga kiritiladi. Demak, korpusning lug‘atchilikdagi vazifasi – **milliy leksikografik bazani boyitish va dalillash**.

4. **Ta’lim va til o‘rgatish:** O‘zbek tili korpusining yana bir ustuvor maqsadi – til ta’limi sohasiga xizmat qilishdir. Korpus materiallari ona tilini yoki chet til sifatida o‘zbek tilini o‘rgatishda innovatsion resurs bo‘la oladi. Xususan, 2021 yilda ishga tushirilgan **“O‘zbek tilining ta’limiy korpusi” (uzschoolcorpara.uz)** ushbu maqsadga yo‘naltirilgan maxsus subkorpusdir. Ta’limiy korpus maktab darsliklari, o‘quv qo‘llanmalar va o‘quv lug‘atlaridan tuzilgan bo‘lib, o‘quvchilar hamda o‘qituvchilar uchun moslashtirilgan. Uning yordamida o‘qituvchilar darsda ishlatish uchun tezda misollar topishi, mashqlar tuzishi mumkin – korpus bir necha soniyada kerakli mavzu bo‘yicha yuzlab real misollarni taqdim eta oladi. Misol uchun, o‘quvchi *“sifat”* mavzusini o‘tayotganda, korpusdan *“yaxshi”, “chiroyli”* kabi sifatlarning haqiqiy matnlardagi qo‘llanishini topib, ular ishtirok etgan gaplarni ko‘rishi mumkin. Bu usul kontekst orqali o‘rgatish tamoyiliga juda mos tushadi. Tilni chet tili sifatida o‘rganuvchilar ham korpusdan foydalansa, ular sun’iy misollardangina emas, balki jonli til materialidan ta’lim oladilar. Korpus asosidagi topshiriqlar, masalan, biror so‘zning kollokatsiyalarini topish, ma’lum grammatik strukturaning qancha va qanday variantlarda qo‘llanishini aniqlash kabi mashqlar talabalar uchun qiziqarli va foydali bo‘ladi. Shuningdek, korpus lingvopedagogik tadqiqotlar uchun ham zamin yaratadi – masalan, xorijiy auditoriya uchun o‘zbek tilidagi eng qiyin so‘zlar yoki tuzilmalarni korpus yordamida statistik tahlil qilish va shu asosda o‘quv materiallarini takomillashtirish mumkin. Demak, milliy korpusning ta’limiy vazifasi – **til o‘qitishda innovatsion elektron vosita** bo‘lish, o‘quv jarayonini boyitish va yengillashtirishdan iborat.

Yuqorida sanab o‘tilgan maqsadlar shuni ko‘rsatadiki, o‘zbek tilining milliy korpusi ko‘p qirrali loyiha bo‘lib, u **ilm-fan, ta’lim, texnologiya va madaniyat sohalarini** bog‘lovchi infratuzilma vazifasini o‘tamoqda. Bir tarafdan, u sof akademik tadqiqotlar uchun qo‘llanilsa, ikkinchi tarafdan jamiyatdagi turli foydalanuvchilar – jurnalistlar, tarjimonlar, yozuvchilar, talabalar – uchun ham foydali axborot manbaidir. Shuning uchun korpusni yaratishda uning ko‘p maqsadli va ko‘p foydalanuvchili tizim ekani inobatga olinadi.

Korpusning asosiy funksiyalari va foydalanuvchilari

Til korpusining funksiyalari deganda, undan foydalangan holda bajarilishi mumkin bo‘lgan amaliy ishlar tushuniladi. O‘zbek tili milliy korpusi quyidagi asosiy funksional imkoniyatlarni taklif qiladi:

1. **Konkordans qidiruvi:** Korpusning eng asosiy funksiyasi – kerakli so‘z yoki iborani barcha kontekstlari (atrofidagi so‘zlari bilan)da topib berishdir. Bu **konkordans qidiruv** deb ataladi. Masalan, foydalanuvchi korpusga *“tarix”* so‘zini kiritib qidirsa, korpus bir necha soniya ichida ushbu so‘z uchragan barcha jumlalarni ro‘yxatlab beradi, bunda *“tarix”* so‘zi har bir misolda ajratib ko‘rsatiladi. Bu usul tadqiqotchiga so‘zning qo‘llanish xilma-xilligini ko‘rish, ma’nosini kontekstda tushunish, hatto **frazeologik birikmalarni aniqlash** imkonini beradi. Konkordans funksiyasi orqali biror adib tilini o‘rganayotgan adabiyotshunos o‘sha adib eng ko‘p ishlatgan so‘zlar ro‘yxatini tuzishi yoki til xususiyatlarini tahlil qilishi mumkin. Shuningdek, til o‘rganuvchilar uchun ham konkordans – boy misollar xazinasi.

2. **So‘zlar statistikasini chiqarish:** Korpusdan nafaqat alohida misollar, balki **umumiy statistik ma’lumotlarni olish** ham mumkin. Masalan, *“mustaqillik”* so‘zi korpusda necha marta uchragan? U qaysi janrda ko‘proq qo‘llanadi – gazeta uslubidami yoki badiiy adabiyotdami? Korpus bunday savollarga javob berish uchun tegishli funksiyani o‘z ichiga oladi. Maxsus hisoblash rejimi orqali berilgan so‘z yoki ibora korpusda *uchrash tezligi (per million ko‘rinishida),* u bilan *eng ko‘p birga keluvchi so‘zlar (kollokatsiyalar)*, hatto vaqt o‘tishi bilan *dinamika (yillarga ko‘ra chastota o‘zgarishi)* kabi statistik ma’lumotlar olinadi. Bu funksiyalar tilshunoslar uchun ayniqsa qimmatlidir, chunki so‘zning trendsini, taraqqiyotini ko‘rish, lug‘at boyligi miqyosini baholash imkonini beradi.

3. **Filtrlash va tanlangan korpuslar (subkorpuslar):** O‘zbek milliy korpusida foydalanuvchilar uchun turli **subkorpuslar** ham ajratilgan. Misol tariqasida, faqat badiiy adabiyot matnlarini yoki faqat ommaviy axborot vositalari matnlarini o‘z ichiga olgan alohida tanlamalarni ko‘rish mumkin. Bu orqali tadqiqotchi o‘ziga kerak yo‘nalishdagi matnlarnigina tahlil qilishi osonlashadi. Misol uchun, lingvist olim rasmiy uslubdagi hujjatlarda yangi paydo bo‘lgan so‘zlarni o‘rganmoqchi bo‘lsa, korpusdan rasmiy hujjatlar subkorpusini tanlab, u yerdan qidiruv va tahlillarni amalga oshiradi. Yana bir muhim filtr – **xronologik filtr**. Bu yordamida ma’lum bir davr (masalan, 1990–2000 yillar) matnlarini ajratib olish va ularda tahlil o‘tkazish mumkin. Natijada, korpus **sinxron (bir davr ichida)** va **diaxron (turli davrlar kesimida)** tadqiqotlar uchun moslashadi. Foydalanuvchilar korpus interfeysida oddiy menyu orqali bunday filtrlarni tanlay oladilar, bu esa murakkab so‘rovlarni sodda usulda shakllantirish imkonini beradi.

4. **Lug‘aviy va grammatik axborot chiqarish:** Korpus tarkibidagi matnlar lingvistik jihatdan teglangani bois, undan grammatik ma’lumotlarni ham izlab topish imkoni bor. Masalan, foydalanuvchi korpusdan faqat fe’l turkumiga oid so‘zlarni

yoki faqat ko‘plik shakldagi otlarni qidirishi mumkin. Bu jarayonda korpusning ichki morphologik teglar bazasi ishlaydi. O‘zbek tilining milliy korpusida, masalan, faqat ko‘plikdagi otlarning ro‘yxatini chiqarish orqali ko‘plik shakllantirish qoidalarining amaliy ko‘rinishini kuzatish mumkin – eng ko‘p *“-lar”* qo‘shimchasi bilan kelgan otlar yoki noodatiy ko‘plik shakllarini topish kabi. Shuningdek, korpusdan lingvistik qurilishlarni izlash imkoniyati ham qo‘shilgan: masalan, *“A dan ko‘ra B”* konstruksiyasini qidirib, uning turli variatsiyalari (masalan, “undan ko‘ra bunday…”) korpus bo‘yicha qanday namunalarda uchrashini ko‘rish mumkin. Bu funksiyalar tilshunoslik izlanishlari uchun ayni muddaodir – ayniqsa, o‘zbek tilidagi **bog‘lanish vositalari, frazeologizmlar** kabi murakkab hodisalarni o‘rganishda korpus noyob imkoniyatlar yaratadi.

Yuqoridagi funksiyalardan turli foydalanuvchilar manfaatdor bo‘lishi mumkin. **Akademik tilshunoslar** korpusdan o‘z izlanishlari uchun asosiy material sifatida foydalanadilar – ular uchun korpus faktlar xazinasi. **Leksikograflar (lug‘at tuzuvchilar)** korpusdan so‘z ma’nolarini tasdiqlovchi misollar, yangi so‘zlarni topish, sinonim va antonimlarni kontekstda aniqlash uchun foydalanadilar. **Adabiyotshunoslar va matnshunoslar** muayyan adib yoki davr tilining xususiyatlarini o‘rganish uchun korpusga murojaat etadilar – masalan, bir yozuvchining uslubini boshqa yozuvchilar uslubi bilan solishtirish kabi. **Tarjimonlar** korpusdan murakkab iboralarning tarjima muqobilini topish yoki ikki til korpusidagi parallellarni ko‘rish orqali o‘z ishlarida yordam oladilar. **O‘qituvchilar va talabalar** korpusdan ilmiy ishlar yozishda, til o‘rganishda interaktiv vosita sifatida foydalanmoqdalar – masalan, talabalar insho yozishda biror atamaning to‘g‘ri qo‘llanishiga ishonch hosil qilish uchun korpusga murojaat qilishi mumkin. **Jurnalistlar va blogerlar** ham korpusdan foyda olishi mumkin: ma’lum bir so‘z ommaviy axborot vositalarida qanday ishlatilayotganini ko‘rib, uslubiy to‘g‘ri qo‘llashga oid xulosa chiqaradilar. Hatto, huquq sohasida, xorij tajribasida bo‘lgani kabi, **til ekspertizasi** uchun ham korpusdan foydalanish imkoniyatlari mavjud – masalan, sud lingvistik tahlillarida matnlarni qiyosiy o‘rganish uchun.

Shunday qilib, o‘zbek milliy korpusining funksional imkoniyatlari keng, foydalanuvchilar doirasi esa turli qatlamlarni qamrab oladi. Bu korpus tilshunos olimdan boshlab oddiy til o‘rganuvchigacha bo‘lgan har bir kishi uchun **foydali elektron manbadir**. Korpus orqali har kim o‘zbek tilining real qo‘llanilishi bilan tanishishi, zarur ma’lumotni olishi va tahlil qilishi mumkin.

O‘zbek tilidagi mavjud korpuslar: baho, ustunlik va kamchiliklar

Hozirgi kunga kelib, o‘zbek tilida bir nechta korpus loyihalari mavjud yoki rivojlantirilmoqda. Ularning har biri o‘ziga xos ustunlik va ayrim cheklovlarga ega. Quyida mavjud korpuslarning qisqacha tavsifi va ularga berilayotgan umumiy baho keltiriladi:

1. **O‘zbek tilining milliy korpusi (uzbekcorpus.uz):** Bu – yuqorida batafsil tavsifi berilgan asosiy milliy korpus. Uning asosiy ustunligi – qamrovi keng va turli, ya’ni o‘zbek adabiy tilining deyarli barcha sohalarini o‘z ichiga olgan. Shuningdek, uning integratsiyalashgan subkorpuslar tizimi mavjud: masalan, parallel korpus, ta’limiy korpus, mualliflik korpusi kabi bo‘limlar milliy korpus tarkibida ajratilgan. Bu esa uni ko‘p funksiyali kompleks tizimga aylantiradi. Milliy korpusning yana bir afzalligi – uning onlayn ochiqligi va bepul foydalanish imkoniyatidir: istalgan foydalanuvchi internet orqali ro‘yxatdan o‘tib, korpusdan foydalanishi mumkin. Shu bilan birga, kamchilik tomonlari ham yo‘q emas. Avvalo, korpus hajmi hali dunyoning yirik korpuslari bilan qiyoslaganda kichik – taxminan o‘n millionlab so‘zlar (aniq raqamlar doimiy o‘sib bormoqda). Mutaxassislar milliy korpus hajmini kelgusida bir necha yuz million so‘zga yetkazishni rejalashtirmoqdalar. Yana bir cheklov – og‘zaki nutq materiallarining kamligi. Hozircha korpus asosan yozma matnlardan iborat bo‘lib, jonli og‘zaki nutq (masalan, audio yozuvlardan transkripsiyalar) bazasi endi-endi shakllanmoqda. Bu esa tilning og‘zaki ko‘rinishini tadqiq etishda noqulaylik tug‘diradi. Milliy korpusning yana bir muammosi – morfologik teglashning to‘liqligi masalasi: hali barcha matnlar to‘liq avtomatik teglangan emas, teglash sifati ham ba’zan oqsaydi (ayniqsa, murakkab qo‘shma so‘zlar yoki dialektal birliklarda). Biroq bu kamchiliklarni bartaraf etish ustida ishlar davom etmoqda – tez orada korpusning yanada takomillashgan yangi versiyasi (“yangi versiya” havolasi mavjud) ishga tushirilishi rejalangan. Umuman, milliy korpus o‘zbek tilini raqamlashtirishdagi tarixiy qadam bo‘lib, uning yaratilishi tilimiz taraqqiyotida muhim madaniy voqelik sifatida baholanmoqda.

2. **Ta’limiy korpus (uzschoolcorpara.uz):** Ushbu korpus milliy korpusning tarkibiy qismi bo‘lsa-da, alohida loyiha sifatida ham e’tirofga loyiq. Ta’limiy korpusning ustun tomoni – maqsadli auditoriyaga mos tanlanganligidir. Maktab darsliklari va o‘quv matnlari asosida tuzilgan bu korpusdan bevosita ta’lim jarayonida foydalanish mumkin. Foydalanuvchi uchun sodda interfeys va didaktik funksiyalar (masalan, sinf darajasi bo‘yicha tekstlarni tanlash, mashq uchun jumlalarni tasodifiy tanlab berish va h.k.) joriy qilingan. Ta’limiy korpus kichikroq hajmga ega (bir necha yuz mingdan bir necha million so‘zgacha), bu esa uni tezkor va yengil qiladi. Kamchilik tomoni – u milliy tilning to‘liq manzarasini bermaydi, chunki faqat darslik materiallari bilan cheklangan. Biroq uning maqsadi boshqa – o‘quvchilar uchun til materiallarini yetkazish. Shuningdek, ayrim matnlarning eskirgan bo‘lishi (eski darsliklardan kiritilgan bo‘lishi) mumkin, bu holatda zamonaviy til holatini aks ettirishda cheklov tug‘iladi. Ta’limiy korpus jamoasi bu muammoni yechish uchun doimiy ravishda yangi darsliklar bilan bazani yangilab bormoqda.

3. **Mualliflik korpusi:** Bu korpus hali rasmiy e’lon qilingan to‘liq mahsulot emas, balki ilmiy tadqiqotlar doirasida yaratilgan kichik korpusdir. Sh.

Hamroyevaning tashabbusi bilan boshlangan ushbu korpusni yaratishdan maqsad – muayyan mualliflar asarlari tilini chuqur o‘rganish, avtorial uslubni statistik tahlil qilishdir. Masalan, dastlabki tajriba sifatida O‘zbekiston xalq shoiri Usmon Azim dramatik asarlarining mualliflik korpusi tuzilgan. Mualliflik korpusining ustunligi shundaki, u adabiyotshunoslik va stilistika uchun noyob material beradi: bitta muallifning barcha asarlari yagona bazada bo‘lgani holda, ularning lug‘at tarkibi, uslubiy qurilishi oson taqqoslanadi. Bu plagiatni aniqlash, mualliflikni aniqlash kabi masalalarda ham qo‘l kelishi mumkin. Kamchiligi esa – qamrov tor, umumtil xususiyatlarini tadqiq etib bo‘lmaydi. Har bir muallif uchun alohida kichik korpus tuzish talab etiladi, bu esa ko‘p mehnat va vaqt oladi. Hozircha mualliflik korpusi loyihasi eksperimental bosqichda, lekin uning nazariy asoslari ishlab chiqilgan (Hamroyevaning PhD tadqiqoti) va kelgusida bu yo‘nalish boyitilishi mumkin.

4. **Parallel korpus:** O‘zbek tilida parallel korpuslar tuzish borasida ham ilk qadamlar tashlangan. Milliy korpus tarkibida parallel bo‘lim mavjud bo‘lsa-da, hozircha unda oz miqdorda material bor (ayniqsa, o‘zbek–ingliz va o‘zbek–rus tarjimalari). Ustunlik shuki, ushbu bazalar mashina tarjimasi va tarjimashunoslik uchun asqatadi – bir jumlaning ikkita tildagi variantini qiyosiy tahlil qilish, adekvat va noadeekvat tarjima namunalarini topish mumkin. Kamchilik tomoni esa – parallel korpus yig‘ish nihoyatda murakkab: sifatli tarjima topish, matnlarni bir-biriga moslashtirib bog‘lash (alignment) ko‘p manbalarni talab qiladi. Shunga qaramay, 2022–2023 yillarda ayrim nodavlat tashabbuslar bilan o‘zbek tilidan ingliz va rus tiliga tarjimalar korpusi yuzaga keldi (masalan, Tatoeba bazasi orqali jamoaviy yig‘ilgan jumla juftliklari). Ayni paytda bu korpuslar kichik hajmda bo‘lsa-da, istiqbolda ularni kengaytirish rejalari bor.

5. **Jahon korpuslari doirasidagi o‘zbek tili korpuslari:** Qizig‘i shundaki, o‘zbek tiliga oid bir necha korpuslar xorijiy ilmiy markazlar tomonidan ham yaratilgan. Masalan, Germaniyadagi Leyptsig universiteti lingvistika instituti Leipzig Corpora Collection dasturi doirasida 2017-yilda “Uzbek web corpus” tuzilgan bo‘lib, unga internetdan avtomatik yig‘ilgan ~9 million so‘zdan iborat matnlar kiritilgan. Bu korpus asosan veb-sahifalardagi jamoat matnlarini o‘z ichiga oladi va alohida qidiruv interfeysi orqali taqdim etilgan. Uning ustunligi – ma’lumot avtomatik to‘plangani sabab tezkor yaratilgan va internet tilini aks ettiruvchi namunalarni beradi. Kamchiligi – tozaligi nisbatan past: chunki avtomatik yig‘ishda reklama, takroriy axborot kabi keraksiz qismlar ham tushgan bo‘lishi mumkin. Shuningdek, Leyptsig korpusida lingvistik annotatsiya chuqur emas – faqat oddiy konkordans qidiruv imkoni bor, xolos. Yana bir misol, Rossiya Fanlar akademiyasi Tilshunoslik institutida “Corpus Turcicum” loyihasi doirasida turkiy tillarning solishtirma korpusi yaratilgan, ularda ham o‘zbekcha matnlar qatnashadi (asosan matnlarning ruscha tarjimalari bilan parallel holatda). Bunday xalqaro loyihalar

o‘zbek tilini jahon korpus maydoniga kiritayotgani bilan ahamiyatli, lekin ular milliy korpusning o‘rnini bosolmaydi – chunki maqsad va mezonlar boshqacha.

Yuqoridagi tahlillar asosida umumiy xulosa qilish mumkinki, **o‘zbek tilining korpuslari** endigina shakllanish bosqichida bo‘lsa-da, qisqa muddatda bir necha turdagi korpuslar paydo bo‘ldi va ular bir-birini to‘ldiruvchi xususiyatga ega. Milliy korpus umumo‘zbektillik manzarani bersa, ixtisoslashgan korpuslar maxsus ehtiyojlarni qondiradi. Ustun jihatlar – tilimiz nihoyat korpusli tillar safiga qo‘shilgani, tadqiqotchilar va mutaxassislar qo‘lida qudratli vosita paydo bo‘lgani. Kamchilik jihatlar – hali hajm va qamrovda cheklilik, annotatsiya darajasining to‘liq emasligi, ayrim korpuslar orasida integratsiya yo‘qligi (masalan, parallel va milliy korpuslar bazasini birlashtirish masalasi). Biroq bu kamchiliklar vaqt o‘tishi bilan bartaraf etilishi tabiiy, zero korpuslarning o‘zi ham doimiy rivojlanishda bo‘ladigan tuzilmalardir.

Jahon tajribasi bilan qisqa qiyosiy tahlil

O‘zbek tilining korpuslari rivoji jahon tajribasi bilan solishtirilganda endigina boshlang‘ich bosqichda ekani ko‘zga tashlanadi. Ingliz, rus, fransuz kabi yirik tillarda korpus lingvistikasining **o‘nlab yillik tarixi mavjud bo‘lib**, ular yaratgan resurslar bugungi kunda billionlab so‘zlarni qamrab oladi. Masalan, Britaniya Milliy Korpusi 1990-yillarda 100 million so‘z bilan tuzilgan bo‘lsa, hozirda **Bank of English** va **Google Ngrams** kabi loyihalar milliardlab so‘zli korpuslarni taqdim etmoqda. Rus tilining milliy korpusi **(RNC)** 300 milliondan ortiq so‘z birligini o‘z ichiga oladi va unda nafaqat grammatik, balki semantik va pragmatik teglash ham joriy etilgan. Bu tillar korpuslarida matnlar soni ko‘pligi, xilma-xilligi, hatto sheva va lahjalar, tarixiy davrlar bo‘yicha alohida bo‘linmalari borligi bilan ajralib turadi.

Shu bilan birga, jahon tajribasi shuni ko‘rsatadiki, **korpus tuzish hech qachon yakunlangan loyiha bo‘lmaydi** – doimo rivojlanadi. Ingliz tili uchun ham keyingi yillarda ixtisoslashgan korpuslar (masalan, Twitter korpusi, huquqiy matnlar korpusi, tibbiy matnlar korpusi) paydo bo‘lgan. O‘zbek tilida ham xuddi shunday yo‘l kuzatilmoqda: umumiy milliy korpus yonida turli sohaviy korpuslar shakllanmoqda (ta’limiy, parallel, mualliflik va hokazo). Bu jihatdan o‘zbek tilining korpus loyihasi jahon andozalarini asta-sekin ergashmoqda, deyish mumkin.

Korpuslarning sifat ko‘rsatkichlari jihatidan qaraganda, yirik tillar korpuslari allaqachon **qo‘shma annotatsiyalangan (morfologik, sintaktik, semantik)** ko‘rinishga o‘tgan. Masalan, Amerika Milliy Korpusi matnlari sintaktik daraxt tuzilmalarigacha belgilangan, shuningdek, korpusdan bevosita mashina tarjima modellarini o‘rgatish, intonatsion tahlillar qilish kabi murakkab funksiyalarni qo‘llab-quvvatlaydi. O‘zbek tilida bunday chuqurlikdagi korpus ishlanmalari hali reja bosqichida, biroq ilmiy izlanishlar boshlangan. B. Elov va hamkorlarining NER (Named Entity Recognition), lemmatizatsiya va sintaktik tahlil bo‘yicha qilayotgan

ishlari buning asosini yaratmoqda. Ya’ni, yaqin yillarda o‘zbek korpusida ham shaxs nomlari, joy nomlari alohida belgilanishi, gaplarning sintaksis daraxti avtomatik chiqarilishi kutilmoqda – bular jahon tajribasida sinovdan o‘tgan va biz uchun yangi bo‘lgan jihatlar.

Qiyosiy tahlilda e’tiborga molik yana bir omil – **resurslar va hamjamiyat** masalasi. Dunyoning yirik korpuslari ortida ko‘pincha katta ilmiy jamoalar, davlat dasturlari yoki universitet laboratoriyalari turadi. O‘zbekistonda korpus yaratish ishlariga so‘nggi yillarda davlat va ilmiy jamiyat e’tibor qaratib, Innovatsion rivojlanish vazirligi grantlari ajratila boshlandi. Masalan, 2020–2022 yillarda “O‘zbek tilining milliy korpusi” bo‘yicha maxsus innovatsion loyiha moliyalashtirildi va bu ishlarni yangi bosqichga olib chiqdi. Bu holat jahon tajribasiga mos – ko‘plab davlatlar milliy korpuslarni **milliy loyihalar darajasida** qo‘llab-quvvatlaydilar (Angliyaning BNC loyihasi Britaniya Kengashi homiyligida bo‘lgani kabi). Bundan tashqari, jahon korpus hamjamiyatlarida (masalan, Corpus Linguistics konferensiyalarida) endi o‘zbek tilining korpusi haqida ham ma’lumotlar paydo bo‘lib, xalqaro mutaxassislar bizning tajribamizni o‘rganishga qiziqish bildirmoqdalar.

Jahon tajribasi bilan qiyoslanganda, o‘zbek korpusining yana bir farqi – **bu tilning xususiyatlari bilan bog‘liq**. Agglutinativ va o‘z kelib chiqishi turkiy bo‘lgan tillar korpuslarini yaratishda turli muammolar bo‘ladi; masalan, turk tilida ham milliy korpus yaratish davomida so‘zlarning juda ko‘p shakllanmasi muammo tug‘dirgan. O‘zbek tilida ham shunday – ingliz yoki rus tillariga qaraganda **morfolgiya ko‘lami keng**, shu bois avtomatik teglash, lemmatizatsiya yechimlari murakkabroq. Qiyosiy o‘rganishlardan ma’lumki, turk tilining milliy korpusini tuzish jarayonida grammatikaviy omillar chuqurroq hisobga olingan, natijada hozir turk milliy korpusida har bir so‘zning butun paradigmatik shakllari bog‘langan holda ko‘rsatiladi (masalan, bir fe’lning barcha zamon va shaxs formalarini bir joyga jamlab ko‘rish mumkin). O‘zbek korpusida hozircha bu jihat cheklangan – lekin aynan shu turdagi funksiyalar ham joriy etish rejalari mavjud.

Xulosa. O‘zbek tili milliy korpusi tilshunoslik va texnologiya sohasida muhim yutuq bo‘lib, uning kelajakdagi rivojlanishi yanada katta imkoniyatlarni ochadi. Korpusning hajmini kengaytirish, annotatsiya sifatini oshirish va zamonaviy bigdata texnologiyalari bilan integratsiya qilish uning samaradorligini oshiradi. Bigdata yordamida korpusni qayta ishlash va tahlil qilish imkoniyatlari tilni o‘rganishda yangi yondashuvlarni keltirib chiqaradi va o‘zbek tilini global miqyosda raqobatbardosh qilishga xizmat qiladi.

O‘zbek tili korpusi tilshunoslik, ta’lim va texnologiya sohalarini integratsiyalashgan holda rivojlantirishning muhim vositasi hisoblanadi. Tarixiy jihatdan u an’anaviy usullardan zamonaviy BigData yechimlariga o‘tgan bo‘lsa-da,

hali ham resurslar etishmovchiligi va metodologik muammolar mavjud. Kelajakda korpusning samaradorligini oshirish uchun xalqaro hamkorlik va AIga asoslangan yondashuvlar asosiy yo‘nalish bo‘lishi kerak. Jahon tajribasi o‘zbek tilining korpusini rivojlantirish uchun o‘ziga xos namuna va motivasiya vazifasini o‘tayapti. O‘zbek korpusi hozircha hajm va chuqurlikda yetakchi tillar korpuslaridan orqada bo‘lsa-da, ularni namunali model sifatida olib, tez sur’atlarda rivojlanmoqda. Bir necha yil ichida o‘zbek tilining milliy korpusi Markaziy Osiyo tillari orasida eng ilg‘or elektron til resursiga aylanishi kutilmoqda. Eng muhimi, milliy korpusimiz jahon korpuslar oilasiga kirib, xalqaro hamkorlik uchun ham ochiq bo‘lmoqda – masalan, ko‘plab xorijiy tadqiqotchilar endi uzbekcorpus.uz ma’lumotlaridan foydalana olishadi. Bu esa tilimizning xalqaro maydondagi nufuzini oshirishga xizmat qilishi shubhasiz. Ushbu paragrafda Milliy korpus loyihasi bo‘yicha ma’lumotlar O‘zbekistonda chop etilgan ilmiy maqolalar va konferensiya materiallariga tayangan holda yoritildi, xususan B. Elov, Sh. Hamroyeva, N.Z. Abdurahmonova va ularning hamkorlarining ishlaridan olindi. Mazkur bo‘limdagi dalillar va tahlillar ushbu manbalardagi faktlarga asoslangan bo‘lib, ilmiy izchillik tamoyiliga amal qilingan. Bu yondashuv o‘zbek tili korpusining tarixiy shakllanishini, maqsad va vazifalarini chuqur anglash hamda uni jahon tajribasi bilan bog‘lash imkonini beradi.

Foydalanilgan adabiyotlar:

1. McEnery, T., & Hardie, A. (2012). Corpus linguistics: Method, theory and practice. Cambridge University Press.
2. Abdurahmonova N.Z. O‘zbek tili elektron korpusining kompyuter modellari: Filol. fan. d-ri. (DSc) diss. avtoreferati. – Toshkent, 2021.
3. https://uzschoolcorpara.uz/uz/Search
4. Elov, B., & Xudayberganov, N. (2024). O‘zbek tili korpusi matnlarini pos teglash usullari. Computer Linguistics: problems, solutions, prospects, 1(1).
5. Hamroyeva, S., Abdullayeva, O., & Uzoqova, M. (2022). O‘zbek tilida pos tegging masalasi: muammo va takliflar. Uzbekistan language and culture, 5(2), P. 51-68.
6. https://scholar.google.com/citations?user=wYCKlsoAAAAJ&hl=ru
7. Qarshiyev, A., Tursunov, M., & Maxmidov, S. (2022). O‘zbek tili milliy korpusini loyihalash. Computer linguistics: problems, solutions, prospects, *1*(1).
8. Kasimova, M. B. (2024). O‘zbek tili milliy korpusida darajalanish hodisasining berilishi. Miasto Przyszłości, *44*, 641-645.
9. Xudayberganov, N. (2024). O‘zbek tili korpusiga morfologik ishlov berish. Computer linguistics: problems, solutions, prospects, *1*(1).
10. Elov, B., Hamroyeva, S., Alayev, R., Xusainova, Z., & Yodgorov, U. (2023). O‘zbek tili korpusi matnlarini qayta ishlash usullari. Digital transformation and artificial intelligence, *1*(3), 117-129.